Part 6 Fundamental Limits in Information Theory

Introduction

- Information Theory is the fundamental theory behind information manipulation, including data compression and data transmission.

Introduction

- For years, researchers wish to seek answers to some fundamental questions on information manipulation:
  - What is the irreducible complexity below which an informational signal cannot be compressed? Entropy.
  - What is the ultimate transmission rate for reliable communication over a noisy channel? Capacity.
  - A more striking result
    - If the entropy of the source is less than the capacity of the channel, then (asymptotic) error-free (or arbitrarily small error) communication over the channel can be achieved.

Uncertainty, Information, and Entropy

- Uncertainty = Information
  - When one gains "information", he/she shall lose "uncertainty".
  - Example. Suppose a discrete random variable S takes value from $\mathcal{S} = \{s_0, s_1, \ldots, s_{K-1}\}$ with probability

$$\Pr(S = s_k) = p_k, \quad k = 0, 1, \ldots, K-1$$

How much information we gain if we observe the outcome of S (provided that we already know the statistics of S)?

Uncertainty, Information, and Entropy

- Case 1: $p_0 = 1$.
  - Since we know that $\boldsymbol{s}_0$ will be observed before we observe it, no uncertainty loses (namely, no information gains) for this observation.
- Case 2: $p_k < 1$ for every k.
  - Can we quantitatively measure the "information" amount we gain after observing S?

Uncertainty, Information, and Entropy

- Axioms for information (uncertainty) measure
  - Monotonicity in event probability
    - If an event is less likely to happen, it is more uncertain that the event would happen. Therefore, its degree of uncertainty should be higher.
  - Additivity
    - If S and T are two independent random variables, then the uncertainty loss due to the observations of both S and T should be equal to the sum of the uncertainty loss due to the observation of S and the uncertainty loss due to the observation of T.

Uncertainty, Information, and Entropy

- Continuity
  - A small adjustment in event probability should induce a small change in event uncertainty.

Uncertainty, Information, and Entropy

- Mathematically, the three axioms are transformed to:
  - $I(p)$ is monotonically decreasing in event probability p.
  - $I(p_1 \times p_2) = I(p_1) + I(p_2)$.
  - $I(p)$ is a continuous function of event probability p for $0 \leq p \leq 1$.
- It can be proved that the only function that satisfies the above three conditions is:

$$I(p) = C \cdot \log\left(\frac{1}{p}\right), \text{ where } C \text{ is a positive constant.}$$

Uncertainty, Information, and Entropy

- Hence, for random variable *S*, each outcome, when it does happen, will respectively give the observer "information"

$$I(p_0) = C \cdot \log(1/p_0)$$
$$I(p_1) = C \cdot \log(1/p_1)$$
$$\vdots$$
$$I(p_{K-1}) = C \cdot \log(1/p_{K-1})$$

- In expectation value, the random variable *S* will give the observer information amount

$$H(S) = C \cdot \sum_{k=0}^{K-1} p_k \log(1/p_k)$$

Uncertainty, Information, and Entropy

- When $p_0 = ½$, s_0 is either observed or not observed with equal probable. Hence, intuitively, one should learn one bit of information after this observation.

$$I(p_0) = C \cdot \log(1/p_0) = C \cdot \log(2) = 1$$

- As a result, take $C = 1/\log(2)$.

$$H(S) = \sum_{k=0}^{K-1} p_k \log_2(1/p_k) \text{ bits}$$

- This is named entropy of S.

Uncertainty, Information, and Entropy

- Relation of entropy and data compression
  - For a single random variable with K possible outcomes, a straightforward representation is to use

$$\lceil \log_2(K) \rceil \text{ bits.}$$

    For example, $K = 8$. Then, use 3 bits to represent each outcome.
  - For a sequence of observations on *independent and identically distributed* (i.i.d.) $S_1, S_2, S_3, \ldots$, we will use

$$\lceil \log_2(K) \rceil \text{ bits per observation}$$

    to represent the sequence. **Can we reduce this number?**

Uncertainty, Information, and Entropy

- (Source-Coding Theorem) In 1948, Shannon proved that
  - (Converse) the minimum average number of bits per observation to **losslessly** and **uniquely-decodably** represent an i.i.d. sequence $S_1, S_2, S_3, \ldots$, is lower-bounded by the entropy $H(S)$.
  - (Achievability) the minimum average number of bits per observation to **losslessly** and **uniquely-decodably** represent an i.i.d. sequence $S_1, S_2, S_3, \ldots$, can be made arbitrarily close to $H(S)$.
- Thus, the information quantitative measure finds its theoretical footing!

- **Unique decodability** is an essential premise for Shannon's source coding theorem
  - Unique decodability = Concatenation of codewords (without punctuation mechanism) can be uniquely decodable.

$$\begin{aligned}
\text{codeword of } A &= 0 \\
\text{codeword of } B &= 1 \\
\text{codeword of } C &= 00 \\
\text{codeword of } D &= 01 \\
\text{codeword of } E &= 10 \\
\text{codeword of } F &= 11
\end{aligned} \quad \Rightarrow 01 = \text{either } AB \text{ or } D$$

For uniform distribution, $H(S) = \log_2(6) \approx 2.58$,
but, average codeword length $10/6 = 1.67 < H(S)$

- A uniquely decodable code must satisfy the Kraft-McMillan inequality (or Kraft's inequality).

$$\sum_{k=0}^{K-1} 2^{-\ell_k} \leq 1, \text{ where } \ell_k = \text{length of the } k\text{th codeword}$$

$$\begin{aligned}
\text{codeword of } A &= 0 \\
\text{codeword of } B &= 1 \\
\text{codeword of } C &= 00 \\
\text{codeword of } D &= 01 \\
\text{codeword of } E &= 10 \\
\text{codeword of } F &= 11
\end{aligned}$$

$$\Rightarrow 2^{-1} \times 2 + 2^{-2} \times 4 = 2 > 1$$

A violation to
the Kraft-McMillan inequality !

- **Arbitrarily close to entropy of average codeword length for a sequence of i.i.d. random variables**

$$\begin{aligned} \text{probability of } A &= 0.8 \\ \text{probability of } B &= 0.1 \\ \text{probability of } C &= 0.1 \end{aligned}$$

Best code for single letter

$$\begin{aligned} \text{codeword of } A &= 0 \\ \text{codeword of } B &= 10 \\ \text{codeword of } C &= 11 \end{aligned}$$

Average codeword length

$0.8 \times 1 + 0.2 \times 2 = 1.2 > H(S) = 0.92$

Best code for double letters

$$
\begin{aligned}
\text{codeword of } AA &= 0 \\
\text{codeword of } AB &= 100 \\
\text{codeword of } AC &= 101 \\
\text{codeword of } BA &= 110 \\
\text{codeword of } BB &= 111100 \\
\text{codeword of } BC &= 111101 \\
\text{codeword of } CA &= 1110 \\
\text{codeword of } CB &= 111110 \\
\text{codeword of } CC &= 111111
\end{aligned}
$$

By increasing the number of letters encoded at a time (at the price of memory consumption), one can make the average codeword length approaching the source entropy.

Average codeword length

$$\frac{0.64(1\times 1)+0.08(3\times 3+4\times 1)+0.01(6\times 4)}{2} = 0.96 > H(S) = 0.92$$

Codeword length

Number of codewords

Uncertainty, Information, and Entropy

- Some properties regarding entropy
  - $H(S) \geq 0$ with equality holding if, and only if (iff), S is deterministic.
  - $H(S) \leq \log_2(K)$ with equality holding iff S is uniformly distributed.

The first item can be proved by
that "$0 \leq p_k \leq 1$" implies "$\log(1/p_k) \geq 0$".
Equality holds if, and only if, "$p_k \log(1/p_k) = 0$ for every k.

$$
\begin{aligned}
\log(K) - H(S) &= \log(K) \times \left(\sum_{k=0}^{K-1} p_k \right) - \sum_{k=0}^{K-1} p_k \log(1/p_k) \\
&= \sum_{k=0}^{K-1} p_k \times \log(K) + \sum_{k=0}^{K-1} p_k \log(p_k) \\
&= \sum_{k=0}^{K-1} p_k \log[K \times p_k] \\
&\geq \sum_{k=0}^{K-1} p_k \left(1 - \frac{1}{K \times p_k} \right) \\
&= \sum_{k=0}^{K-1} \left(p_k - \frac{1}{K} \right) \\
&= 1 - 1 = 0.
\end{aligned}
$$

$(\forall y > 0)\ \log(y) \geq 1 - (1/y)$
with equality holding iff $y = 1$

Equality holds if, and only if, $(\forall\ 0 \leq k \leq K-1)$, $K \times p_k = 1$, which means S is uniformly distributed.

Uncertainty, Information, and Entropy

- Definition of discrete memoryless source (DMS) S_1, S_2, S_3, …
  - Discrete = The alphabet of each S_i is discrete.
  - Memoryless = Independent among S_i (from the text)
    - Memoryless = "Identically distributed" in addition to "independent" (otherwise we need to check the time instance i in order to identify the statistics)

Data Compaction

- Data compaction or lossless data compression
  - To remove the redundancy with no loss of information

- A sample uniquely decodable code – Prefix code
  - Prefix condition: No codeword is the prefix of other codeword.
  - A prefix code satisfies the prefix condition.

Data Compaction

- A prefix code is uniquely decodable, thereby satisfying the Kraft-McMillan inequality.
- Converse is not necessary true.
  - A uniquely decodable code is not necessarily a prefix code. For example,

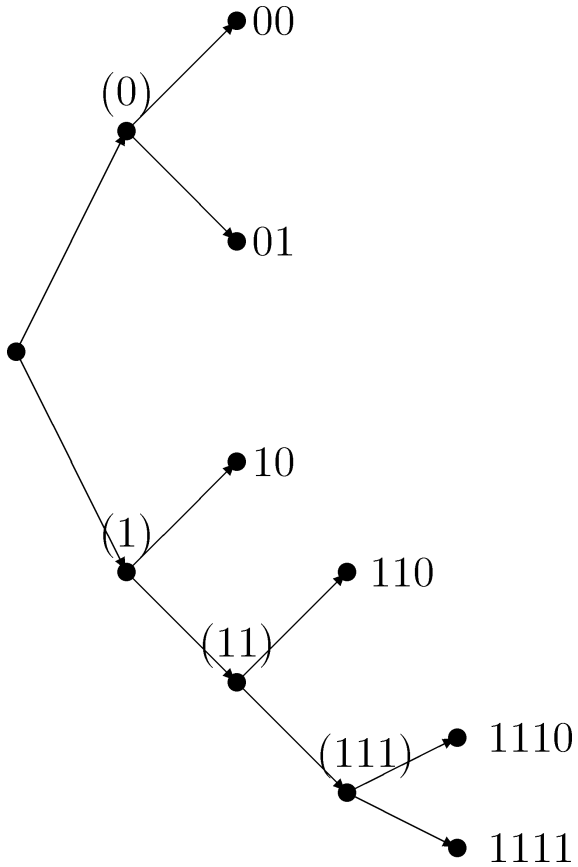
    a uniquely decodable non-prefix code

    Codeword of A = 0
    Codeword of B = 01
    Codeword of C = 011
    Codeword of D = 0111

Data Compaction

The codewords of this prefix code are 00, 01, 10, 110, 1110 and 1111.

- The codewords of a prefix code, when the code is represented by a code tree, are always located at the leaves.
- Conversely, a prefix code can be formed by selecting the binary sequences corresponding to leaves on a code tree.



Data Compaction

- Enforced by the prefix condition (or more specifically, enforced by the fact that the codewords are all residing at the leaves of a code tree), the prefix codeword can be ***instantaneously*** decoded upon the reception of the last bit.

- For this reason, it is also named the **instantaneous code**.

Data Compaction

- Due to the tree-leave graphical representation of prefix codes, it can be shown that:

  For any positive integers $\ell_0, \ell_1, \cdots, \ell_{K-1}$ that satisfy the Kraft-MaMillan inequality, there exists a prefix code that takes these numbers as its codeword lengths.

- With this property, we can prove that there exists a prefix code whose average codeword length satisfies

$$H(S) \leq \bar{L} = \sum_{k=0}^{K-1} p_k \ell_k \leq H(S) + 1$$

Take $\ell_k = \lfloor \log_2(1/p_k) \rfloor + 1 \geq \log_2(1/p_k)$.

Then, $2^{-\ell_k} \leq p_k$.

$$\Rightarrow \sum_{k=0}^{K-1} 2^{-\ell_k} \leq \sum_{k=0}^{K-1} p_k = 1.$$

$\Rightarrow$ There exists such a prefix code with $\ell_0, \ell_1, \cdots, \ell_{K-1}$ as its codeword lengths.

