Delivering Information Reliably

I-Hsiang Wang

Department of Electrical Engineering
National Taiwan University

ihwang@ntu.edu.tw

October 19, 2023

The information processing task motivating the study of this lecture:

For a length-k i.i.d. $\mathrm{Ber}(\frac{1}{2})$ bit sequence, given a noisy channel with input alphabet $\mathcal{X}$ and output alphabet $\mathcal{Y}$, design a good encoding scheme to represent the k-bit sequence using n symbols in $\mathcal{X}$, and a good decoding scheme to reconstruct the k-bit sequence reliably from the observed n symbols in $\mathcal{Y}$ at the output of the channel.

Note: how the encoding and decoding scheme work are known *a priori*.

Fundamental Questions:

- What is the maximum possible ratio $\frac{k}{n}$ (transmission rate) ?
- How to achieve that fundamental limit?

The channel coding problem (Shannon's abstraction)

Meta Description

1. **Message**: Random message $W \sim \mathrm{Unif}\{1, \ldots, 2^k\}$.
2. **Channel**: Consist of an input alphabet $\mathcal{X}$, an output alphabet $\mathcal{Y}$, and a family of conditional laws $\left\{\mathsf{P}_{Y_k|X^k,Y^{k-1}} \,\middle|\, k \in \mathbb{N}\right\}$ determining the stochastic relationship between the output symbol Y_k and the input symbol X_k along with all past signals (X^{k-1}, Y^{k-1}).
3. **Encoder**: Encode the message w by a length-n codeword $x^n \in \mathcal{X}^n$.
4. **Decoder**: Reconstruct message $\hat{w}$ from the channel output y^n.
5. **Efficiency**: Maximize the code rate $\mathrm{R} := \frac{k}{n}$ bits/channel use, given a certain decoding criterion.

Decoding criterion: small error probability

A key performance measure: Error Probability $\mathsf{P}_\mathsf{e}^{(n)} := \Pr\{W \neq \hat{W}\}$.

Question: Is it possible to get zero error probability?

Answer: Probably not, unless the channel noise has some special structure.

Following the development of lossless source coding, Shannon turned the attention to answering the following question:

> *Is it possible to have a sequence of encoder/decoder pairs such that $\mathsf{P}_\mathsf{e}^{(n)} \to 0$ as $n \to \infty$? If so, what is the largest possible code rate R where vanishing error probability is possible?*

Encoding and decoding

Decoder takes the observation Y^n generated from the noisy channel and make a decision

$$\hat{W} = \mathrm{dec}(Y^n) \in \mathcal{W} \equiv \{1, \ldots, 2^k\}.$$

This is just a 2^k-ary detection (hypothesis testing) problem, with

$$\mathcal{H}_w : Y^n \sim \mathsf{P}_w \equiv \mathsf{P}_{Y^n|X^n}(\cdot|x^n(w)),\ w = 1, \ldots, 2^k.$$

Looks pretty simple – just use maximum likelihood.

But the analysis of the error probability is tricky.

Encoding and decoding (cont'd)

The even more tricky part is the encoder.

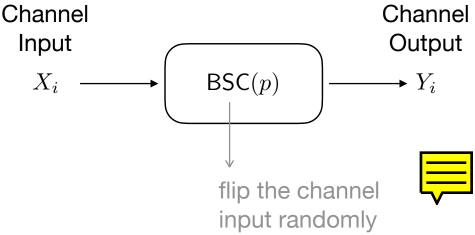
Encoder places the 2^k *codewords* in $\mathcal{X}^n$ so that after passing through the noisy channel $\mathsf{P}_{Y^n|X^n}$, the distributions of the observations

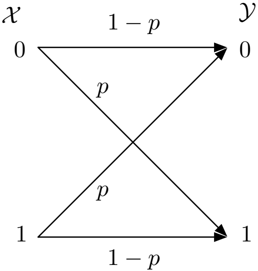
$$\left\{\mathsf{P}_w(\cdot) \equiv \mathsf{P}_{Y^n|X^n}(\cdot|x^n(w)) \,\middle|\, w = 1, \ldots, 2^k\right\}$$

are well-separated.

