Channel Coding with a Cost Constraint

I-Hsiang Wang

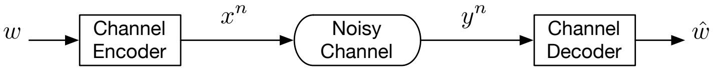
Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

November 12, 2024

1. Discrete Memoryless Channel with an Input Cost Constraint

2. Extension to Continuous Channels
   - Information measures for continuous distributions
   - Channel coding theorems for continuous memoryless channels



w → Channel Encoder → x^n → Noisy Channel → y^n → Channel Decoder → $\hat{w}$

We have studied the fundamental limit on the rate of reliable communication over DMC $\left(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y}\right)$.

The only constraint on the design of coding schemes is specified by the channel – *channel inputs must lie in* $\mathcal{X}$ and *channel outputs must lie in* $\mathcal{Y}$.

In practice, there might be additional constraints on certain **costs** such as power consumption in transmitting/receiving certain symbols.

Intuition: capacity becomes larger when the cost is less constrained.

Question (quantitative): What is the capacity under certain cost constraints?

Next, we take **average cost constraint** into account and derive the **capacity-cost function**.

Input cost and average cost constraints

We begin the treatment with input cost (not hard to extend the framework to incorporate output cost).

A non-negative **input cost function** $b : \mathcal{X} \to [0, \infty)$ is defined over the input alphabet $\mathcal{X}$ of the channel.

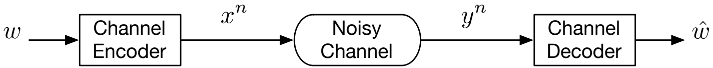
In this lecture, we focus on average cost constraints as follows:

$\forall\, n \in \mathbb{N}$, over n channel uses, the average input cost constraint B requires the coding scheme satisfy

$$\frac{1}{n} \sum_{i=1}^{n} b(x_i) \le \mathsf{B}.$$

Remark: WLOG one can shift $b\left(\cdot\right)$ such that there exists a symbol $x_o \in \mathcal{X}$ with $b\left(x_o\right) = 0$, and the constraint B is shifted by the same amount. Hence, we may assume the existence of such zero-cost symbol $x_o \in \mathcal{X}$.

Channel coding with input cost over DMC



1. A $(n, \lceil n\mathrm{R} \rceil, \mathsf{B})$ channel code consists of
   - $\mathrm{enc}_n : \mathcal{W} \to \mathcal{X}^n$ that maps each w to a length n codeword x^n, $k := \lceil n\mathrm{R} \rceil$.
     The codeword follows the input cost constraint $\frac{1}{n}\sum_{i=1}^{n} b(x_i) \le \mathsf{B}$.
   - $\mathrm{dec}_n : \mathcal{Y}^n \to \mathcal{W}$ that maps a channel output y^n to a reconstructed $\hat{w}$.

2. The error probability is defined as $\mathsf{P}_\mathsf{e}^{(n)} := \mathsf{Pr}\{W \neq \hat{W}\}$.

3. A rate R is said to be achievable **with input cost** B**,** $\mathsf{B} \ge 0$**,** if there exist a sequence of $(n, \lceil n\mathrm{R} \rceil, \mathsf{B})$ codes such that $\mathsf{P}_\mathsf{e}^{(n)} \to 0$ as $n \to \infty$.

   Channel capacity $\mathrm{C}(\mathsf{B}) := \sup\{\mathrm{R} \,|\, \mathrm{R} : \text{achievable}\}$ is a function of the cost constraint.

Coding theorem with average input cost constraint

Theorem 1 (Channel Coding for DMC with Average Input Cost Constraint)

The capacity of DMC $(\mathcal{X}, \mathsf{P}_{Y|X}, \mathcal{Y})$ *with average input cost constraint* B *is*

$$\mathrm{C}(\mathsf{B}) = \mathrm{C}^{\mathrm{I}}(\mathsf{B}) := \max_{\mathsf{P}_X : \, \mathsf{E}_{\mathsf{P}_X}[b(X)] \le \mathsf{B}} \mathrm{I}(X;Y) \,. \tag{1}$$

Remark: Compared to Theorem 4 of Lecture 3 (the channel coding theorem *without* input cost constraint), there is an additional constraint in the extremal problem (1) laid on the expected cost $\mathsf{E}[b(X)]$.

Before we prove this theorem using standard arguments (Converse: Fano, data processing, single letterization; Achievability: random coding, typicality), let us first discuss some functional properties of $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$.

These properties will be useful in proving the converse and the achievability.

Properties of the capacity-cost function $\mathrm{C}^{\mathrm{I}}(\mathrm{B})$

Proposition 1

1. $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ *is non-decreasing in* B*.*
2. $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ *is concave in* B*.*
3. $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ *is continuous in* B*.*

pf: $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ is non-decreasing since (1) becomes more constrained as $\mathsf{B} \searrow$.

To show that $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ is concave, we'd like to show for any $\lambda \in [0,1]$,

$$\lambda \mathrm{C}^{\mathrm{I}}(\mathsf{B}_1) + (1-\lambda)\mathrm{C}^{\mathrm{I}}(\mathsf{B}_2) \le \mathrm{C}^{\mathrm{I}}(\lambda \mathsf{B}_1 + (1-\lambda)\mathsf{B}_2).$$

To prove this, let

$$\mathsf{P}_i := \operatorname*{arg\,max}_{\mathsf{P}_X:\ \mathsf{E}[b(X)] \le \mathsf{B}_i} \mathrm{I}(X;Y)$$

be the capacity-achieving distribution under cost constraint B_i, for $i = 1, 2$.

Define $\mathsf{P}_\lambda := \lambda \mathsf{P}_1 + (1-\lambda)\mathsf{P}_2$. Let $X_i \sim \mathsf{P}_i$ for $i = 1, 2$ and $X_\lambda \sim \mathsf{P}_\lambda$ be the mixture of X_1 and X_2.

By the fact that $\mathrm{I}(X;Y)$ is concave in P_X for a fixed $\mathsf{P}_{Y|X}$, we have

$$\lambda \mathrm{C}^{\mathrm{I}}(\mathsf{B}_1) + (1-\lambda)\mathrm{C}^{\mathrm{I}}(\mathsf{B}_2) = \lambda \mathrm{I}(X_1;Y) + (1-\lambda)\mathrm{I}(X_2;Y) \le \mathrm{I}(X_\lambda;Y).$$

Note that $\mathsf{E}[b(X_\lambda)] = \lambda \mathsf{E}[b(X_1)] + (1-\lambda)\mathsf{E}[b(X_2)] \le \lambda \mathsf{B}_1 + (1-\lambda)\mathsf{B}_2$.

Since the average cost of X_λ is $\le \lambda \mathsf{B}_1 + (1-\lambda)\mathsf{B}_2$, by the definition of $\mathrm{C}^{\mathrm{I}}(\cdot)$, we have

$$\mathrm{I}(X_\lambda;Y) \le \mathrm{C}^{\mathrm{I}}(\lambda \mathsf{B}_1 + (1-\lambda)\mathsf{B}_2).$$

Finally, since a convex/concave function on an open interval is continuous on that open interval, the continuity of $\mathrm{C}^{\mathrm{I}}(\cdot)$ in B, $\mathsf{B} \in (0, \infty)$ is established. □

Note: for right-continuity at $\mathsf{B} = 0$, some additional slight efforts are needed.