$\Rightarrow$ Uniquely decodable $\Rightarrow \bar{L} \geq H(S)$

On the other hand, $\ell_k \leq \log_2(1/p_k) + 1$.

$$\Rightarrow \bar{L} = \sum_{k=0}^{K-1} p_k \ell_k \leq \sum_{k=0}^{K-1} p_k \log_2\left(\frac{1}{p_k}\right) + \sum_{k=1}^{K-1} p_k = H(S) + 1.$$

With this property, if we compress two symbols at a time, then

$$H(S^2) \leq \bar{L}_2 \leq H(S^2) + 1$$

With this property, if we compress three symbols at a time, then

$$H(S^3) \leq \bar{L}_3 \leq H(S^3) + 1$$

$$\vdots$$

With this property, if we compress n symbols at a time, then

$$H(S^n) \leq \bar{L}_n \leq H(S^n) + 1$$

As a result, we can make the average codeword length per observation arbitrarily close to the source entropy.

$$H(S) = \frac{1}{n} H(S^n) \leq \frac{1}{n} \bar{L}_n \leq \frac{1}{n} H(S^n) + \frac{1}{n} = H(S) + \frac{1}{n}$$

Data Compaction

- The optimal code for lossless compression – Huffman code

Give a source with probability $\{p_0, \ldots, p_{K-1}\}$.
Let ℓ_k be the binary codeword length of the kth symbol.
Then there exists an optimal uniquely decodable variable-length code satisfying:

1. $p_i > p_j$ implies $\ell_i \leq \ell_j$.
2. The two longest codewords have the same length.
3. The two longest codewords differ only in the last bit and correspond to the two least-frequent symbols.

The idea behind the proof is to show that any code that violates any of the three conditions cannot be the one with the smallest average codeword length.

Huffman encoding algorithm:

1. Combine the two least probable source symbols into a new single symbol, whose probability is equal to the sum of the probabilities of the original two.

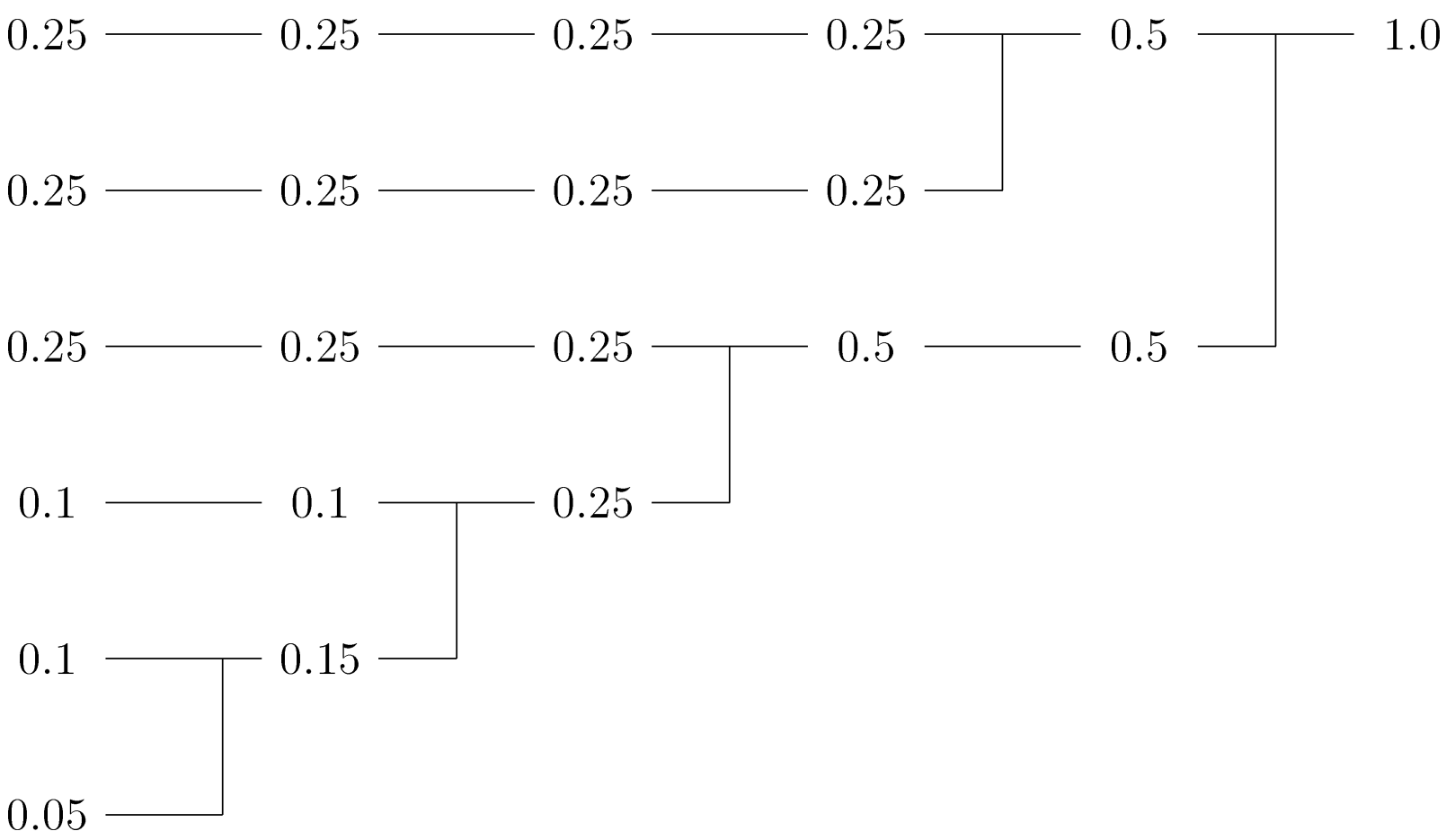
   - Thus we have to encode a new source alphabet of one less symbol.

   Repeat this step until we get down to the problem of encoding just two symbols in a source alphabet, which can be encoded merely using 0 and 1.

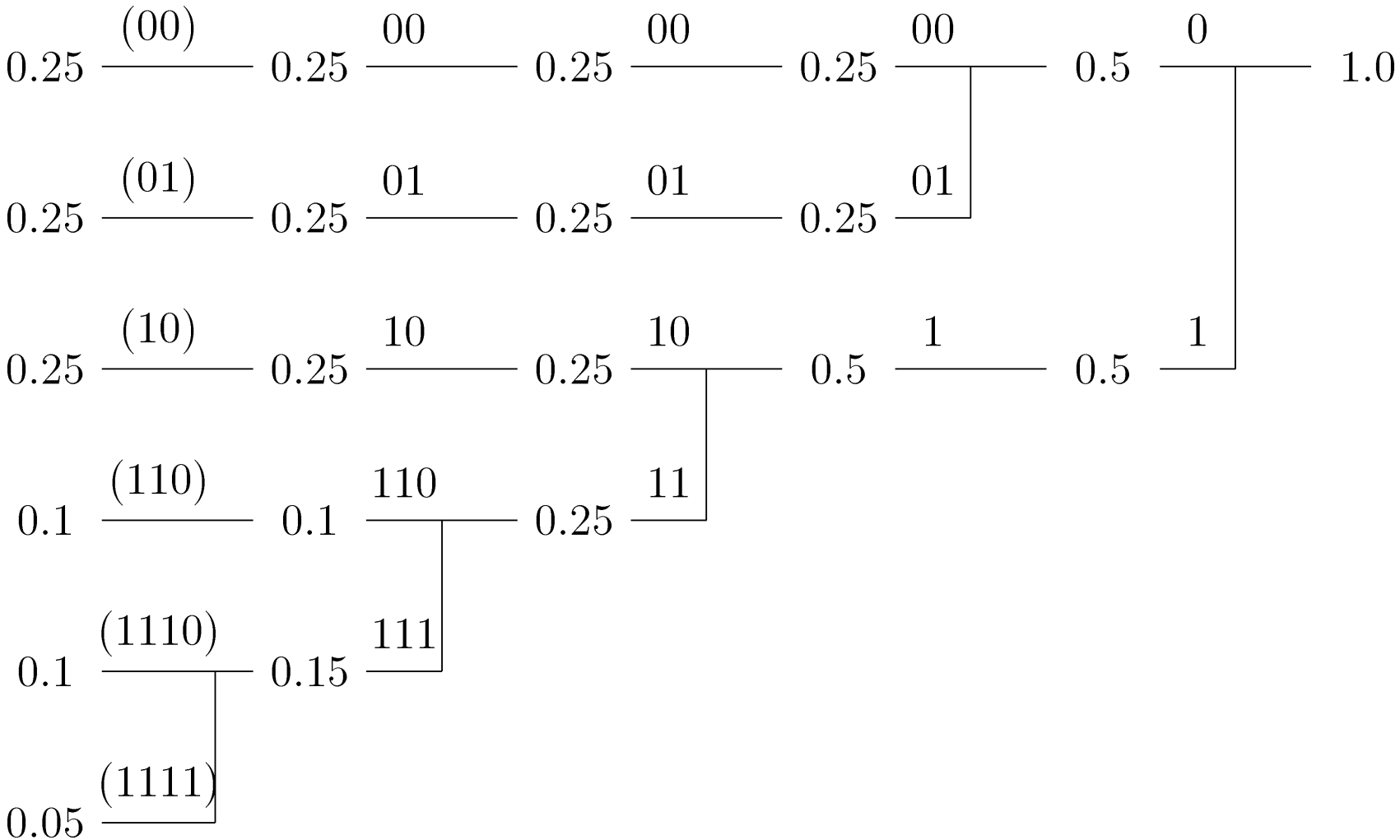
2. Go backward by splitting one of the two (combined) symbols into two original symbols, and the codewords of the two split symbols are formed by appending 0 for one of them and 1 for the other from the codeword of their combined symbol. Repeat this step until all the original symbols have been recovered, and obtain a codeword.

Example. Consider a source with alphabet {1, 2, 3, 4, 5, 6} with probability 0.25, 0.25, 0.25, 0.1, 0.1, 0.05, respectively.

Step 1:



Step 2:


0.25
(00)
0.25
00
0.25
00
0.25
00
0.5
0
1.0
0.25
(01)
0.25
01
0.25
01
0.25
01
0.25
(10)
0.25
10
0.25
10
0.5
1
0.5
1
0.1
(110)
0.1
110
0.25
11
0.1
(1110)
0.15
111
0.05
(1111)

Data Compaction

- Variance of average codeword length of Huffman code

$$\sigma^2 = \sum_{k=0}^{K-1} p_k(\ell_k - \bar{L})^2$$

  - When the probability of a newly combined symbol is found to be equal to another probability in the list, the selection of any one of them as the next symbol to be combined will yield the same average codeword length but different variance.
  - A trick is to avoid using the newly combined symbol in order to minimize the variance. (See the example in the text.)

Data Compaction

- Lempel-Ziv code – An asymptotically optimal (i.e., achieving source entropy) universal code
  - Huffman coding requires the knowledge of probability of occurrence for each symbol; so, it is not "universally" good.
  - Can we design a coding scheme that is universally good (achieving source entropy) for a class of sources, such as, for all i.i.d. sources?

Lempel-Ziv coding scheme

1. Parse the input sequence into strings that have never appeared before.

2. Let L be the number of distinct strings of the parsed source. Then we need $\lceil \log_2(L) \rceil$ bits to index these strings (starting from one). The codeword of each string is the index of its prefix concatenated with the last bit in its source string.

- **Example.** The input sequence is 10110101000010;
- Step 1:
  - With default strings 0 and 1 in the buffer, the algorithm first eats the first letter 1 and finds that it is one of the default strings. So, it eats another letter 0, and yields a new string 10. So 10 is the *first string*.
  - Then the algorithm eats the third letter 1, and finds that it has appeared before. Hence, it keeps eating the next letter until a new string is formed.
  - By repeating these procedures, the source sequence is parsed into strings as

$$0, 1, 10, 11, 01, 010, 00, 10.$$

- Step 2:
  - $L = 8$. So the indices will be:

$$\begin{array}{rcccccccc} \textit{parsed source}: & 0 & 1 & 10 & 11 & 01 & 010 & 00 & 10 \\ \textit{index}: & 001 & 010 & 011 & 100 & 101 & 110 & 111 & - \end{array}.$$

  - E.g., the codeword of source string 010 will be the index of 01, i.e. 101, concatenated with the last bit of the source string, i.e., 0.
- The resultant codeword string of $10, 11, 01, 010, 00, 10$ is:

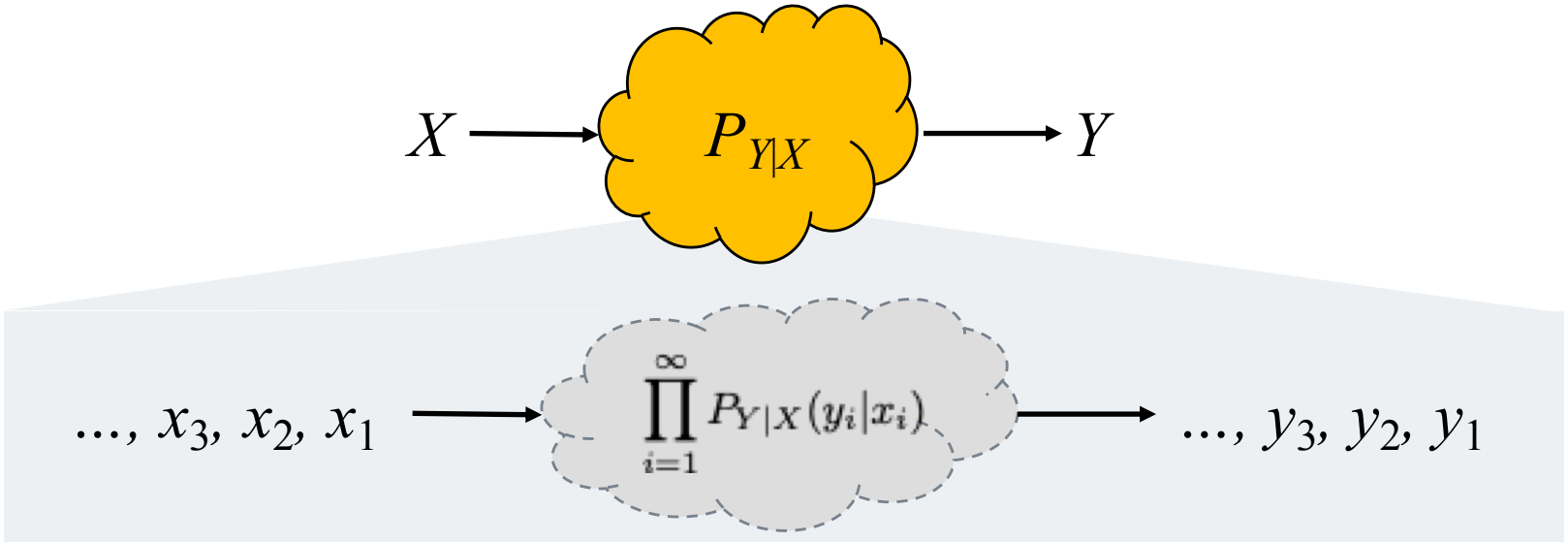
$$(010, 0)(010, 1)(001, 1)(101, 0)(001, 0)(010, 0)$$

or equivalently,

$$010001010011101000100100.$$

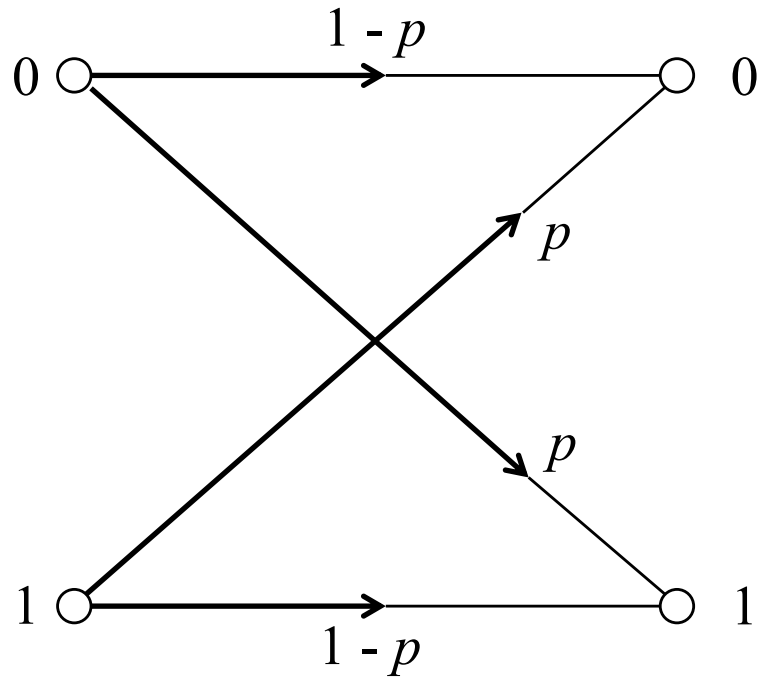
Discrete Memoryless Channels

- Discrete memoryless channels (DMCs)
  - Discrete input and output alphabets
  - Current output only depends on current input, where the "dependence" is time invariant.



Discrete Memoryless Channels

- Example. (Memoryless) Binary symmetric channel
  - The simplest discrete memoryless channel



Mutual Information

- Assumptions in transceiving operation
  - The receiver
    - Does not know the channel input x
    - Does not know the channel noise as well as how the channel noise affects x
    - Do know the channel output y
    - Do know the distribution of input x
    - Do know the transition probability of output y given input x

Mutual Information

- Assumptions in transceiving operation
  - The receiver
    - Do know the information content of channel input to be transmitted, i.e., the input entropy $H(X)$.
    - After observing y, i.e., after transmission, the remaining uncertainty for receiver about the input is $H(X|Y)$.

The information that successfully conveys from input to output is $H(X) - H(X|Y)$.

Mutual Information

- Conditional entropy $H(X|Y)$

Self-information $\log_2\left(\frac{1}{p(x|y)}\right)$

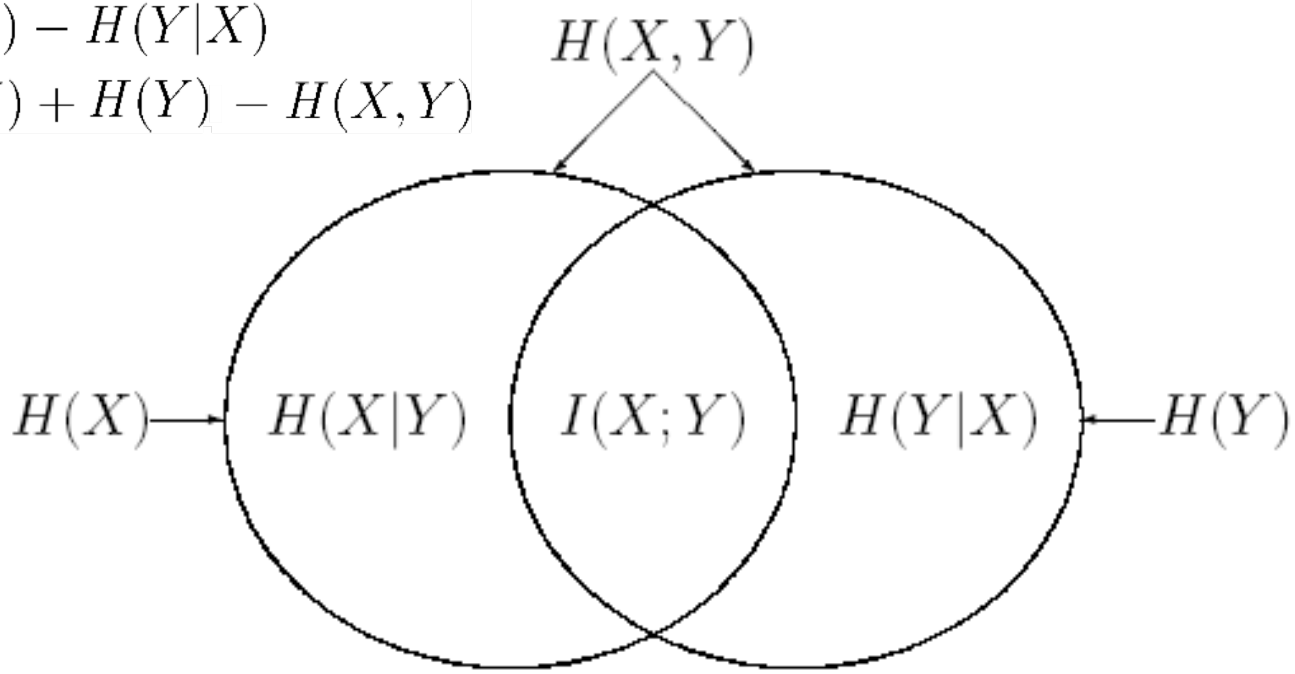
Average self-information $E\left[\log_2\left(\frac{1}{p(X|Y)}\right)\right]$

$$= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log_2\left(\frac{1}{p(x|y)}\right)$$

Mutual Information

- Mutual information

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



Mutual Information

- Properties of mutual information
  - $I(X;Y) = I(Y;X)$
  - $I(X;Y) \geq 0$ with equality holding iff $X \perp\!\!\!\perp Y$

    No information ($H(X) - H(X|Y)$) can be conveyed from input to output if Y is independent of X.

$$
\begin{aligned}
I(X;Y) &= H(X) - H(X|Y) \text{ (measured in unit of nats)} \\
&= \sum_{x\in\mathcal{X}} p(x)\log\left(\frac{1}{p(x)}\right) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{1}{p(x|y)}\right) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{1}{p(x)}\right) - \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{1}{p(x|y)}\right) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x|y)}{p(x)}\right) \\
&= \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\log\left(\frac{p(x,y)}{p(x)p(y)}\right) \\
&\geq \sum_{x\in\mathcal{X}}\sum_{y\in\mathcal{Y}} p(x,y)\left(1-\frac{p(x)p(y)}{p(x,y)}\right) \\
&= 1-1=0
\end{aligned}
$$

$(\forall y>0)\ \log(y)\geq 1-(1/y)$ with equality holding iff $y=1$

Equality holds if, and only if, for $x\in\mathcal{X}$ and $y\in\mathcal{Y}$, $p(x,y)=p(x)p(y)$.

Channel Capacity

- Channel capacity
  - A new terminology introduced by Shannon

$$I(X;Y) = \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p(y|x) \log_2 \frac{p(y|x)}{\sum_{\tilde{x}\in\mathcal{X}} p(\tilde{x})p(y|\tilde{x})}$$

  - $p(y|x)$ is fixed and is given by the channel.
  - $I(X;Y)$ is the information that can be conveyed through the channel.
  - $p(x)$ is the way we use the channel.
  - How about we choose the right $p(x)$ to maximize the information conveyed through the channel!

Channel Capacity

- So, Shannon calculated

$$
\begin{aligned}
C &= \max_{p(x)} I(X;Y) \\
&= \max_{p(x)} \sum_{x\in\mathcal{X}} \sum_{y\in\mathcal{Y}} p(x)p(y|x) \log_2 \frac{p(y|x)}{\sum_{\tilde{x}\in\mathcal{X}} p(\tilde{x})p(y|\tilde{x})}
\end{aligned}
$$

- Then, he asked himself "What is the operational meaning of this quantity?"

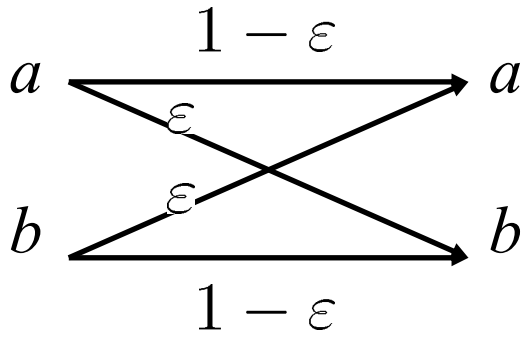
Channel Coding Theorem

- Shannon found that
  - if R (bits/channel usage) $> C \Rightarrow P_e$ (word error rate) cannot be made arbitrarily small;
  - if R (bits/channel usage) $< C \Rightarrow P_e$ (word error rate) can be made arbitrarily small;

Word error rate (WER)= Codeword error rate

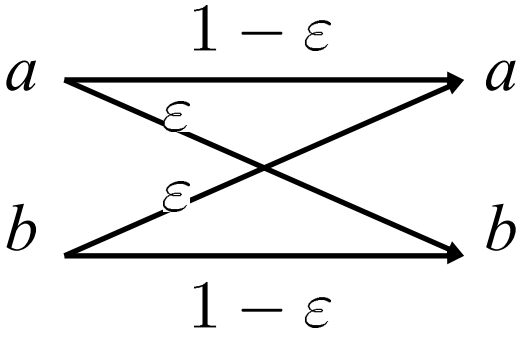
Channel Coding Theorem

- Code rate in data transmission
  - Binary symmetric channel



$\varepsilon = 0.1$

  - WER for one information bit per channel usage (i.e., for an uncoded system) is ε.



$$\text{encoder} \begin{cases} 0 \to aaa \\ 1 \to bbb \end{cases} \Rightarrow \text{decoder} = \text{majority law}$$

$$\text{Error occurs when, e.g., transmit } aaa \text{ but receive } \begin{cases} abb \\ bab \\ bba \\ bbb \end{cases}$$

- WER for 1/3 information bit per channel usage is given by:

$$\begin{aligned} P_e &= \varepsilon^2(1-\varepsilon) + \varepsilon^2(1-\varepsilon) + \varepsilon^2(1-\varepsilon) + \varepsilon^3 \\ &= 3(0.1)^2(0.9) + 0.1^3 = 0.028 < 0.1 \end{aligned}$$

- WER is reduced at the price of code rate reduction (i.e., transmission speed reduction).

- How about comparing the WERs of all codes with the same code rate (i.e., the same transmission speed)? For example, $R = 1/3$.

- Shannon proved that under $\varepsilon = 0.1$, if

$$R < C = 1 - H(\varepsilon) = 1 - 0.468996 = 0.531004$$

P_e can be made arbitrarily close to zero,

where $H(\varepsilon) = \varepsilon \log_2 \frac{1}{\varepsilon} + (1-\varepsilon) \log_2 \frac{1}{1-\varepsilon}$.

- In terminology, "reliable" in data transmission means that WER can be made arbitrarily small.
- Thus $R = 1/3$ is a reliable code rate (i.e., a reliable data transmission rate).

Channel Coding Theorem

- Note that the left-hand-side (code rate) and the right-hand-side (channel capacity) should be measured (or calculated) in the same unit.

$$R \text{ (information bits generated per unit time } T_s)$$
$$C \text{ (information bits per channel usage period } T_c)$$

We can compare $R < C$ only when $T_s = T_c$.