To make these challenges concrete and see that there are ways to resolve them, let us begin with an example – the binary symmetric channel (BSC), one of the simplest yet non-trivial noisy channel.

Motivating example: the binary symmetric channel





- Binary input/output: $x_i, y_i \in \{0,1\}$, $i = 1, 2, \ldots$.
- Channel flips the input bit i.i.d. with probability $p \in (0, 1/2)$:

$$Y_i = X_i \oplus Z_i,\ Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p),\ i = 1, 2, \ldots.$$

Encoding

$$W \equiv \begin{bmatrix} B_1 & B_2 & \dots & B_k \end{bmatrix} \longrightarrow \boxed{\text{enc}} \longrightarrow \begin{bmatrix} X_1 & X_2 & \dots & \dots & X_n \end{bmatrix}$$

message

$\boldsymbol{B} \sim \mathrm{Unif}\big(\{0,1\}^k\big)$

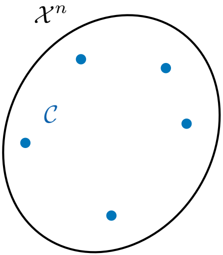
codeword

$\boldsymbol{X} \in \{0,1\}^n$

- Goal: introduce redundancy to combat channel noises (for BSC, bit flips).
- Codebook: the range of the encoding function enc:

$$\mathcal{C} = \left\{\mathrm{enc}(w) \equiv \boldsymbol{x}(w) \,\middle|\, w \in \mathcal{W} \equiv \{1, \dots, 2^k\}.\right\}.$$

- eg. Repetition coding $\mathcal{C} = \{[0 \ \dots \ 0], [1 \ \dots \ 1]\}$.



Pictorially, the encoding function should place the codewords far apart so that they do not get confused with one another.

Decoding

$$\begin{bmatrix} Y_1 & Y_2 & \dots & \dots & Y_n \end{bmatrix} \longrightarrow \boxed{\text{dec}} \longrightarrow \begin{bmatrix} \hat{B}_1 & \hat{B}_2 & \dots & \hat{B}_k \end{bmatrix} \equiv \hat{W}$$

received signal $\quad\quad\quad\quad$ decoded message

$\boldsymbol{Y} \in \{0,1\}^n \quad\quad\quad\quad \hat{W} \in \mathcal{W}$

- Performance metic: probability of error $\mathsf{P}_{\mathsf{e}}^{(n)} = \mathsf{Pr}\{\hat{W} \neq W\}$.
- Since $W \sim \mathrm{Unif}(\mathcal{W})$ (uniform prior), *maximum likelihood* decoding is the optimal: $\hat{w}_{\mathrm{ML}}(\boldsymbol{y}) = \arg\max_{w \in \mathcal{W}} \mathsf{P}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{y}\middle|\boldsymbol{x}(w)\right)$.
- For BSC, likelihood function is simple and determined by the *Hamming distance* between $\boldsymbol{y}$ and the candidate codeword $\boldsymbol{x}(w)$:

$$\mathsf{P}_{\boldsymbol{Y}|\boldsymbol{X}}\left(\boldsymbol{y}\middle|\boldsymbol{x}(w)\right) = p^{\mathrm{d_H}(\boldsymbol{y},\boldsymbol{x}(w))}(1-p)^{n-\mathrm{d_H}(\boldsymbol{y},\boldsymbol{x}(w))} = (1-p)^n \left(\tfrac{p}{1-p}\right)^{\mathrm{d_H}(\boldsymbol{y},\boldsymbol{x}(w))}.$$

Since $0 < p < 1/2$, we have

$$\hat{w}_{\mathrm{ML}}(\boldsymbol{y}) = \arg\min_{w \in \mathcal{W}} \mathrm{d_H}\left(\boldsymbol{y}, \boldsymbol{x}(w)\right)$$