Converse proof of Theorem 1

Following the converse proof of DMC without input cost, we arrive at

$$\mathrm{R} - \varepsilon_n \le \frac{1}{n}\sum_{i=1}^{n} \mathrm{I}\left(X_i;Y_i\right),$$

where $\varepsilon_n \to 0$ as $n \to \infty$. From the capacity formula (1), it can be seen that

$$\forall\, i \in \{1,...,n\},\ \mathrm{I}(X_i;Y_i) \le \mathrm{C}^{\mathrm{I}}(\mathsf{E}[b(X_i)])\,. \quad \text{(let } \mathsf{B}_i := \mathsf{E}\left[b\left(X_i\right)\right]\text{)}$$

Since $\frac{1}{n}\sum_{i=1}^{n} b(x_i) \le \mathsf{B}$ for all x^n, we have $\frac{1}{n}\sum_{i=1}^{n} \mathsf{B}_i \le \mathsf{B}$. As a result,

$$\mathrm{R} - \varepsilon_n \le \frac{1}{n}\sum_{i=1}^{n} \mathrm{C}^{\mathrm{I}}(\mathsf{B}_i) \overset{(a)}{\le} \mathrm{C}^{\mathrm{I}}\Big(\frac{1}{n}\sum_{i=1}^{n} \mathsf{B}_i\Big) \overset{(b)}{\le} \mathrm{C}^{\mathrm{I}}(\mathsf{B}).$$

- (a) is due to concavity of $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ in B.
- (b) is due to $\frac{1}{n}\sum_{i=1}^{n} \mathsf{B}_i \le \mathsf{B}$, and $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ is non-decreasing.

Therefore, if R is achievable, it must be the case that $\mathrm{R} \le \mathrm{C}^{\mathrm{I}}(\mathsf{B})$. □

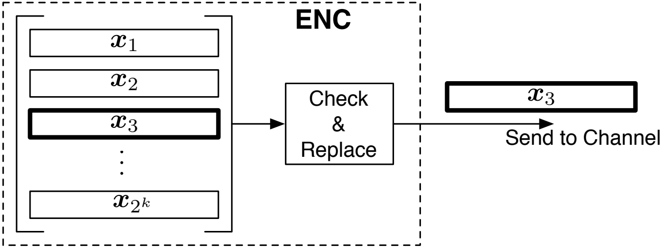
Ideas of the Achievability

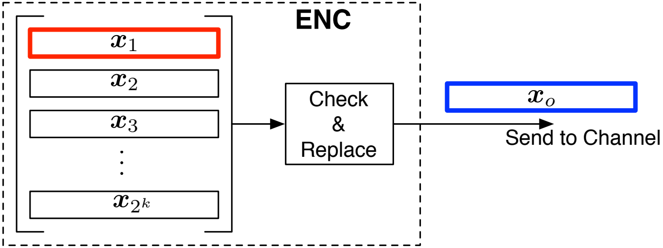
Achievability proof mostly follows that of DMC *without* input cost constraints.

However, there is an issue:

In random codebook generation, how to ensure all codewords satisfy the cost constraint?

Idea: If a generated codeword violates the input cost constraint, then we replace it by the zero-cost codeword $\boldsymbol{x}_o := \begin{bmatrix} x_o & \cdots & x_o \end{bmatrix}$ when we send it.


ENC
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_3$
$\boldsymbol{x}_{2^k}$
Check & Replace
$\boldsymbol{x}_3$
Send to Channel


ENC
$\boldsymbol{x}_1$
$\boldsymbol{x}_2$
$\boldsymbol{x}_3$
$\boldsymbol{x}_{2^k}$
Check & Replace
$\boldsymbol{x}_o$
Send to Channel

So, why not form an equivalent channel and absorb the **“check-and-replace”** procedure into it?

It is then tempting to apply the analysis and bounding techniques used in DMC without input cost constraints on this equivalent channel, which contains the **“check-and-replace”** procedure.

However, this procedure involves computing $\frac{1}{n}\sum_{i=1}^{n} b(x_i)$, which is **not symbol-by-symbol**, and introduce **memory** into the equivalent channel.

The previous analysis and bounding techniques for **memoryless** channels cannot be used.

To circumvent the issue, we shall modify the proposed scheme into another one with worse error probability, using the following steps:

- Do not “replace” when a codeword violates the cost constraint.
- Instead, send it directly and take care of the violation **at the decoder**.
- If the decoded codeword violates the constraint, it declares an error.
- This way, if a violation happens, an error always occurs under the modified scheme, while in the original scheme, there is a slight chance that an error does not occur.

In the following we focus on the modified scheme where the violation of cost constraints is taken care by the decoder, which introduces another kind of error event.

Note: Once we show the existence of a sequence of modified schemes with vanishing error probability, we can add the “check-and-replace” procedure back and then obtain a valid scheme.

Achievability proof of Theorem 1

Achievability proof mostly follows that of DMC without input cost constraints.

Keep in mind that if a chosen codeword violates the cost constraint, it results in a decoding error.

How to control the probability of such violations? **Typicality**.

Lemma 1 (Typical Average Lemma)

For any nonnegative function $g(x)$ *on* $\mathcal{X}$*, if* $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathsf{P})$*, then*

$$(1-\varepsilon)\mathsf{E}_{\mathsf{P}}[g(X)] \le \tfrac{1}{n}\textstyle\sum_{i=1}^{n} g(x_i) \le (1+\varepsilon)\mathsf{E}_{\mathsf{P}}[g(X)].$$

The proof is straightforward due to the definition of typical sequences and left as exercise.

Hence, we should generate random codewords with a slightly smaller average cost $\frac{\mathsf{B}}{1+\varepsilon}$.

Random Codebook Generation:

Generate the random codebook $\mathbf{C}$ with i.i.d. entries according to

$$\mathsf{P}_X = \underset{\mathsf{P}:\ \mathsf{E}_{\mathsf{P}}[b(X)] \le \frac{\mathsf{B}}{1+\varepsilon}}{\arg\max} \mathrm{I}(X;Y).$$

Observe that:

- If $x^n \in \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_X)$, it satisfies the cost constraint due to Lemma 1:

$$\tfrac{1}{n}\sum_{i=1}^{n} b(x_i) \le (1+\varepsilon)\mathsf{E}[b(X)] = (1+\varepsilon)\tfrac{\mathsf{B}}{1+\varepsilon} = \mathsf{B}.$$

- If the generated $x^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_X)$, it may violate the constraint.

  Nevertheless, the probability that this happens vanishes as $n \to \infty$, so the (relaxed) decoder can declare an error whenever the decoded codeword $x^n \notin \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_X)$.