Otherwise, we should compare

$$\frac{R}{T_s} \lessgtr \frac{C}{T_c} \text{ information bits per second}$$

Differential Entropy and Mutual Information for Continuous Ensembles

- Entropy of a continuous source
  - Suppose a continuous source X has continuous density f.
  - Transform the continuous source X into a discrete one X^{Δ} by quantization with step size D.

$$X^{\Delta} = i \quad \text{if } i\Delta \le X < (i+1)\Delta$$

$$\text{and } \Pr[X^{\Delta} = i] = \int_{i\Delta}^{(i+1)\Delta} f(x)dx = f(x_i)\Delta$$

$$\text{for some } i\Delta \le x_i < (i+1)\Delta.$$

(Mean-value theorem)

The entropy of X^{Δ} is therefore given by:

$$\begin{aligned} H(X^{\Delta}) &= \sum_{i=-\infty}^{\infty} f(x_i)\Delta \log_2 \frac{1}{f(x_i)\Delta} \\ &= \sum_{i=-\infty}^{\infty} f(x_i)\Delta \log_2 \frac{1}{f(x_i)} + \sum_{i=-\infty}^{\infty} f(x_i)\Delta \log_2 \frac{1}{\Delta} \\ &= \sum_{i=-\infty}^{\infty} \Delta \left(f(x_i) \log_2 \frac{1}{f(x_i)} \right) + \log_2 \frac{1}{\Delta} \end{aligned}$$

$$\lim_{\Delta \downarrow 0} H(X^{\Delta}) = \underbrace{\int_{\Re} f(x) \log_2 \frac{1}{f(x)} dx}_{h(X)} + \lim_{\Delta \downarrow 0} \left(\log_2 \frac{1}{\Delta} \right).$$

- Two observations can be made on the entropy of a continuous source.
  - Its entropy is infinite (due to the second term $\log_2(1/\Delta)$); so, a continuous source contains infinite number of information bits (amount).

$$\lim_{\Delta \downarrow 0} \left(\log_2 \frac{1}{\Delta} \right) = \infty$$

  - The first term may be viewed as the quantization efficiency for the source, and is named the differential entropy $h(X)$.
    - To uniformly quantize X up to n-bit accuracy requires $\Delta = 2^{-n}$.
    - However, to losslessly express the quantization result (so that the accuracy remains) requires (approximately)

$$h(X) + n \text{ bits}$$

Differential Entropy and Mutual Information for Continuous Ensembles

- Richness in information content in Gaussian source
  - Example:
    - Let X be a Gaussian random variable with mean μ and variance σ^2.
    - Let Y be a random variable with mean μ and variance σ^2.
    - Then,

$$h(Y) \leq h(X) = \frac{1}{2}\log_2(2\pi e\sigma^2)$$

$$
\begin{aligned}
h(X) &= \int_{\Re} f_X(x) \log_2 \frac{1}{f_X(x)} dx \\
&= \int_{\Re} f_X(x) \left[\frac{1}{2} \log_2(2\pi\sigma^2) + \log_2(e) \cdot \frac{(x-\mu)^2}{2\sigma^2} \right] dx \\
&= \int_{\Re} f_Y(x) \left[\frac{1}{2} \log_2(2\pi\sigma^2) + \log_2(e) \cdot \frac{(x-\mu)^2}{2\sigma^2} \right] dx \\
&= \int_{\Re} f_Y(x) \log_2 \frac{1}{f_X(x)} dx \\
\Rightarrow h(X) - h(Y) &= \int_{\Re} f_X(x) \log_2 \frac{1}{f_X(x)} dx - \int_{\Re} f_Y(y) \log_2 \frac{1}{f_Y(y)} dy \\
&= \int_{\Re} f_Y(x) \log_2 \frac{1}{f_X(x)} dx - \int_{\Re} f_Y(x) \log_2 \frac{1}{f_Y(x)} dx \\
&= \int_{\Re} f_Y(x) \log_2 \frac{f_Y(x)}{f_X(x)} dx \\
&\geq \log_2(e) \int_{\Re} f_Y(x) \left[1 - \frac{f_X(x)}{f_Y(x)} \right] dx = 0.
\end{aligned}
$$

$$
\begin{aligned}
h(X) &= \int_a^b f_X(x) \log_2 \frac{1}{f_X(x)} dx \\
&= \int_a^b f_X(x) \log_2 |b-a| dx \\
&= \int_a^b f_Y(x) \log_2 |b-a| dx \\
&= \int_a^b f_Y(x) \log_2 \frac{1}{f_X(x)} dx
\end{aligned}
$$

We can similarly prove that among all distributions with support in $[a, b)$, the uniform distribution has the largest differential entropy.

$$
\begin{aligned}
\Rightarrow h(X) - h(Y) &= \int_a^b f_X(x) \log_2 \frac{1}{f_X(x)} dx - \int_a^b f_Y(y) \log_2 \frac{1}{f_Y(y)} dy \\
&= \int_a^b f_Y(x) \log_2 \frac{1}{f_X(x)} dx - \int_a^b f_Y(x) \log_2 \frac{1}{f_Y(x)} dx \\
&= \int_a^b f_Y(x) \log_2 \frac{f_Y(x)}{f_X(x)} dx \\
&\geq \log_2(e) \int_a^b f_Y(x) \left[1 - \frac{f_X(x)}{f_Y(x)} \right] dx = 0.
\end{aligned}
$$

Differential Entropy and Mutual Information for Continuous Ensembles

- Mutual information for continuous ensembles

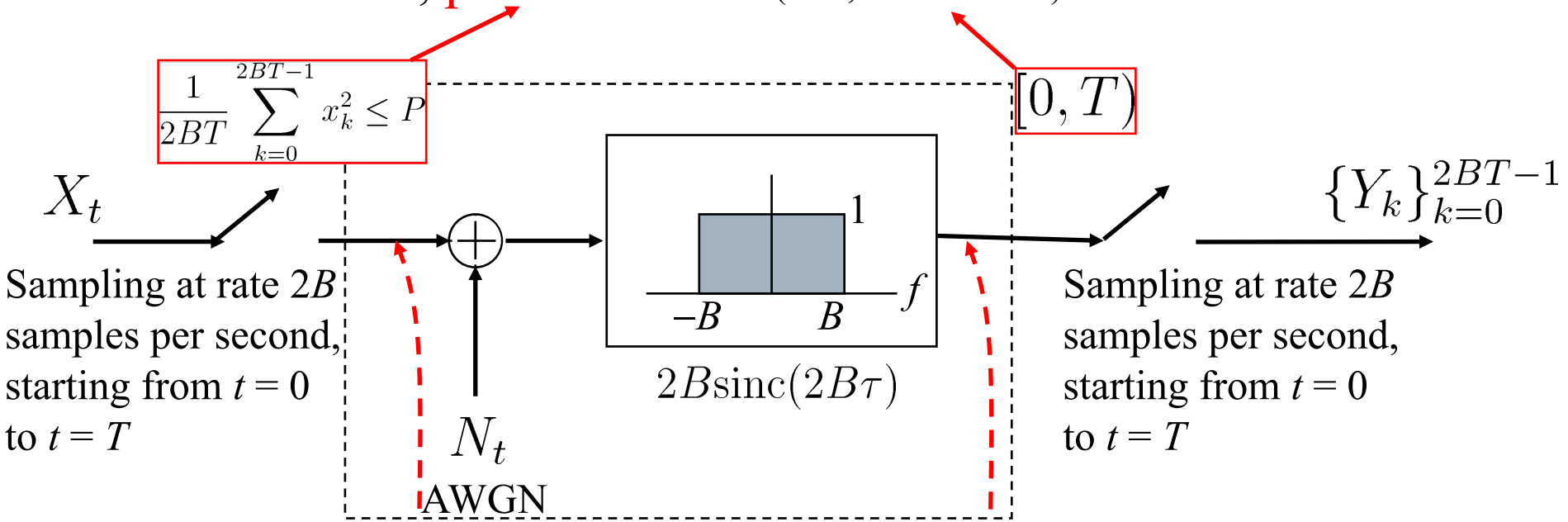
$$\begin{aligned} I(X^{\Delta};Y^{\Delta}) &= H(X^{\Delta}) - H(X^{\Delta}|Y^{\Delta}) \\ &\approx \left(h(X) + \log_2 \frac{1}{\Delta}\right) - \left(h(X|Y) + \log_2 \frac{1}{\Delta}\right) \\ &= h(X) - h(X|Y) \end{aligned}$$

Therefore,

$$I(X;Y) = \lim_{\Delta \downarrow 0} I(X^{\Delta};Y^{\Delta}) = h(X) - h(X|Y)$$

Information Transmission over Gaussian Channels

- Band-limited, power-limited (also, time-limited) Gaussian channels



$\Rightarrow Y_k = X_k + N_k$, where N_k zero-mean i.i.d. Gaussian with variance $N_0 B$

Information Transmission over Gaussian Channels

- What is the channel capacity (information bit per channel usage) for band-limited, power-limited (time-limited) Gaussian channels?

  Answer:

$$C = \max_{\{P_X : E[X^2] \leq P\}} I(X_k; Y_k)$$

- Optimization of mutual information

$$\begin{aligned} I(X_k; Y_k) &= h(Y_k) - h(Y_k|X_k) \\ &= h(Y_k) - h(N_k) \text{ (See the next slide.)} \end{aligned}$$

$$
\begin{aligned}
h(Y|X) &= \int_{\Re} f_X(x) h(Y|X=x) dx \\
&= \int_{\Re} f_X(x) \left(\int_{\Re} f_{Y|X}(y|x) \log_2 \frac{1}{f_{Y|X}(y|x)} dy \right) dx \\
&= \int_{\Re} f_X(x) \left(\int_{\Re} f_N(y-x) \log_2 \frac{1}{f_N(y-x)} dy \right) dx \\
&= \int_{\Re} f_X(x) \left(\int_{\Re} f_N(n) \log_2 \frac{1}{f_N(n)} dn \right) dx \\
&= h(N) = \frac{1}{2} \log_2 \left(2\pi e \sigma^2 \right) \\
\Rightarrow I(X_k; Y_k) &= h(Y_k) - h(N_k) = h(Y_k) - \frac{1}{2} \log_2 \left(2\pi e \sigma^2 \right) \\
&\leq \frac{1}{2} \log_2 \left(2\pi e \mathrm{Var}(Y_k) \right) - \frac{1}{2} \log_2 \left(2\pi e \sigma^2 \right).
\end{aligned}
$$

The upper bound can be achieved by making Y_k Gaussian (i.e., by taking X_k Gaussian with variance P)

Note that $E[X_k^2] \leq P$.

$$\Rightarrow \max_{\{X_k : E[X_k^2] \leq P\}} I(X_k; Y_k)$$

$$= \frac{1}{2}\log_2\left(2\pi e\left[P+\sigma^2\right]\right) - \frac{1}{2}\log_2\left(2\pi e\sigma^2\right)$$

$$= \boxed{\frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) \text{ bits per channel usage}}$$

$$= \frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) \frac{\text{bits}}{\text{channel usage}} \times \frac{1 \text{ channel usage}}{\frac{1}{2B} \text{ seconds}}$$

$$= \boxed{B\log_2\left(1+\frac{P}{N_0 B}\right) \text{ bits per second}}$$

Information Transmission over Gaussian Channels

- Sphere packing argument

$$Y_k = X_k + N_k, k = 1, 2, \cdots, n$$

where $\{X_k\}$ and $\{N_k\}$ are zero-mean i.i.d. respectively with variances P and σ^2.

With high probability

$$Y_1^2 + Y_2^2 + \cdots + Y_n^2 \approx n(P + \sigma^2)$$

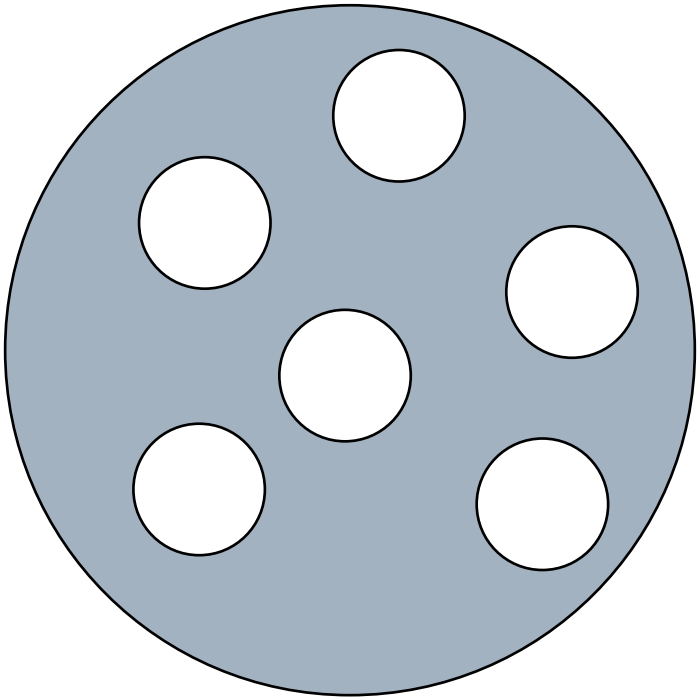
Although the receiver does not know the transmitted $\boldsymbol{X}$, it knows that with high probability, the transmitted one will be in the sphere

$$(Y_1 - X_1)^2 + (Y_2 - X_2)^2 + \cdots + (Y_n - X_n)^2 \approx n\sigma^2$$

Information Transmission over Gaussian Channels

- Hence, if the spheres centered at each (possibly) transmitted $\boldsymbol{X}$ with radius $(n\sigma^2)^{1/2}$ do not overlap with each other, the decoding error will be small.

$$Y_1^2 + \cdots + Y_n^2 \approx n(P + \sigma^2)$$



Information Transmission over Gaussian Channels

- Question: What the maximum number of spheres we can place inside the *Y*-space is ?