Error probability analysis

Under ML decoding, the error probability is given as

$$\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)} = \frac{1}{2^k} \sum_{w \in \mathcal{W}} \underbrace{\Pr\left\{\exists\, \tilde{w} \neq w \text{ such that } \mathrm{d_H}(\boldsymbol{Y}, \boldsymbol{x}(\tilde{w})) \le \mathrm{d_H}(\boldsymbol{Y}, \boldsymbol{x}(w)) \,\middle|\, W = w\right\}}_{\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)}(w)}$$

The event that $\boldsymbol{x}(\tilde{w})$ is closer to $\boldsymbol{Y}$ than the actual $\boldsymbol{x}(w)$ is just the event that BSC flips more than half of the bits at which $\boldsymbol{x}(w)$ and $\boldsymbol{x}(\tilde{w})$ differ.

Hence by the union bound, we may get to

$$\mathsf{P}_{\mathsf{e},\mathrm{ML}}^{(n)}(w) \le \sum_{\substack{\tilde{w} \in \mathcal{W} \\ \tilde{w} \neq w}} \sum_{j = \left\lceil \frac{\mathrm{d_H}(\boldsymbol{x}(\tilde{w}), \boldsymbol{x}(w))}{2} \right\rceil}^{\mathrm{d_H}(\boldsymbol{x}(\tilde{w}), \boldsymbol{x}(w))} \binom{\mathrm{d_H}(\boldsymbol{x}(\tilde{w}), \boldsymbol{x}(w))}{j} p^j (1-p)^{\mathrm{d_H}(\boldsymbol{x}(\tilde{w}), \boldsymbol{x}(w)) - j}.$$

It largely depends on how the encoding function populates the codewords in the codebook and hard to analyze, let alone the asymptote as $n \to \infty$.

Reflections

- Recall: for lossless source coding, we leveraged the *concentration of probability* (AEP) when $n \to \infty$ to achieve data compression.

  Reason: distribution of the random source is known and "well-behaved".

- But for the channel coding problem, the encoding function enc makes the distribution of coded symbols non-i.i.d. in general, and it becomes hard to control the statistical behavior of the codewords.

$$\underset{\text{i.i.d. over time}}{\boldsymbol{B}} \longrightarrow \boxed{\mathrm{enc}} \longrightarrow \boldsymbol{x}(\boldsymbol{B}) = \big[x_1(\boldsymbol{B}) \quad \underset{\text{not i.i.d. over time}}{x_2(\boldsymbol{B}) \quad \dots} \quad x_n(\boldsymbol{B})\big]$$

Work-around:

1. "Random" encoding
2. "Typicality" decoding

Typicality decoding

Observation: bit-flips are i.i.d. $\mathrm{Ber}(p)$, and hence the number of bit-flips follows $\mathrm{Binom}(n, p)$ and concentrates at np when $n \to \infty$. So, it is "typical" that there are $\approx np$ bit-flips, that is, $\forall\, \varepsilon > 0$,

$$\mathsf{Pr}\Big\{ \overbrace{\mathrm{w}(\boldsymbol{Z})}^{\text{\# of bit-flips}} \le n(p+\varepsilon) \Big\} \to 1 \quad \text{as } n \to \infty.$$

Typicality decoding:

$$\hat{w}_{\mathrm{T}} = \text{the unique } w \in \mathcal{W} \text{ such that } \mathrm{d_H}(\boldsymbol{y}, \boldsymbol{x}(w)) \le n(p+\varepsilon).$$