Error Probability Analysis:

Following the same procedure in the error probability analysis for DMC without input cost constraints, we arrive at upper bounding

$$\mathsf{Pr}\left\{\mathcal{E} \,|\, W = 1\right\} := \mathsf{P}_1(\mathcal{E}).$$

The error event now can be decomposed as

$$\mathcal{E} = \mathcal{E}_0 \cup \mathcal{A}_1^c \cup \left(\cup_{w \neq 1} \mathcal{A}_w\right),$$

where $\mathcal{A}_w := \{(X^n(w), Y^n) \in \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_{X,Y})\}$, and $\mathcal{E}_0 := \{X^n(1) \notin \mathcal{T}_\varepsilon^{(n)}(\mathsf{P}_X)\}$.

Upper bounding $\mathsf{P}_1\left(\mathcal{A}_1^c\right)$ and $\mathsf{P}_1\left(\mathcal{A}_w\right)$ for $w \neq 1$ remains the same.

No need to worry about $\mathsf{P}_1(\mathcal{E}_0)$ because $\mathcal{E}_0 \subseteq \mathcal{A}_1^c$.

Hence, following the error probability analysis in our achievability proof of the noisy channel coding theorem without cost constraints, we conclude that for any $\mathrm{R} < \mathrm{C}^{\mathrm{I}}(\mathsf{B}/(1+\varepsilon))$, R is achievable.

Since $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ is continuous in B, we make $\mathrm{C}^{\mathrm{I}}(\mathsf{B}/(1+\varepsilon))$ arbitrarily close to $\mathrm{C}^{\mathrm{I}}(\mathsf{B})$ from below, and conclude that for all $\mathrm{R} < \mathrm{C}^{\mathrm{I}}(\mathsf{B})$, R is achievable. □

1 Discrete Memoryless Channel with an Input Cost Constraint

2 Extension to Continuous Channels

- Information measures for continuous distributions
- Channel coding theorems for continuous memoryless channels

Entropy of a continuous random variable

Question: What is the entropy of a continuous real-valued r.v. X ?

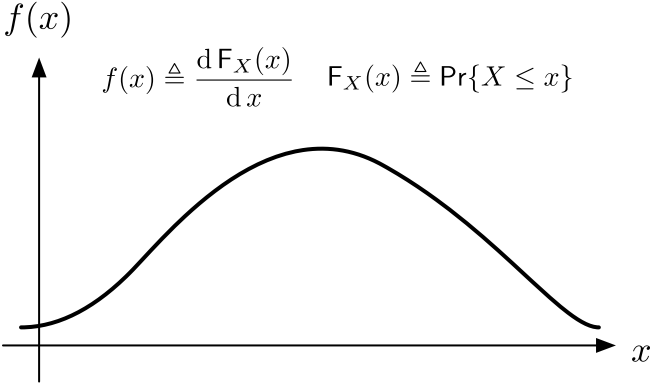
Suppose X has the probability density function (PDF) $\mathsf{f}_X(\cdot)$.

Let us discretize X to answer this question, as follows:

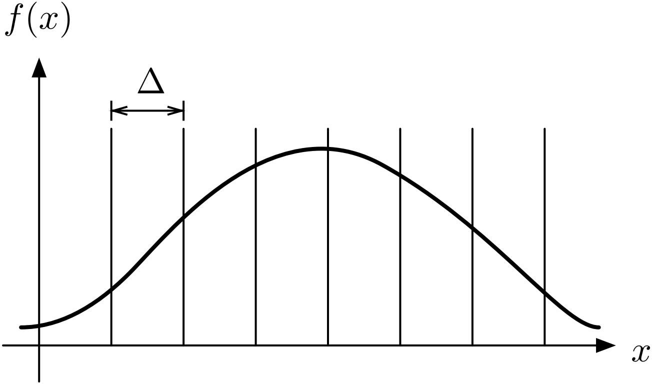
- Partition $\mathbb{R}$ into length-Δ intervals: $\mathbb{R} = \bigcup_{k=-\infty}^{\infty} [k\Delta, (k+1)\Delta)$.
- Suppose that $\mathsf{f}_X(\cdot) \equiv f(\cdot)$ is continuous, then by the mean-value theorem (MVT),

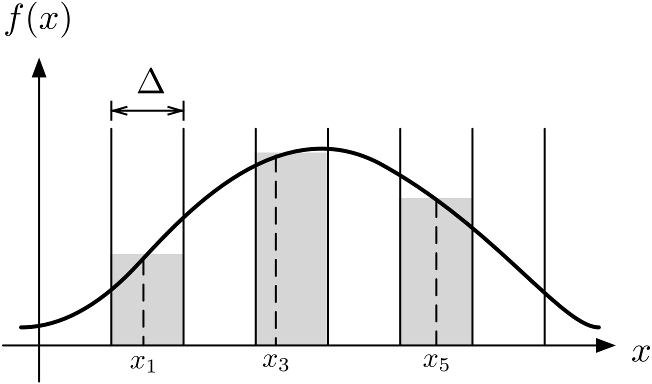
$$\forall\, k \in \mathbb{Z},\ \exists\, x_k \in [k\Delta, (k+1)\Delta) \text{ such that } f(x_k) = \frac{1}{\Delta} \int_{k\Delta}^{(k+1)\Delta} f(x)\, \mathrm{d}x.$$

- Set $[X]_\Delta := x_k$ if $X \in [k\Delta, (k+1)\Delta)$, with PMF $\mathsf{P}(x_k) = f(x_k)\Delta$.



$$f(x) \triangleq \frac{\mathrm{d}\,\mathsf{F}_X(x)}{\mathrm{d}\,x} \qquad \mathsf{F}_X(x) \triangleq \mathsf{Pr}\{X \le x\}$$


$f(x)$
Δ
x


$f(x)$
Δ
x_1
x_3
x_5
x

Observation: $\lim_{\Delta \to 0} \mathrm{H}([X]_\Delta) = \mathrm{H}(X)$ (intuitively), while

$$\begin{aligned}
\mathrm{H}([X]_\Delta) &= -\sum_{k=-\infty}^{\infty} (f(x_k)\Delta) \log (f(x_k)\Delta) \\
&= -\Delta \sum_{k=-\infty}^{\infty} f(x_k) \log f(x_k) - \log \Delta \\
&\to -\int_{\infty}^{\infty} f(x) \log f(x) \, \mathrm{d}x + \infty = \infty \qquad \text{as } \Delta \to 0
\end{aligned}$$

Hence, $\mathrm{H}(X) = \infty$ if $-\int_{\infty}^{\infty} f(x) \log f(x) \, \mathrm{d}x = \mathsf{E}\left[\log \frac{1}{f(X)}\right]$ exists.

It is quite intuitive that the entropy of a continuous random variable can be arbitrarily large, because it can take infinitely many possible values.

The term $\mathsf{E}_{X \sim \mathsf{f}_X}[\log 1/\mathsf{f}_X(X)]$ shares a similar form as the Shannon entropy $\mathsf{E}_{X \sim \mathsf{p}_X}[\log 1/\mathsf{p}_X(X)]$, and it is called the *differential entropy* of the density f_X (or the continuous random variable X).