$$Y_1^2 + Y_2^2 + \cdots + Y_n^2 \approx n(P + \sigma^2)$$
$$\Rightarrow \text{volumn} = A_n([n(P + \sigma^2)]^{1/2})^n$$

$$(Y_1 - X_1)^2 + (Y_2 - X_2)^2 + \cdots + (Y_n - X_n)^2 \approx n\sigma^2$$
$$\Rightarrow \text{volume} = A_n([n\sigma^2]^{1/2})^n$$

$\Rightarrow$ The maximum number of codewords approximately equals

$$\frac{A_n[n(P + \sigma^2)]^{n/2}}{A_n[n\sigma^2]^{n/2}} = \left(1 + \frac{P}{\sigma^2}\right)^{n/2}$$

Information Transmission over Gaussian Channels

- Hence, the code rate is given by

$$\frac{1}{n}\log_2\left(1+\frac{P}{\sigma^2}\right)^{n/2} \text{ bits per channel usage}$$

$$=\frac{1}{2}\log_2\left(1+\frac{P}{\sigma^2}\right) \text{ bits per channel usage}$$

$$=C \text{ channel capacity}$$

Implications of the Information Capacity for Gaussian Channels

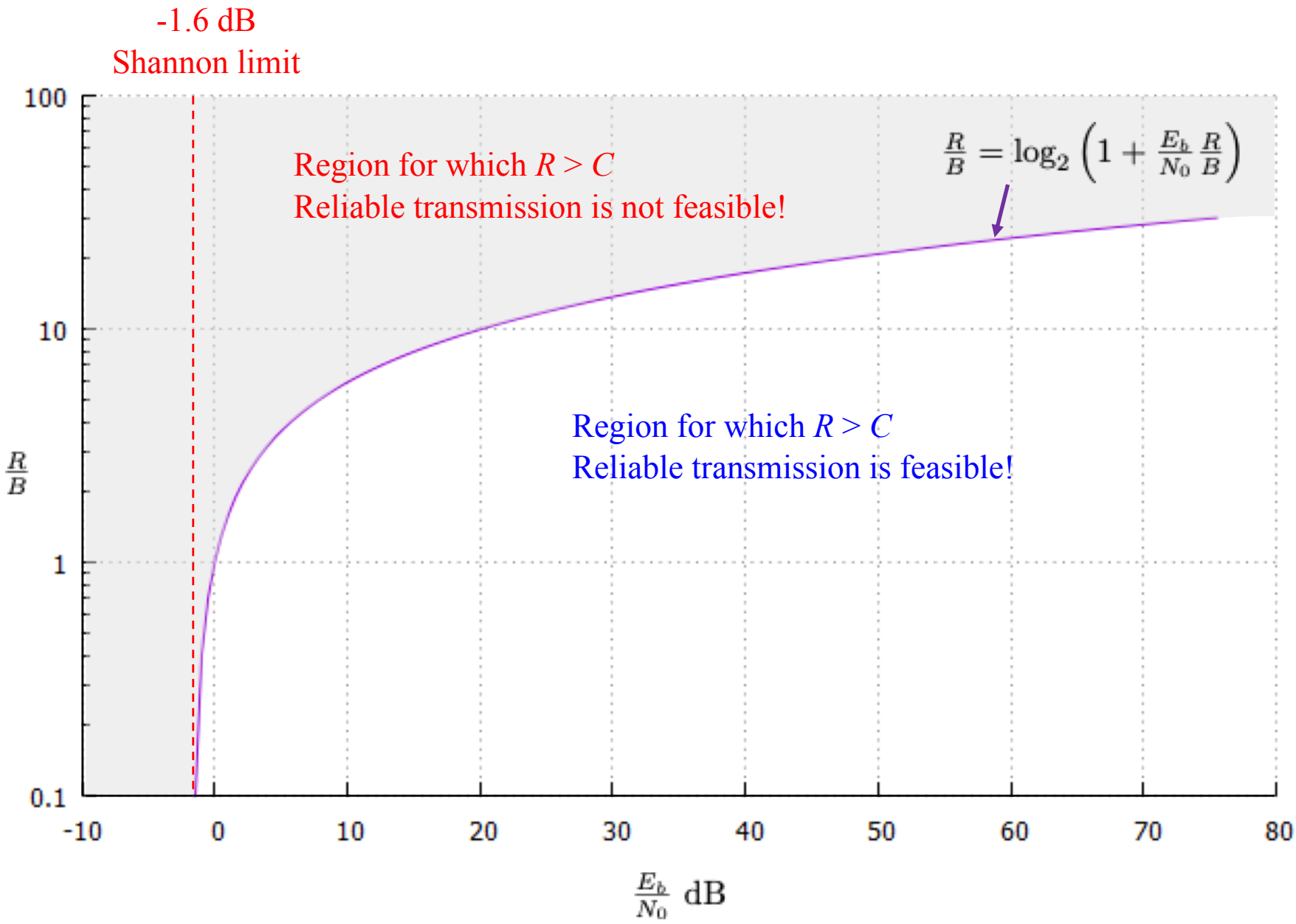
- In order to have arbitrarily small error, it requires:

$$R < B \log_2 \left(1 + \frac{P}{N_0 B}\right) \text{ bits per second}$$

- With $P = E_b R$, the above requirement is equivalent to:

$$\frac{R}{B} < \log_2 \left(1 + \frac{E_b}{N_0}\frac{R}{B}\right) \text{ bits per second per Hertz}$$

- We can then plot the relation between R/B (bits per second per Hertz) and E_b/N_0.


-1.6 dB
Shannon limit
Region for which $R > C$
Reliable transmission is not feasible!
$\frac{R}{B} = \log_2\left(1 + \frac{E_b}{N_0}\frac{R}{B}\right)$
Region for which $R > C$
Reliable transmission is feasible!
100
10
1
0.1
$\frac{R}{B}$
-10
0
10
20
30
40
50
60
70
80
$\frac{E_b}{N_0}$ dB

Implication 1:

$$\frac{E_b}{N_0} > \frac{2^{R/B}-1}{R/B} \text{ and } \lim_{R/B \downarrow 0} \frac{2^{R/B}-1}{R/B} = \log(2) = -1.6 \text{ dB}$$

implies E_b/N_0 must exceed -1.6 dB in order to make possible the arbitrarily small error transmission.

-1.6 dB is named the Shannon limit for an AWGN channel.

Implication 2:

The lower the E_b/N_0 (even if it exceeds -1.6 dB), the lower the bandwidth efficiency R/B for reliable transmission.

Implication 3*:*

$$C_\infty = \lim_{B\uparrow\infty} B\log_2\left(1+\frac{P}{N_0 B}\right) = \frac{P}{N_0}\log_2 e$$

implies that when bandwidth B is very large, the capacity is proportional to P/N_0.

Implications of the Information Capacity for Gaussian Channels

- Example: *M*-ary PCM
  - Examination of the relation between transmission rate, bandwidth and power.

In order for M-ary PCM to have a small error, the separation of amplitude levels are chosen to be $\pm(1/2)k\sigma$, $\pm(3/2)k\sigma, \cdots, \pm[(M-1)/2]k\sigma$, where $\sigma^2 = N_0 B$, and k is some chosen constant.

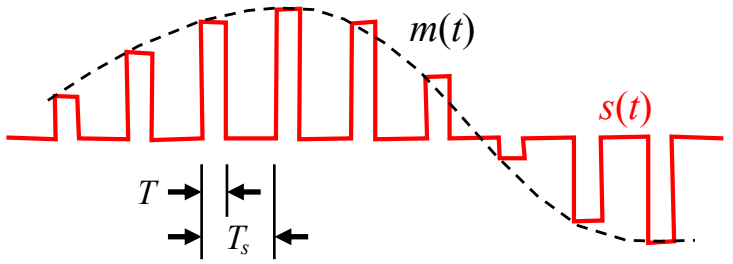
$$P = \frac{2}{M}\left[\left(\frac{1}{2}\right)^2 + \left(\frac{3}{2}\right)^2 + \cdots + \left(\frac{M-1}{2}\right)^2\right](k\sigma)^2 = k^2\sigma^2\left(\frac{M^2-1}{12}\right)$$

$$M = \left(1 + \frac{12P}{k^2\sigma^2}\right)^{1/2} = \left(1 + \frac{12P}{k^2 N_0 B}\right)^{1/2}$$

The transmission rate of n M-ary symbols (for n interleaving PCM users) with sampling rate $2W$ is given by

$$R = \frac{\log_2(M^n) \text{ bits}}{\frac{1}{2W} \text{ seconds}} = 2Wn\log_2(M) = Wn\log_2\left(1 + \frac{12P}{k^2 N_0 B}\right)$$

For n interleaving PCM users, $B = \kappa nW$, where often $1 \leq \kappa \leq 2$. In an ideal case, we take $\kappa = 1$.



$$R = B \log_2 \left(1 + \frac{\tilde{P}}{N_0 B} \right) \text{ bps, where } \tilde{P} = 12P/k^2.$$

This implies exactly the same relation between transmission rate, power and bandwidth as Shannon's capacity formula!

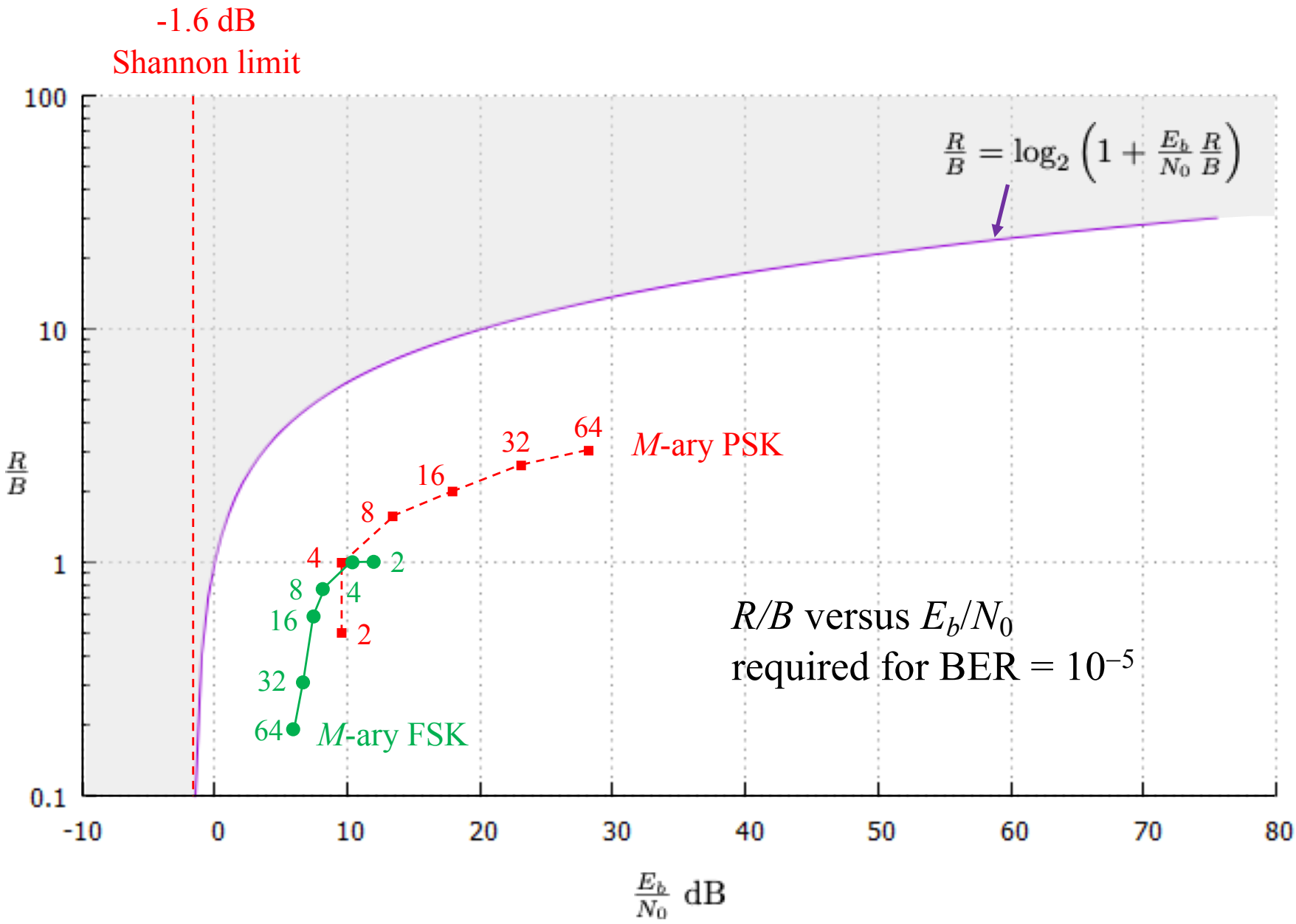
Implications of the Information Capacity for Gaussian Channels

- Comparison of bandwidth efficiency
  - Example: *M*-ary PSK and *M*-ary FSK
    - From Slide IDC1-65, the bandwidth efficiency of *M*-ary PSK satisfy:

$$\frac{R}{B} = \frac{\log_2(M)}{2}$$

    - From Slides IDC2-64, the bandwidth efficiency and error rate of *M*-ary FSK satisfy:

$$\frac{R}{B} = \frac{2\log_2(M)}{M}$$


-1.6 dB
Shannon limit
100
10
1
0.1
-10
0
10
20
30
40
50
60
70
80
$\frac{R}{B}$
$\frac{E_b}{N_0}$ dB
$\frac{R}{B} = \log_2\left(1 + \frac{E_b}{N_0}\frac{R}{B}\right)$
2
4
8
16
32
64
M-ary PSK
2
4
8
16
32
64
M-ary FSK
R/B versus E_b/N_0
required for BER = 10^{-5}

Implications of the Information Capacity for Gaussian Channels

- Remark
  - By increasing M, M-ary FSK approaches Shannon limit, while M-ary PSK deviates from Shannon limit.

Implications of the Information Capacity for Gaussian Channels

- Example: Capacity of binary-input AWGN channel

$$Y = X + N$$

where $X \in \{\pm 1\}$ and N zero-mean Gaussian with variance σ^2

$$\begin{aligned} C &= \max_{0 \le p \le 1} I(X;Y), \text{ where } \Pr[X = -1] = p \\ &= \frac{\log_2(e)}{\sigma^2} - \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-y^2/2} \log_2 \left[\frac{1}{\sigma^2} + y\sqrt{\frac{1}{\sigma^2}} \right] dy \end{aligned}$$

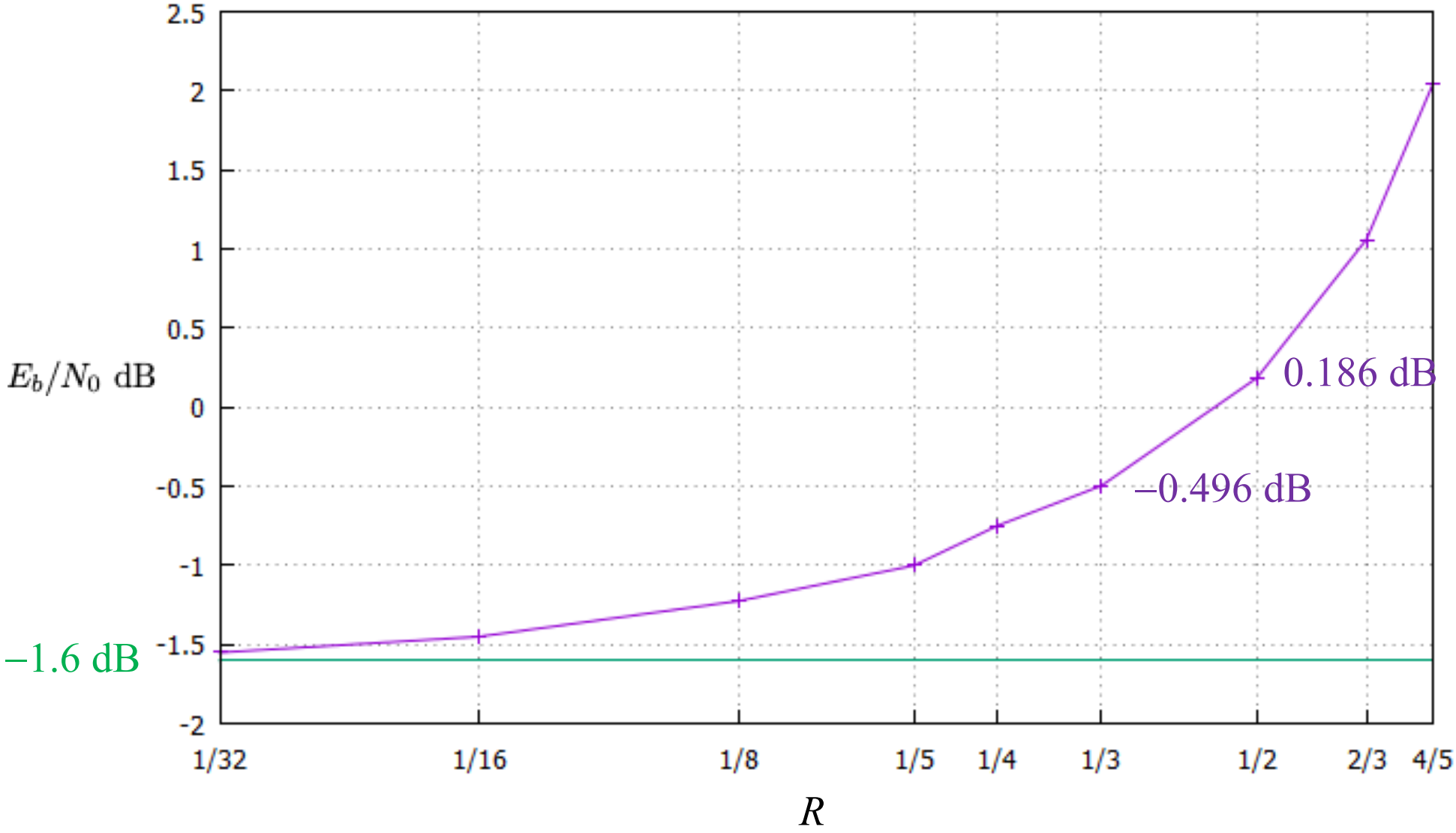
$$\frac{E_b}{N_0} = \frac{1 \text{ (Joule/channel usage) } n \text{ (channel usages) } /k \text{ (bits)}}{N_0}$$

$$= \frac{1}{N_0\left(\frac{k}{n}\right)} = \frac{1}{N_0 R} = \frac{1}{2\sigma^2 R}$$

$$\Rightarrow \sigma^2 = \frac{N_0}{2E_b R}$$

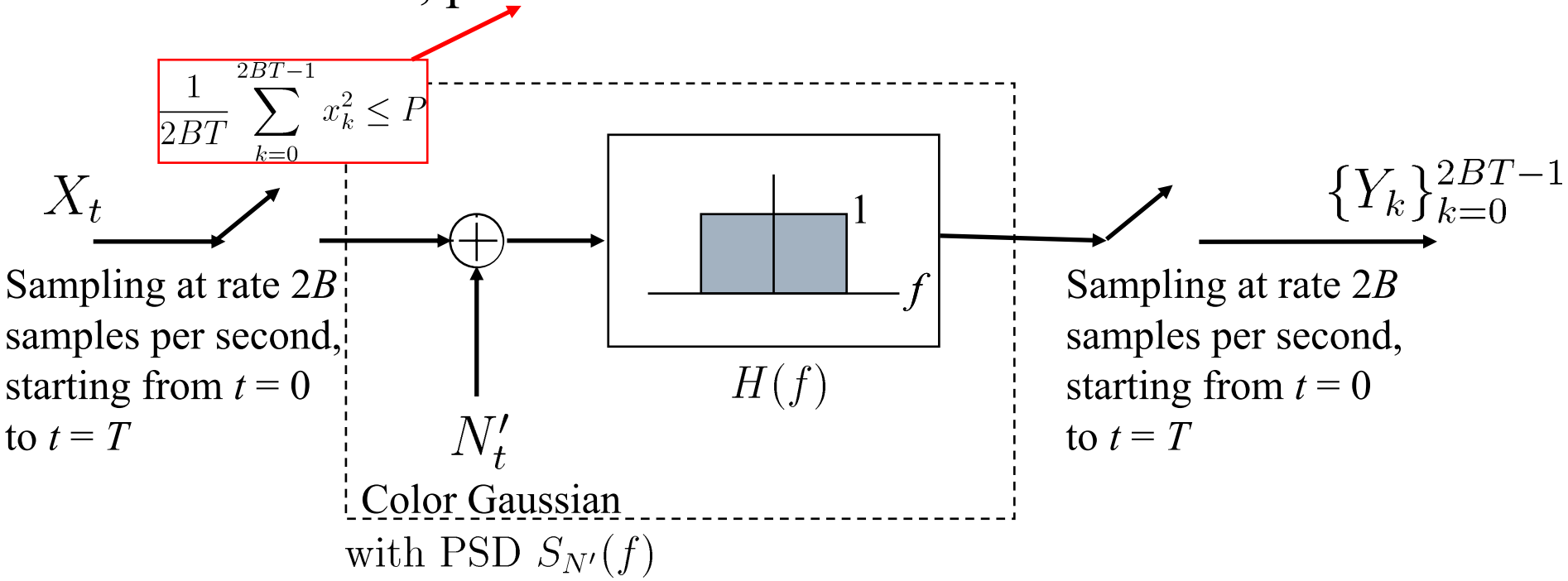
$$R < C = 2R\log_2(e)\frac{E_b}{N_0} - \frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-y^2/2}\log_2\left[2R\frac{E_b}{N_0} + y\sqrt{2R\frac{E_b}{N_0}}\right]dy$$

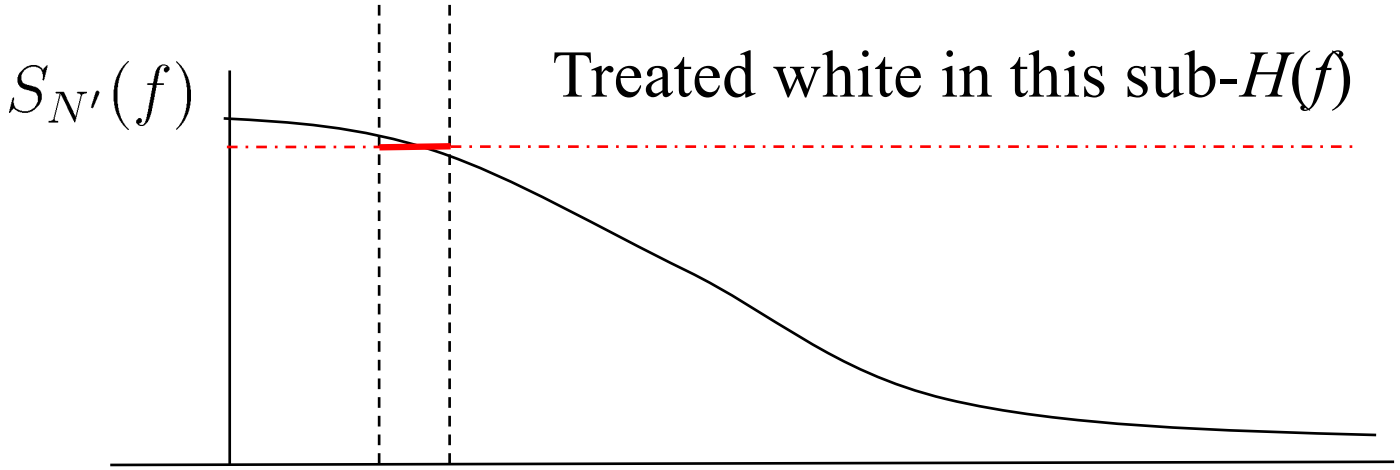
We then plot the minimum E_b/N_0 that is required to validate the above inequality for a given code rate R.


2.5
2
1.5
1
0.5
0
-0.5
-1
-1.5
-2
E_b/N_0 dB
−1.6 dB
0.186 dB
−0.496 dB
1/32
1/16
1/8
1/5
1/4
1/3
1/2
2/3
4/5
R

Information Capacity of Colored Noise Channel

- Band-limited, power-limited color Gaussian channels

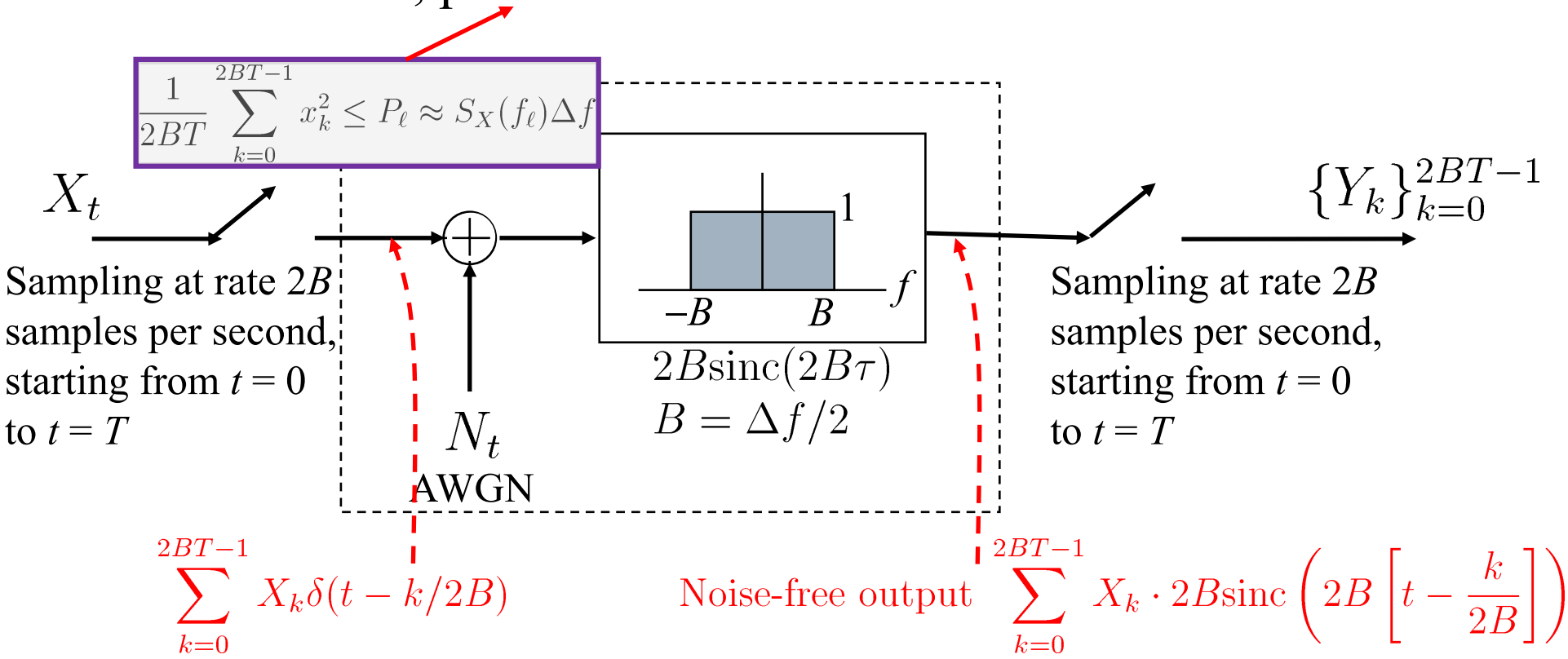




Assume that there are N sub-intervals.