Differential entropy

Definition 1 (Differential Entropy)

The differential entropy of a continuous r.v. X with PDF f_X is defined as

$$\mathrm{h}(X) := \mathsf{E}_{X\sim \mathsf{f}_X}\left[\log \frac{1}{\mathsf{f}_X(X)}\right]$$

if the (improper) integral exists.

The previous quantization argument implies that the Shannon entropy of the quantized version of the continuous r.v. X to n-bit precision ($\Delta = 2^{-n}$) is roughly $n + \mathrm{h}(X)$ bits.

In words, $\mathrm{h}(X)$ is the *extra* number of bits on the average required to describe X to n-bit precision. It can be positive, zero, or negative.

Example 1 (Differential entropy of a uniform r.v.)

For a r.v. $X \sim \mathrm{Unif}([a,b])$, that is, its PDF $\mathsf{f}_X(x) = \frac{1}{b-a}\mathbb{1}\{a \le x \le b\}$, its differential entropy

$$\mathrm{h}(X) = \log(b-a).$$

Interpretation via quantization:

- Suppose $X \sim \mathrm{Unif}([0,1])$. $\mathrm{h}(X) = 0$. To describe X to n-bit precision, n bits are needed.
- Suppose $X \sim \mathrm{Unif}([0,2^{-m}])$, $m > 0$. $\mathrm{h}(X) = -m$. To describe X to n-bit precision, just $n-m$ bits are needed.
- Suppose $X \sim \mathrm{Unif}([0,2^{m}])$, $m > 0$. $\mathrm{h}(X) = m$. To describe X to n-bit precision, an extra m bits are needed.

Remark: if $b-a$ approaches 0, X becomes deterministic, and its differential entropy becomes $-\infty$.

Example 2 (Differential entropy of $\mathrm{N}(0,1)$)

For a r.v. $X \sim \mathrm{N}(0,1)$, that is, its PDF $\mathsf{f}_X(x) = \frac{1}{\sqrt{2\pi}} \exp(-\frac{x^2}{2})$, its differential entropy

$$\mathrm{h}(X) = \frac{1}{2} \log(2\pi e).$$

Conditional differential entropy

Suppose X is jointly distributed with another random variable Y, and given $Y = y$, X is still a continuous r.v. with density $\mathsf{f}_{X|Y}(\cdot|y)$. The differential entropy of this density is

$$\mathrm{h}(X|Y=y) = \int_{-\infty}^{\infty} \mathsf{f}_{X|Y}(x|y) \log \frac{1}{\mathsf{f}_{X|Y}(x|y)}\ \mathrm{d}x.$$

Averaging it over Y, the conditional differential entropy emerges:

$$\begin{aligned}\mathrm{h}(X|Y) &= \begin{cases} \displaystyle\sum_{y\in\mathcal{Y}} \mathsf{p}_Y(y)\mathrm{h}(X|Y=y)\,, & \text{if } Y \text{ is discrete with PMF } \mathsf{p}_Y. \\ \displaystyle\int_{-\infty}^{\infty} \mathsf{f}_Y(y)\mathrm{h}(X|Y=y)\ \mathrm{d}y, & \text{if } Y \text{ is continuous with PDF } \mathsf{f}_Y. \end{cases} \\ &= \mathsf{E}_{X,Y}\left[\log \frac{1}{\mathsf{f}_{X|Y}(X|Y)}\right].\end{aligned}$$

Information divergence

Recall that the information divergence of a distribution with density $\mathsf{f}_1(\cdot)$ from another distribution with density $\mathsf{f}_2(\cdot)$ is just

$$\mathrm{D}\left(\mathsf{f}_1 \| \mathsf{f}_2\right) := \mathsf{E}_{X \sim \mathsf{f}_1}\left[\log \frac{\mathsf{f}_1(X)}{\mathsf{f}_2(X)}\right] = \int_{x \in \mathsf{supp}_{\mathsf{f}_1}} \mathsf{f}_1(x) \log \frac{\mathsf{f}_1(x)}{\mathsf{f}_2(x)}\, \mathrm{d}x$$

if the (improper) integral exists. Other properties remain to hold such as non-negativity, DPI, chain rule, conditioning increases divergence, etc..

Differential entropy turns out to have a strong connection to KL divergence. In $\mathrm{D}\left(\mathsf{f} \| \mathrm{g}\right)$, suppose we replace the second PDF g by the “uniform” density $\mathbb{1}\left\{x \in \mathbb{R}\right\}$, that is, instead of a probability measure, we use the *Lebesgue measure*. Then,

$$\mathrm{D}\left(\mathsf{f} \| \mathsf{g}\right) = \int_{-\infty}^{\infty} \mathsf{f}(x) \log \mathsf{f}(x)\ \mathrm{d}x = -\mathrm{h}\left(\mathsf{f}\right).$$

Hence, for a probability law that contains a discrete component, its differential entropy is $-\infty$.

Mutual information

The mutual information between two jointly distributed r.v.'s X and Y can be defined via information divergence when they are not discrete:

- If (X, Y) is jointly distributed with PDF $\mathsf{f}_{X,Y}$, then

$$\begin{aligned} \mathrm{I}(X;Y) &= \mathrm{D}(\mathsf{f}_{X,Y} \| \mathsf{f}_X \times \mathsf{f}_Y) \\ &= \mathsf{E}\left[\log \frac{\mathsf{f}_{X,Y}(X,Y)}{\mathsf{f}_X(X)\mathsf{f}_Y(Y)}\right] \\ &= \mathrm{h}(X) + \mathrm{h}(Y) - \mathrm{h}(X,Y) = \mathrm{h}(X) - \mathrm{h}(X|Y) = \mathrm{h}(Y) - \mathrm{h}(Y|X). \end{aligned}$$

- More generally, if $Y|X = x$ has a density for all x, then

$$\mathrm{I}(X;Y) = \mathrm{D}\left(\mathsf{f}_{Y|X} \middle\| \mathsf{f}_Y \middle| \mathsf{P}_X\right) = \mathrm{h}(Y) - \mathrm{h}(Y|X).$$

Properties that extend to continuous r.v.'s

Proposition 2 (Chain rule)

Suppose (X, Y) *is jointly distributed with joint PDF* $\mathrm{f}_{X,Y}$ *and* $\mathrm{h}(X, Y)$ *exists, then* $\mathrm{h}(X, Y) = \mathrm{h}(X) + \mathrm{h}(Y|X)$*. More generally,*

$$\mathrm{h}(X^n) = \textstyle\sum_{i=1}^{n} \mathrm{h}\left(X_i \middle| X^{i-1}\right).$$

Proposition 3 (Conditioning reduces differential entropy)

$$\mathrm{h}(X|Y) \le \mathrm{h}(X), \quad \mathrm{h}(X|Y, Z) \le \mathrm{h}(X|Z).$$

New properties of differential entropy

Differential entropy can be negative.

Since $b-a$ can be made arbitrarily small, $\mathrm{h}(X) = \log(b-a)$ can be negative. Hence, the non-negative property of entropy *cannot* be extended to differential entropy.