Information Capacity of Colored Noise Channel

- Band-limited, power-limited Gaussian channels



$\Rightarrow Y_k = X_k + N_k$, where N_k zero-mean i.i.d. Gaussian with variance $N_0 B$

$$C_\ell \approx B \log_2 \left(1 + \frac{P_\ell}{\sigma_\ell^2}\right) \text{ bits per second}$$

$$= \left(\frac{\Delta f}{2}\right) \log_2 \left(1 + \frac{P_\ell}{\sigma_\ell^2}\right) \text{ for } 1 \leq \ell \leq N$$

where $\sigma_\ell^2 \approx S_{N'}(f_\ell)\Delta f$.

$$C = \sum_{\ell=1}^{N} C_\ell = \frac{1}{2} \sum_{\ell=1}^{N} \Delta f \log_2 \left(1 + \frac{P_\ell}{\sigma_\ell^2}\right)$$

where $\sum_{\ell=1}^{N} P_\ell = P$.

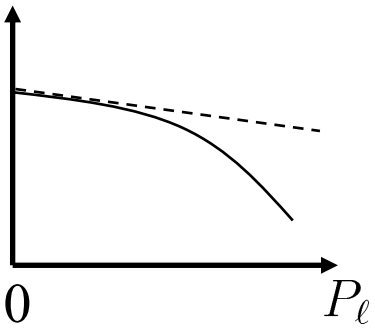
Question: How to maximize C subject to $\sum_{\ell=1}^{N} P_\ell = P$?

Lagrange multipliers technique

$$J = \sum_{\ell=1}^{N} C_\ell = \frac{1}{2}\sum_{\ell=1}^{N} \Delta f \log_2\left(1+\frac{P_\ell}{\sigma_\ell^2}\right) + \lambda\left(P - \sum_{\ell=1}^{N} P_\ell\right)$$

Deriving $\partial J/\partial P_\ell$ yields:

$$\begin{cases} \frac{1}{2\log(2)}\Delta f \frac{\frac{1}{\sigma_\ell^2}}{1+\frac{P_\ell}{\sigma_\ell^2}} - \lambda \le 0, \quad P_\ell^* = 0 \\ \frac{1}{2\log(2)}\Delta f \frac{\frac{1}{\sigma_\ell^2}}{1+\frac{P_\ell}{\sigma_\ell^2}} - \lambda = 0, \quad 0 < P_\ell^* < P \\ \frac{1}{2\log(2)}\Delta f \frac{\frac{1}{\sigma_\ell^2}}{1+\frac{P_\ell}{\sigma_\ell^2}} - \lambda \ge 0, \quad P_\ell^* = P \end{cases}$$



$$\begin{cases} P_\ell^* + \sigma_\ell^2 \geq \frac{\Delta f}{2\log(2)\lambda} = K\Delta f, & P_\ell^* = 0 \\ P_\ell^* + \sigma_\ell^2 = \frac{\Delta f}{2\log(2)\lambda} = K\Delta f, & P_\ell^* > 0 \end{cases}$$

where $K = \frac{1}{2\log(2)\lambda}$.

$$\begin{cases} S_X^*(f_\ell)\Delta f + S_{N'}(f_\ell)\Delta f \geq K\Delta f, & S_X^*(f_\ell)\Delta f = 0 \\ S_X^*(f_\ell)\Delta f + S_{N'}(f_\ell)\Delta f = K\Delta f, & S_X^*(f_\ell)\Delta f > 0 \end{cases}$$

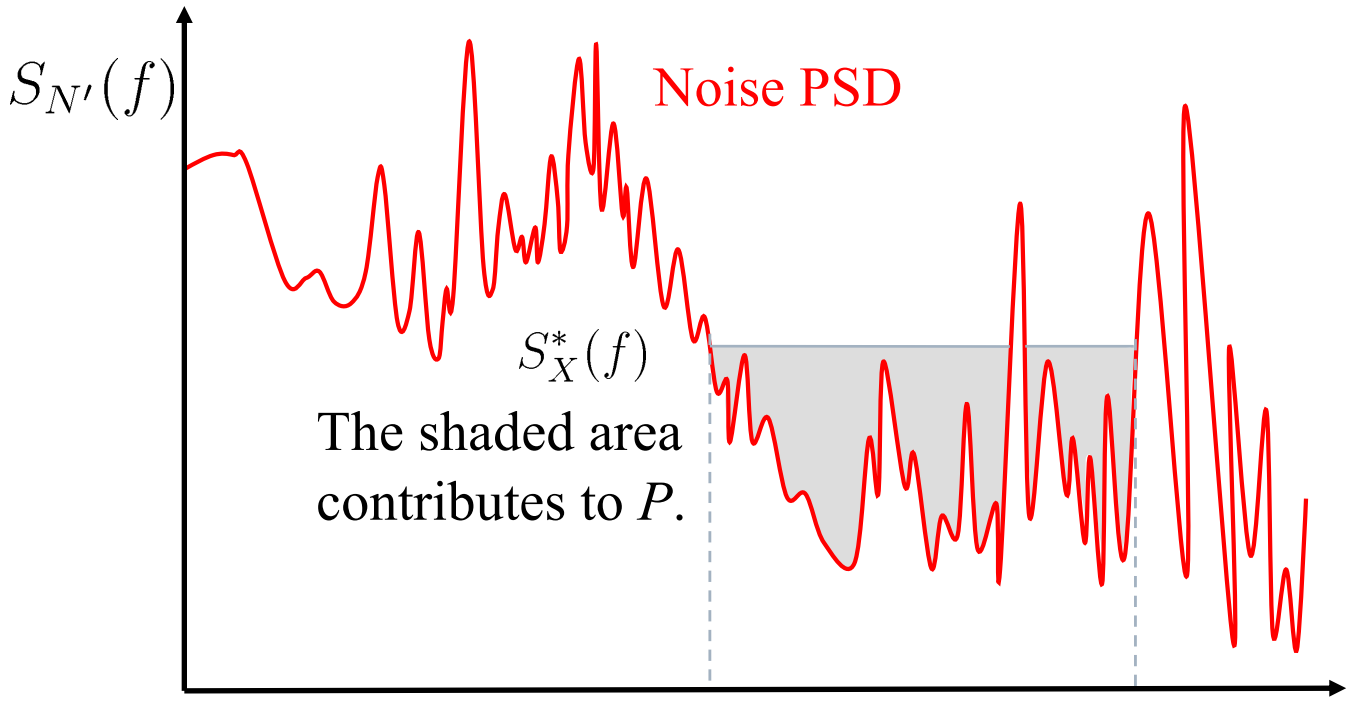
Equivalently,

$$\begin{cases} 0 \geq K - S_{N'}(f_\ell), & S_X^*(f_\ell) = 0 \\ S_X^*(f_\ell) = K - S_{N'}(f_\ell), & S_X^*(f_\ell) > 0 \end{cases}$$

$$\Rightarrow S_X^*(f) = \max\{K - S_{N'}(f), 0\}$$

Information Capacity of Colored Noise Channel

- Water-filling interpretation of information-capacity theorem for a colored noise channel

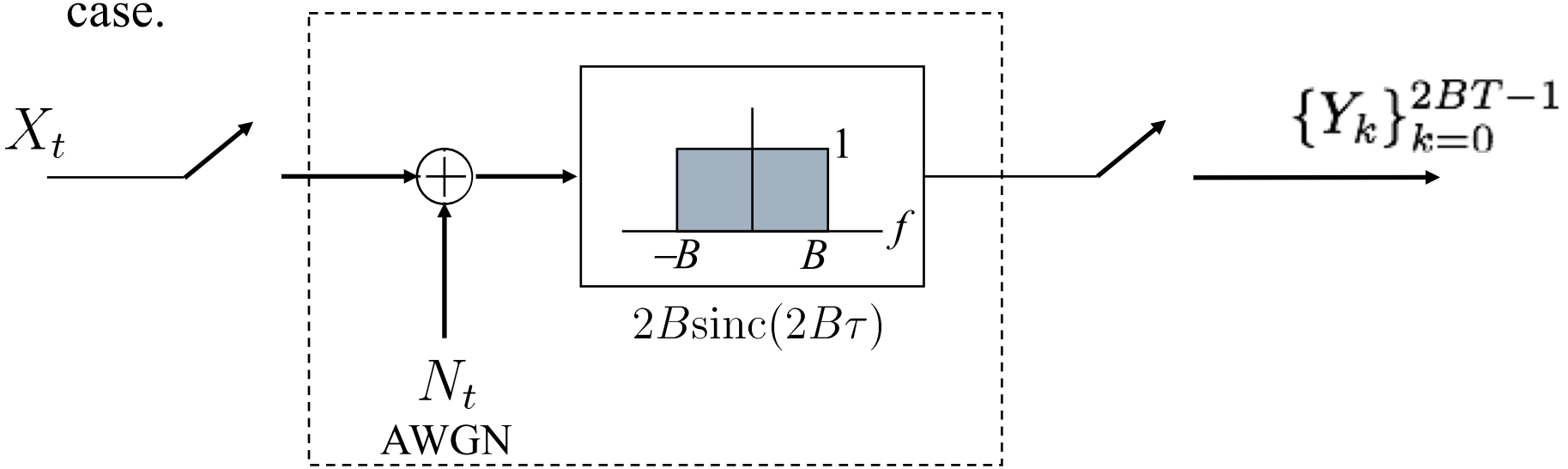


$$
\begin{aligned}
C &= \frac{1}{2}\sum_{\ell=1}^{N}\Delta f \log_2\left(1+\frac{P_\ell^*}{\sigma_\ell^2}\right) \\
&= \frac{1}{2}\sum_{\ell=1}^{N}\Delta f \log_2\left(1+\frac{S_X^*(f_\ell)\Delta f}{S_{N'}(f_\ell)\Delta f}\right) \\
&= \frac{1}{2}\sum_{\ell=1}^{N}\Delta f \log_2\left(1+\frac{S_X^*(f_\ell)}{S_{N'}(f_\ell)}\right) \\
&\approx \frac{1}{2}\int_{\mathcal{F}} \log_2\left(1+\frac{\max\{K-S_{N'}(f),0\}}{S_{N'}(f)}\right) df \\
&= \frac{1}{2}\int_{\mathcal{F}} \max\left\{\log_2\left(\frac{K}{S_{N'}(f)}\right),0\right\} df
\end{aligned}
$$

where K is chosen to satisfy $P=\int_{\mathcal{F}}\max\{K-S_{N'}(f),0\}df$, and $\mathcal{F}$ is the signal band.

Information Capacity of Colored Noise Channel

- This makes "Band-limited, power-limited Gaussian channels" a special case.



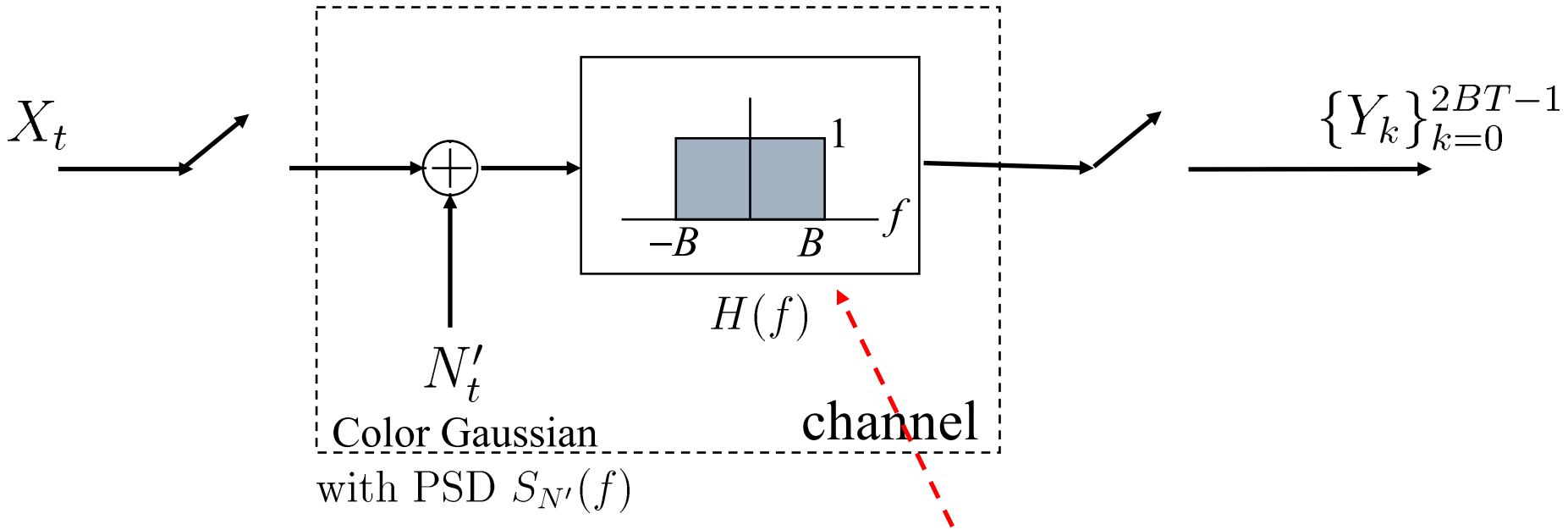
$$\Rightarrow S_X^*(f) = \max\{K - N_0/2, 0\}, |f| \leq B$$

$$\text{where } P = \int_{-B}^{B} S_X^*(f)df = 2B(K - N_0/2).$$

$$K = \frac{P}{2B} + \frac{N_0}{2}$$

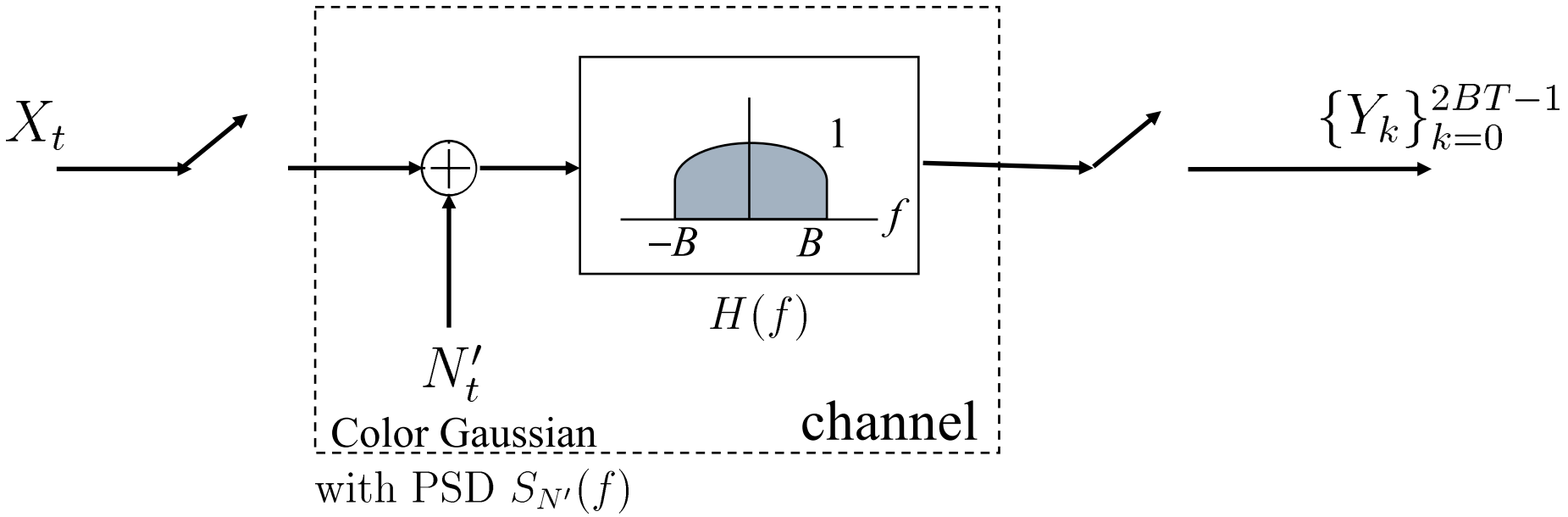
$$\begin{aligned} C &= \frac{1}{2}\int_{-B}^{B} \max\left\{\log_2\left(\frac{K}{S_{N'}(f)}\right), 0\right\} df \\ &= \frac{1}{2}\int_{-B}^{B} \max\left\{\log_2\left(\frac{K}{N_0/2}\right), 0\right\} df \\ &= \frac{1}{2}\int_{-B}^{B} \max\left\{\log_2\left(1+\frac{P}{N_0 B}\right), 0\right\} df \\ &= B\log_2\left(1+\frac{P}{N_0 B}\right) \end{aligned}$$

Just a reminder. My $S_{N'}(f)$ is exactly $S_N(f)/|H(f)|^2$ in text.



In the previous derivation, we assume that *H*(*f*) is an ideal lowpass filter with bandwidth *B*.

What if *H*(*f*) does not assume the shape of an ideal lowpass filter?
Will channel capacity be affected?

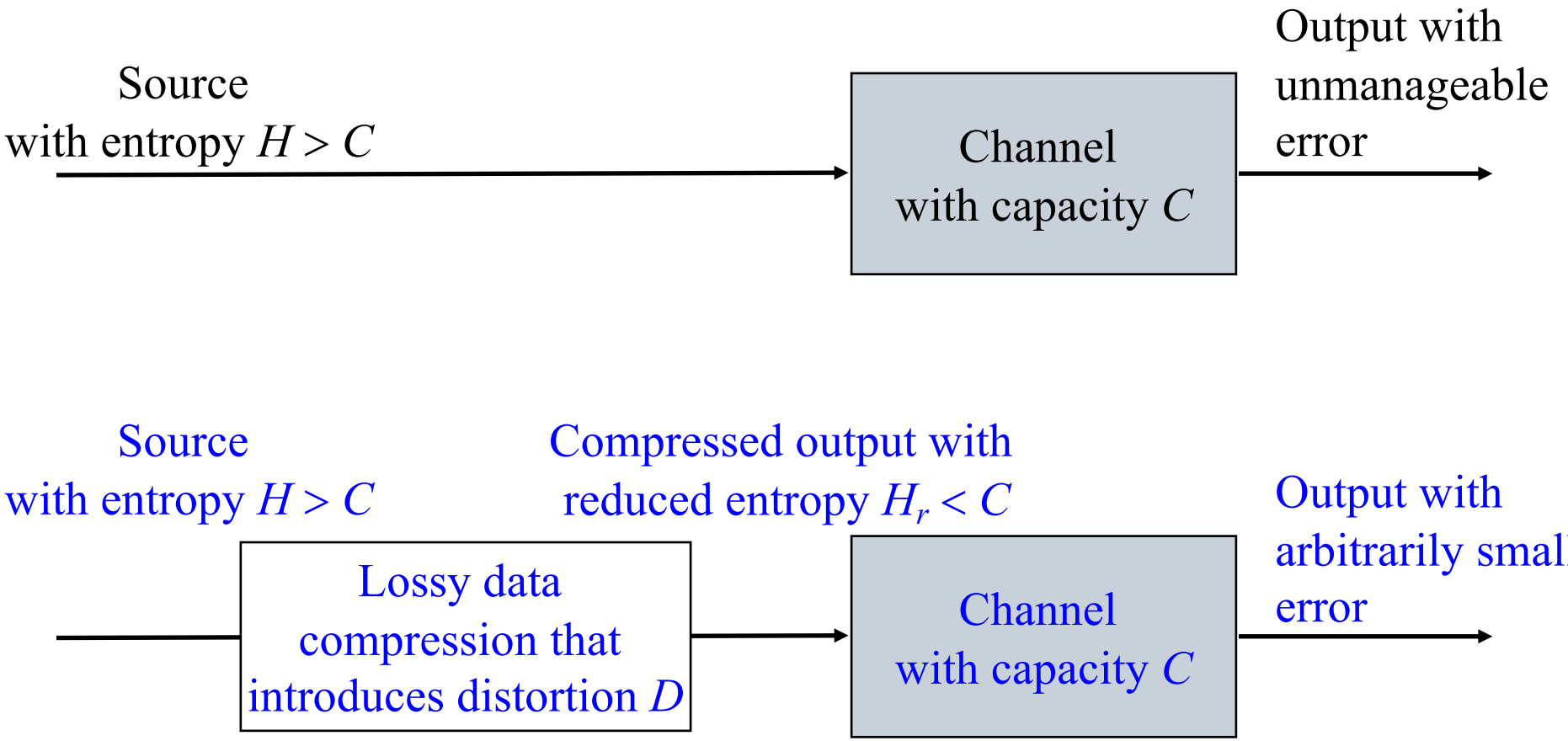


Answer: Since we assume that *H*(*f*) is known to the system, we can add an equalizer 1/*H*(*f*) at the receiver to compensate it.

Accordingly, under such perfectly known *H*(*f*) assumption, the non-ideal-shape of *H*(*f*) will not affect the channel capacity, namely, channel capacity is completely determined by the power spectrum of the color Gaussian.

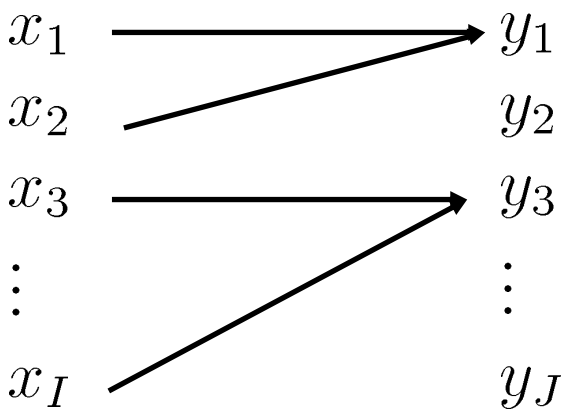
As a final remark, my $S_{N'}(f)$ is exactly $S_N(f)/|H(f)|^2$ in the textbook, in which the noise N is placed after the filter. Hence, equivalent $S_{N'}(f)$ will be affected by $H(f)$ after it is equivalently modeled to be placed before the filter. My comment above indicates that once $S_{N'}(f)$ is fixed, $H(f)$ becomes irrelevant to the capacity.

Rate Distortion Function



Rate Distortion Function

- Lossy data compression with a fidelity criterion



Fidelity criterion $d(x_i, y_j)$

Non-reversible function mapping f

$$\text{Average distortion} = \sum_{i=1}^{I} p(x_i) d(x_i, f(x_i))$$

$$\text{Define } p(y_j|x_i) = \begin{cases} 1, & y_j = f(x_i) \\ 0, & \text{otherwise} \end{cases}$$

$$\text{Average distortion} = \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i)p(y_j|x_i)d(x_i,y_j)$$

Question: What is the smallest data compression rate (number of output bits per input symbol) such that the average distortion is less than D?

Answer:

$$R(D) = \min_{\{p(y|x) \ : \ \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i)p(y_j|x_i)d(x_i,y_j) \le D\}} I(X;Y)$$

$$R(D) = \min_{\{p(y|x) \ : \ \sum_{i=1}^{I}\sum_{j=1}^{J} p(x_i)p(y_j|x_i)d(x_i,y_j)\le D\}} I(X;Y)$$

Example. Derive the rate-distortion function for binary memoryless source $X_1, X_2, \cdots$ with

$$\Pr[X=0] = 1 - \Pr[X=1] = p \in (0, 1/2)$$

$$\text{Reproduction alphabet } \mathcal{Y} = \{0, 1\}$$

$$\text{Hamming distortion } d(x,y) = \begin{cases} 1, & x \neq y \\ 0, & x = y \end{cases}$$

Solution.

$$\begin{aligned} R(D) &= \min_{\{p(y|x) \ : \ E[d(X,Y)]\le D\}} I(X;Y) \\ &= \min_{\{p(y|x) \ : \ E[d(X,Y)]= D\}} I(X;Y) \end{aligned}$$

$$
\begin{aligned}
R(D) &= \min_{\{p(y|x)\ :\ \Pr[X\neq Y]=D\}} I(X;Y) \\
&= \min_{\{p(y|x)\ :\ \Pr[X\neq Y]=D\}} \left[H(X) - H(X|Y)\right] \\
&= \min_{\{p(y|x)\ :\ \Pr[X\neq Y]=D\}} \left[H(X) - H(X\oplus Y|Y)\right] \\
&\geq \min_{\{p(y|x)\ :\ \Pr[X\neq Y]=D\}} \left[H(X) - H(X\oplus Y)\right] \\
&= H_b(p) - H_b(D)
\end{aligned}
$$

where

$$
H_b(z) = z\log_2\frac{1}{z} + (1-z)\log_2\frac{1}{1-z}.
$$

$$R(D) = \min_{\{p(y|x)\ :\ E[d(X,Y)]\leq D\}} I(X;Y) = \begin{cases} H_b(p) - H_b(D), & 0 \leq D < p \\ 0, & D \geq p \end{cases}$$

For $0 \leq D < p$, $R(D) = H(X) - H_b(D)$ can be achieved with $H(X|Y) = H_b(D)$, which can be satisfied by letting:

$$p_{X|Y}(1|0) = p_{X|Y}(0|1) = D.$$

For $D \geq p$, $R(D) = 0$ can be achieved with $H(X|Y) = H_b(p) = H(X)$, i.e., $Y \perp\!\!\!\perp X$ with marginal distribution satisfying

$$E[d(X,Y)] = \Pr[X=0]\Pr[Y=1] + \Pr[X=1]\Pr[Y=0]$$

$$= p \cdot \Pr[Y=1] + (1-p) \cdot \Pr[Y=0] \leq D.$$

- Remarks
  - $R(D)$ reduces to $H(X)$ when $D = 0$.
  - As D (the acceptable distortion level) increases, the data compression rate can be reduced to nearly zero.