Scaling will change the differential entropy.

Consider $X \sim \mathrm{Unif}[0,1]$. Then, $2X \sim \mathrm{Unif}[0,2]$. Hence,

$$\mathrm{h}(X) = \log 1 = 0,\ \mathrm{h}(2X) = \log 2 = 1 \implies \mathrm{h}(X) \neq \mathrm{h}(2X).$$

This is in sharp contrast to entropy: $\mathrm{H}(X) = \mathrm{H}(g(X))$ as long as $g(\cdot)$ is an invertible function.

Scaling and translation

Proposition 4 (Scaling and Translation in the Scaler Case)

Let X be a continuous random variable with differential entropy $\mathrm{h}(X)$.

- *Translation does not change the differential entropy: For a constant c, $\mathrm{h}(X+c) = \mathrm{h}(X)$.*
- *Scaling shifts the differential entropy: For a constant $a \neq 0$, $\mathrm{h}(aX) = \mathrm{h}(X) + \log|a|$.*

Proposition 5 (Scaling and Translation in the Vector Case)

Let $\boldsymbol{X}$ be a continuous random vector with differential entropy $h(\boldsymbol{X})$.

- *For a constant vector $\boldsymbol{c}$, $\mathrm{h}(\boldsymbol{X}+\boldsymbol{c}) = \mathrm{h}(\boldsymbol{X})$.*
- *For an invertible matrix $\mathbf{a} \in \mathbb{R}^{n\times n}$, $\mathrm{h}(\mathbf{a}\boldsymbol{X}) = \mathrm{h}(\boldsymbol{X}) + \log|\det(\mathbf{a})|$.*

The proof of these propositions are left as exercises (simple calculus).

Differential entropy of Gaussian random vectors

Example 3 (Differential entropy of Gaussian random vectors)

For a n-dim random vector $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{m}, \mathbf{k})$, its differential entropy $\mathrm{h}(\boldsymbol{X}) = \frac{1}{2}\log\left((2\pi e)^n \det(\mathbf{k})\right)$.

sol: For an n-dim random vector $\boldsymbol{X} \sim \mathrm{N}(\boldsymbol{m}, \mathbf{k})$, we can rewrite $\boldsymbol{X}$ as

$$\boldsymbol{X} = \mathbf{a}\boldsymbol{W} + \boldsymbol{m},$$

where $\mathbf{a}\mathbf{a}^{\intercal} = \mathbf{k}$ and $\boldsymbol{W}$ consists of $W_i \overset{\text{i.i.d.}}{\sim} \mathrm{N}(0,1)$, $i = 1, ..., n$.

Hence, by the translation and scaling properties of differential entropy:

$$\begin{aligned}\mathrm{h}(\boldsymbol{X}) &= \mathrm{h}(\boldsymbol{W}) + \log|\det(\mathbf{a})| = \textstyle\sum_{i=1}^{n} \mathrm{h}(W_i) + \frac{1}{2}\log\det(\mathbf{k}) \\ &= \frac{n}{2}\log(2\pi e) + \frac{1}{2}\log\det(\mathbf{k}) = \frac{1}{2}\log\left((2\pi e)^n \det(\mathbf{k})\right).\end{aligned}$$

Maximum differential entropy

Uniform distribution maximizes entropy for r.v. with finite support.

For differential entropy, the maximization problem needs to be associated with constraints on the distribution. (otherwise, it is simple to make it infinite)

It turns out that, under a **second moment constraint**, zero-mean Gaussian maximizes the differential entropy.

Theorem 2 (Maximum Differential Entropy under Covariance Constraint)

Let $\boldsymbol{X}$ be a random vector with mean $\boldsymbol{m}$ and covariance matrix

$$\mathsf{E}\left[(\boldsymbol{X}-\boldsymbol{m})(\boldsymbol{X}-\boldsymbol{m})^{\intercal}\right]=\mathbf{k},$$

and $\boldsymbol{X}^{\mathrm{G}}$ be Gaussian with the same covariance $\mathbf{k}$. Then,

$$\mathrm{h}(\boldsymbol{X}) \le \mathrm{h}\left(\boldsymbol{X}^{\mathrm{G}}\right)=\frac{1}{2}\log\left((2\pi e)^n \det(\mathbf{k})\right).$$

pf: First, we can assume WLOG that both $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{G}}$ are zero-mean, since translation does not change the differential entropy.

Let the PDF of $\boldsymbol{X}$ be $\mathsf{f}(\boldsymbol{x})$ and the PDF of $\boldsymbol{X}^{\mathrm{G}}$ be $\mathsf{f}^{\mathrm{G}}(\boldsymbol{x})$. Hence,

$$0 \le \mathrm{D}\left(\mathsf{f}\middle\|\mathsf{f}^{\mathrm{G}}\right) = \mathsf{E}\left[\log \mathsf{f}(\boldsymbol{X})\right] - \mathsf{E}\left[\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{X})\right] = -\mathrm{h}\left(\boldsymbol{X}\right) - \mathsf{E}_{\boldsymbol{X}\sim\mathsf{f}}\left[\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{X})\right].$$

Note that $\log \mathsf{f}^{\mathrm{G}}\left(\boldsymbol{x}\right)$ is a quadratic function of $\boldsymbol{x}$. $\boldsymbol{X}$ and $\boldsymbol{X}^{\mathrm{G}}$ have the same second moment. Hence,

$$\begin{aligned} &\mathsf{E}_{\boldsymbol{X}\sim\mathsf{f}}\left[\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{X})\right] = \mathsf{E}_{\boldsymbol{X}\sim\mathsf{f}^{\mathrm{G}}}\left[\log \mathsf{f}^{\mathrm{G}}(\boldsymbol{X})\right] = -\mathrm{h}\left(\boldsymbol{X}^{\mathrm{G}}\right) \\ \Longrightarrow\ &0 \le \mathrm{D}\left(\mathsf{f}\middle\|\mathsf{f}^{\mathrm{G}}\right) = -\mathrm{h}\left(\boldsymbol{X}\right) + \mathrm{h}\left(\boldsymbol{X}^{\mathrm{G}}\right) \\ \Longrightarrow\ &\mathrm{h}\left(\boldsymbol{X}\right) \le \mathrm{h}\left(\boldsymbol{X}^{\mathrm{G}}\right). \end{aligned}$$

□

In general, maximum (differential) entropy problems subject to moment constraints can be solved via the non-negativity of KL divergence.

1 Discrete Memoryless Channel with an Input Cost Constraint

2 Extension to Continuous Channels

- Information measures for continuous distributions
- Channel coding theorems for continuous memoryless channels

Coding theorems: from discrete to continuous

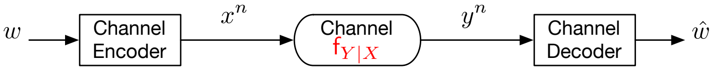
Two main techniques for extending the achievability part of coding theorems from the discrete world to the continuous world:

1. **Discretization**: Discretize the source and channel input/output to create a discrete system, and then make the discretization finer and finer to prove the achievability.
2. **New typicality**: Extend weak typicality for continuous r.v. and repeat the arguments in a similar way. In particular, replace the entropy terms in the definitions of weakly typical sequences by differential entropy terms.

Using discretization to derive the achievability of Gaussian channel capacity follows Gallager [Gal68]. Cover&Thomas [CT06] uses weak typicality for continuous r.v.'s. Moser [Mos18] uses threshold decoder, similar to weak typicality.

In this lecture, we sketch how to use **discretization** for the achievability proof. The proof, however, is not the focus of the lecture.

Continuous memoryless channel



1. Input/output alphabet $\mathcal{X} = \mathcal{Y} = \mathbb{R}$.
2. Continuous Memoryless Channel (CMC):
   - **Channel Law**: Governed by the conditional density (PDF) $\mathsf{f}_{Y|X}$.
   - **Memoryless**: $Y_i - X_i - \left(X^{i-1}, Y^{i-1}\right)$.
3. Average input cost constraint B: $\frac{1}{n}\sum_{i=1}^{n} b(x_i) \le$ B, where $b : \mathbb{R} \to [0, \infty)$ is the (single-letter) cost function.

The definitions of error probability, achievable rate, and capacity, are the same as those in DMC.

Channel coding theorem

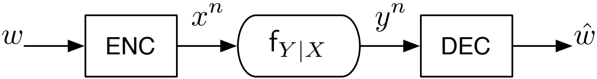
Theorem 3 (Continuous Memoryless Channel Capacity)

The capacity of the CMC $\left(\mathbb{R}, \mathsf{f}_{Y|X}, \mathbb{R}\right)$ *with input cost constraint* B *is*

$$\mathrm{C}(\mathsf{B}) = \sup_{X:\ \mathsf{E}[b(X)] \le \mathsf{B}} \mathrm{I}(X;Y)\,. \tag{2}$$

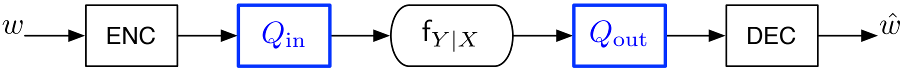
Converse proof: Exactly the same as that in the DMC case.

Sketch of the achievability (1): discretization



Achievability proof makes use of discretization —we can apply the result in DMC with input cost:

Sketch of the achievability (1): discretization

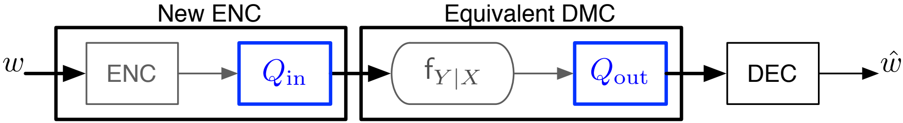


Achievability proof makes use of discretization —we can apply the result in DMC with input cost:

- Q_{in}: (single-letter) discretization that maps $X \in \mathbb{R}$ to $X_{\text{d}} \in \mathcal{X}_{\text{d}}$.
- Q_{out}: (single-letter) discretization that maps $Y \in \mathbb{R}$ to $Y_{\text{d}} \in \mathcal{Y}_{\text{d}}$.

Note that both $\mathcal{X}_{\text{d}}$ and $\mathcal{Y}_{\text{d}}$ are discrete (countable) alphabets.

Sketch of the achievability (1): discretization



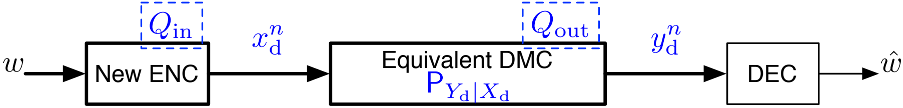
Achievability proof makes use of discretization – we can apply the result in DMC with input cost:

- Q_{in}: (single-letter) discretization that maps $X \in \mathbb{R}$ to $X_{\text{d}} \in \mathcal{X}_{\text{d}}$.
- Q_{out}: (single-letter) discretization that maps $Y \in \mathbb{R}$ to $Y_{\text{d}} \in \mathcal{Y}_{\text{d}}$.

Note that both $\mathcal{X}_{\text{d}}$ and $\mathcal{Y}_{\text{d}}$ are discrete (countable) alphabets.

Idea: With the two discretization blocks Q_{in} and Q_{out}, one can build an *equivalent DMC* $\left(\mathcal{X}_{\text{d}}, \mathsf{P}_{Y_{\text{d}}|X_{\text{d}}}, \mathcal{Y}_{\text{d}}\right)$ as shown above.

Sketch of the achievability (2): arguments

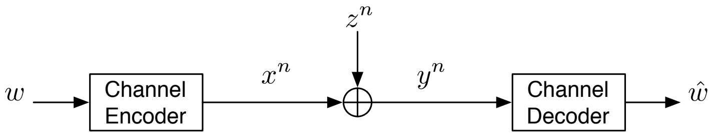


1. **Random codebook generation**: Generate the codebook randomly based on the original (continuous) r.v. X, satisfying $\mathsf{E}[b(X)] \le \mathsf{B}$.
2. **Choice of discretization**: Choose Q_{in} such that the cost constraint will not be violated after discretization. Specifically, $\mathsf{E}\left[b(X_{\text{d}})\right] \le \mathsf{B}$.
3. **Achievability in the equivalent DMC**: By the achievability part of the channel coding theorem for DMC with input constraint, any rate $\mathrm{R} < \mathrm{I}(X_{\text{d}}; Y_{\text{d}})$ is achievable.
4. **Achievability in the original CMC**: Prove that when the discretization in Q_{in} and Q_{out} gets finer and finer, $\mathrm{I}(X_{\text{d}}; Y_{\text{d}}) \to \mathrm{I}(X; Y)$.

   $\mathrm{I}(X; Y) \ge \limsup \mathrm{I}(X_{\text{d}}; Y_{\text{d}})$: easy by DPI.

   $\mathrm{I}(X; Y) \le \liminf \mathrm{I}(X_{\text{d}}; Y_{\text{d}})$: check Chapter 4.6 of Polyanskiy& Wu.

Additive white Gaussian noise (AWGN) channel



1. Input/output alphabet $\mathcal{X} = \mathcal{Y} = \mathbb{R}$.
2. AWGN Channel:
   - Conditional PDF $\mathsf{f}_{Y|X}$ is given by $Y = X + Z$, $Z \sim \mathrm{N}(0, \sigma^2) \perp\!\!\!\perp X$.
   - $\{Z_i\}$ form an i.i.d. (white) Gaussian r.p. with $Z_i \sim \mathrm{N}(0, \sigma^2)$, $\forall\, i$.
   - Memoryless: $Z_i \perp\!\!\!\perp \left(W, X^{i-1}, Z^{i-1}\right)$.
   - Without feedback: $Z^n \perp\!\!\!\perp X^n$.
3. Average input power constraint B: $\frac{1}{n}\sum_{i=1}^{n} |x_i|^2 \le$ B.

Channel coding theorem for the Gaussian channel

Example 4 (Gaussian Channel Capacity)

The capacity of the AWGN channel with input power constraint B and noise variance σ^2 is given by

$$\mathrm{C}(\mathsf{B}) = \sup_{X:\ \mathsf{E}[|X|^2] \le \mathsf{B}} \mathrm{I}(X;Y) = \frac{1}{2}\log\left(1 + \frac{\mathsf{B}}{\sigma^2}\right). \tag{3}$$

Note: For the AWGN channel, the supremum is actually attainable with Gaussian input $X \sim \mathrm{N}(0, \mathrm{B})$, that is, the input has density

$$\mathsf{f}_X(x) = \frac{1}{\sqrt{2\pi\mathsf{B}}} e^{-\frac{x^2}{2\mathsf{B}}},$$

as shown in the next slide.

sol: Let us compute the capacity of AWGN channel (3) as follows:

$$\begin{aligned} \mathrm{I}(X;Y) &= \mathrm{h}(Y) - \mathrm{h}(Y|X) = \mathrm{h}(Y) - \mathrm{h}(X+Z|X) \\ &= \mathrm{h}(Y) - \mathrm{h}(Z|X) = \mathrm{h}(Y) - \mathrm{h}(Z) & \text{(since } Z \perp\!\!\!\perp X\text{)} \\ &= \mathrm{h}(Y) - \tfrac{1}{2}\log\left(2\pi e\,\sigma^2\right) \\ &\overset{(\mathrm{a})}{\le} \tfrac{1}{2}\log\left((2\pi e)(\mathsf{B}+\sigma^2)\right) - \tfrac{1}{2}\log\left(2\pi e\,\sigma^2\right) \\ &= \tfrac{1}{2}\log\left(1+\tfrac{\mathsf{B}}{\sigma^2}\right). \end{aligned}$$

Here (a) is due to $\mathrm{h}(Y) \le \frac{1}{2}\log(2\pi e\,\mathsf{Var}[Y])$ and

$$\mathsf{Var}[Y] = \mathsf{Var}[X] + \mathsf{Var}[Z] \le \mathsf{B} + \sigma^2,$$

since $\mathsf{Var}[X] \le \mathsf{E}\left[X^2\right] \le \mathsf{B}$.

Finally, note that the above inequalities hold with equality when $X \sim \mathrm{N}(0,\mathsf{B})$.

Practical relevance of the Gaussian noise model

In communication engineering, the additive Gaussian noise is the most widely used model for a noisy channel with real (complex) input/output.

Reasons:

1. Due to CLT, Gaussian well models noise that is the aggregation of many small perturbations.
2. Analytically Gaussian is highly tractable.
3. Consider a input-power-constrained channel with independent additive noise. Within the family of noise distributions that have the same variance, Gaussian noise is the worst case noise.

The last point is important: for a additive-noise-channel with input power constraint B and noise variance σ^2, its capacity is lower bounded by the Gaussian channel capacity $\frac{1}{2}\log\left(1+\frac{\mathsf{B}}{\sigma^2}\right)$.

Gaussian noise is the worst-case noise

Proposition 6

Consider a Gaussian r.v. $X^{\mathrm{G}} \sim \mathrm{N}(0, \mathsf{B})$ *and* $Y = X^{\mathrm{G}} + Z$*, where* Z *has density* f_Z*, variance* $\mathsf{Var}[Z] = \sigma^2$ *and* $Z \perp\!\!\!\perp X^{\mathrm{G}}$*. Then,*

$$\mathrm{I}\left(X^{\mathrm{G}}; Y\right) \ge \frac{1}{2}\log\left(1 + \frac{\mathsf{B}}{\sigma^2}\right).$$

With Proposition 6, we immediately obtain the following theorem:

Theorem 4 (Gaussian is the Worst-Case Additive Noise)

Consider a CMC $\mathsf{f}_{Y|X}$*:* $Y = X + Z$*,* $Z \perp\!\!\!\perp X$*, with input power constraint* B *and noise variance* σ^2*. The additive noise has density. Then, the capacity* $\mathrm{C}(\mathsf{B})$ *is minimized when* $Z \sim \mathrm{N}(0, \sigma^2)$*, that is,*

$$\mathrm{C}(\mathsf{B}) \ge \mathrm{C}^{\mathrm{G}}(\mathsf{B}) := \frac{1}{2}\log\left(1 + \frac{\mathsf{B}}{\sigma^2}\right).$$

Proof of Proposition 6

Let $Z^{\mathrm{G}} \sim \mathrm{N}(0, \sigma^2)$, and denote $Y^{\mathrm{G}} := X^{\mathrm{G}} + Z^{\mathrm{G}}$. We aim to prove

$$\mathrm{I}\left(X^{\mathrm{G}}; Y\right) \ge \mathrm{I}\left(X^{\mathrm{G}}; Y^{\mathrm{G}}\right).$$

First note that $\mathrm{I}\left(X^{\mathrm{G}}; Y\right) = \mathrm{h}(Y) - \mathrm{h}(Z)$ does not change if we shift Z by a constant. Hence, WLOG assume $\mathsf{E}[Z] = 0$. Since both X^{G} and Z are zero-mean, so is Y.

Note that $Y^{\mathrm{G}} \sim \mathrm{N}(0, \mathsf{B} + \sigma^2)$ and $Z^{\mathrm{G}} \sim \mathrm{N}(0, \sigma^2)$. Hence,

$$\begin{aligned}
\mathrm{h}\left(Y^{\mathrm{G}}\right) &= \mathsf{E}_{Y^{\mathrm{G}}}\left[-\log \mathsf{f}_{Y^{\mathrm{G}}}\left(Y^{\mathrm{G}}\right)\right] \\
&= \tfrac{1}{2}\log\left(2\pi(\mathsf{B}+\sigma^2)\right) + \tfrac{\log e}{2(\mathsf{B}+\sigma^2)}\mathsf{E}_{Y^{\mathrm{G}}}\left[\left(Y^{\mathrm{G}}\right)^2\right] \\
&= \tfrac{1}{2}\log\left(2\pi(\mathsf{B}+\sigma^2)\right) + \tfrac{\log e}{2(\mathsf{B}+\sigma^2)}\mathsf{E}_{Y}\left[(Y)^2\right] \\
&= \mathsf{E}_{Y}\left[-\log \mathsf{f}_{Y^{\mathrm{G}}}(Y)\right]
\end{aligned}$$

Key: Y and Y^{G} have the same variance.

Following a similar derivation, $\mathrm{h}\left(Z^{\mathrm{G}}\right) = \mathsf{E}_Z\left[-\log \mathsf{f}_{Z^{\mathrm{G}}}(Z)\right]$. Therefore,

$$
\begin{aligned}
&\mathrm{I}\left(X^{\mathrm{G}};Y^{\mathrm{G}}\right) - \mathrm{I}\left(X^{\mathrm{G}};Y\right) \\
&= \left\{\mathrm{h}\left(Y^{\mathrm{G}}\right) - \mathrm{h}(Y)\right\} - \left\{\mathrm{h}\left(Z^{\mathrm{G}}\right) - \mathrm{h}(Z)\right\} \\
&= \left\{\mathsf{E}_Y\left[-\log \mathsf{f}_{Y^{\mathrm{G}}}(Y)\right] - \mathsf{E}_Y\left[-\log \mathsf{f}_Y(Y)\right]\right\} \\
&\quad - \left\{\mathsf{E}_Z\left[-\log \mathsf{f}_{Z^{\mathrm{G}}}(Z)\right] - \mathsf{E}_Z\left[-\log \mathsf{f}_Z(Z)\right]\right\} \\
&= \mathsf{E}_Y\left[\log \frac{\mathsf{f}_Y(Y)}{\mathsf{f}_{Y^{\mathrm{G}}}(Y)}\right] - \mathsf{E}_Z\left[\log \frac{\mathsf{f}_Z(Z)}{\mathsf{f}_{Z^{\mathrm{G}}}(Z)}\right] \\
&= \mathsf{E}_{Y,Z}\left[\log \frac{\mathsf{f}_Y(Y)\mathsf{f}_{Z^{\mathrm{G}}}(Z)}{\mathsf{f}_{Y^{\mathrm{G}}}(Y)\mathsf{f}_Z(Z)}\right] \\
&\le \log\left(\mathsf{E}_{Y,Z}\left[\frac{\mathsf{f}_Y(Y)\mathsf{f}_{Z^{\mathrm{G}}}(Z)}{\mathsf{f}_{Y^{\mathrm{G}}}(Y)\mathsf{f}_Z(Z)}\right]\right). \quad \text{(Jensen's Inequality)}
\end{aligned}
$$

To finish the proof, we shall prove that $\mathsf{E}_{Y,Z}\left[\frac{\mathsf{f}_Y(Y)\mathsf{f}_{Z^{\mathrm{G}}}(Z)}{\mathsf{f}_{Y^{\mathrm{G}}}(Y)\mathsf{f}_Z(Z)}\right] = 1$. This is done is the next slide.

$$
\begin{aligned}
&\mathsf{E}_{Y,Z}\left[\frac{\mathsf{f}_Y(Y)\mathsf{f}_{Z^{\mathrm{G}}}(Z)}{\mathsf{f}_{Y^{\mathrm{G}}}(Y)\mathsf{f}_Z(Z)}\right] \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathsf{f}_{Y,Z}(y,z)\frac{\mathsf{f}_Y(y)\mathsf{f}_{Z^{\mathrm{G}}}(z)}{\mathsf{f}_{Y^{\mathrm{G}}}(y)\mathsf{f}_Z(z)}\,\mathrm{d}z\,\mathrm{d}y \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathsf{f}_Z(z)\mathsf{f}_{X^{\mathrm{G}}}(y-z)\frac{\mathsf{f}_Y(y)\mathsf{f}_{Z^{\mathrm{G}}}(z)}{\mathsf{f}_{Y^{\mathrm{G}}}(y)\mathsf{f}_Z(z)}\,\mathrm{d}z\,\mathrm{d}y && (\because Y = X^{\mathrm{G}} + Z) \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \left[\mathsf{f}_{X^{\mathrm{G}}}(y-z)\mathsf{f}_{Z^{\mathrm{G}}}(z)\right]\frac{\mathsf{f}_Y(y)}{\mathsf{f}_{Y^{\mathrm{G}}}(y)}\,\mathrm{d}z\,\mathrm{d}y \\
&= \int_{\mathbb{R}}\int_{\mathbb{R}} \mathsf{f}_{Y^{\mathrm{G}},Z^{\mathrm{G}}}(y,z)\frac{\mathsf{f}_Y(y)}{\mathsf{f}_{Y^{\mathrm{G}}}(y)}\,\mathrm{d}z\,\mathrm{d}y && (\because Y^{\mathrm{G}} = X^{\mathrm{G}} + Z^{\mathrm{G}}) \\
&= \int_{\mathbb{R}} \frac{\mathsf{f}_Y(y)}{\mathsf{f}_{Y^{\mathrm{G}}}(y)}\left(\int_{\mathbb{R}} \mathsf{f}_{Y^{\mathrm{G}},Z^{\mathrm{G}}}(y,z)\,\mathrm{d}z\right)\mathrm{d}y \\
&= \int_{\mathbb{R}} \frac{\mathsf{f}_Y(y)}{\mathsf{f}_{Y^{\mathrm{G}}}(y)}\mathsf{f}_{Y^{\mathrm{G}}}(y)\,\mathrm{d}y = \int_{\mathbb{R}} \mathsf{f}_Y(y)\,\mathrm{d}y = 1.
\end{aligned}
$$

Hence, the proof is complete. □

Summary

- DMC with average input cost constraint:

$$\mathrm{C}(\mathsf{B}) = \max_{\mathsf{P}_X :\, \mathsf{E}_{\mathsf{P}_X}[b(X)] \le \mathsf{B}} \mathrm{I}(X;Y)\,.$$

- Continuous memoryless channel capacity:

$$\mathrm{C}(\mathsf{B}) = \sup_{X:\ \mathsf{E}[b(X)] \le \mathsf{B}} \mathrm{I}(X;Y)\,.$$

- Gaussian channel capacity:

$$\mathrm{C}(\mathsf{B}) = \frac{1}{2}\log\left(1 + \frac{\mathsf{B}}{\sigma^2}\right).$$

- For an additive noise channel, Gaussian noise is the worst noise under a second moment constraint.

- Mutual information between two continuous r.v.'s X and Y with joint density $\mathsf{f}_{X,Y}$: $\mathrm{I}(X;Y) = \mathsf{E}\left[\log \frac{\mathsf{f}_{X,Y}(X,Y)}{\mathsf{f}_X(X)\mathsf{f}_Y(Y)}\right]$.

- Differential entropy and conditional differential entropy: $\mathrm{h}(X) := \mathsf{E}\left[\log \frac{1}{\mathsf{f}_X(X)}\right]$, $\mathrm{h}(X|Y) := \mathsf{E}\left[\log \frac{1}{\mathsf{f}_{X|Y}(X|Y)}\right]$.

- $\mathrm{I}(X;Y) = \mathrm{h}(X) - \mathrm{h}(X|Y) = \mathrm{h}(Y) - \mathrm{h}(Y|X)$.

- Chain rule, conditioning reduces differential entropy, non-negativity of mutual information and KL divergence: remain to hold.

- Differential entropy can be negative; $\mathrm{h}(X) \not\leq \mathrm{h}(X,Y)$.

- Translating a r.v. does not change the differential entropy, while scaling it shifts its differential entropy.

- Under a second moment constraint, zero-mean Gaussian maximizes the differential entropy.

References

[CT06] Thomas M. Cover and Joy A. Thomas.
Elements of Information Theory (Wiley Series in Telecommunications and Signal Processing).
Wiley-Interscience, USA, 2006.

[Gal68] Robert G. Gallager.
Information Theory and Reliable Communication.
John Wiley & Sons Inc., USA, 1968.

[Mos18] Stefan M. Moser.
Information Theory (Lecture Notes).
ISI Lab, ETH Zürich, Switzerland, and ICE, NYCU, Hsinchu, Taiwan, 2018.