Obviously, $\mathsf{P}^{(n)}_{\mathsf{e},\mathrm{ML}}(w) \le \mathsf{P}^{(n)}_{\mathsf{e},\mathrm{T}}(w)$. Furthermore, by the union bound,

$$\begin{aligned}\mathsf{P}^{(n)}_{\mathsf{e},\mathrm{T}}(w) \le\ & \underbrace{\mathsf{Pr}\Big\{ \overbrace{\mathrm{w}(\boldsymbol{Z})}^{\mathrm{d_H}(\boldsymbol{x}(w) \oplus \boldsymbol{Z}, \boldsymbol{x}(w))} > n(p+\varepsilon) \Big\}}_{\to 0 \text{ as } n \to \infty} \\ & + \textstyle\sum_{\tilde{w} \in \mathcal{W}, \tilde{w} \neq w} \mathsf{Pr}\Big\{ \mathrm{d_H}(\boldsymbol{x}(w) \oplus \boldsymbol{Z}, \boldsymbol{x}(\tilde{w})) \le n(p+\varepsilon) \Big\} \qquad (*)\end{aligned}$$

Random encoding

$$\Pr\left\{\mathrm{d_H}(\boldsymbol{x}(w) \oplus \boldsymbol{Z}, \boldsymbol{x}(\tilde{w})) \le n(p+\varepsilon)\right\}$$

This term remains difficult to analyze as it depends on the structure of the codebook (more specifically, pairwise Hamming distance between codewords).

To overcome the difficulty, Shannon came up with a *random coding* idea:

> *To prove the existences of codebook a codeb with* $\mathsf{P}_\mathsf{e}^{(n)} \le \epsilon$*, it suffices to show that over a set of codebooks, the averaged* $\mathsf{P}_\mathsf{e}^{(n)} \le \epsilon$*.*

One may consider the codebook $\mathcal{C} = \{\boldsymbol{x}(1), \dots, \boldsymbol{x}(2^k)\}$ as a $2^k \times n$ matrix

$$\mathbf{c} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^k) & x_2(2^k) & \cdots & x_n(2^k) \end{bmatrix}$$

Random codebook simply means that the codebook matrix is random and follows a certain distribution: $\mathbf{C} \sim \mathsf{P}_\mathbf{C} \in \mathcal{P}(\{0,1\}^{2^k \times n})$.

With the random codebook matrix $\mathbf{C}$, the goal now turns to proving

$$\mathsf{E}_{\mathbf{C}\sim\mathsf{P}_{\mathbf{C}},W\sim\mathrm{Unif}(\mathcal{W})}\left[\mathsf{P}^{(n)}_{\mathsf{e},\mathrm{T}}(W;\mathbf{C})\right] \le \epsilon.$$

Clearly, $\mathbf{C}$ is chosen to be independent of W. Hence, it suffices to show that

$$\forall\, w \in \mathcal{W},\ \mathsf{E}_{\mathbf{C}\sim\mathsf{P}_{\mathbf{C}}}\left[\mathsf{P}^{(n)}_{\mathsf{e},\mathrm{T}}(w;\mathbf{C})\right] \le \epsilon.$$

It boils down to show that $\mathsf{E}_{\mathbf{C}\sim\mathsf{P}_{\mathbf{C}}}\left[(*)\right] \le \epsilon/2$, that is,

$$\sum_{\tilde{w}\in\mathcal{W},\tilde{w}\neq w} \mathsf{P}_{\mathbf{C},\boldsymbol{Z}}\Big\{\underbrace{\mathrm{d_H}(\boldsymbol{X}(w)\oplus\boldsymbol{Z},\boldsymbol{X}(\tilde{w}))}_{\mathrm{w}(\boldsymbol{X}(w)\oplus\boldsymbol{Z}\oplus\boldsymbol{X}(\tilde{w}))} \le n(p+\varepsilon)\Big\} \le \epsilon/2.$$

Let's pick a distribution on $\mathbf{C}$ so that the analysis becomes simple:

$$X_i(w) \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(1/2) \quad \forall\, i = 1, 2, \ldots, n,\ \forall\, w = 1, 2, \ldots, 2^k,$$

that is, the entries of $\mathbf{C}$ are i.i.d. $\mathrm{Ber}(1/2)$. Then, each codeword (row of $\mathbf{C}$) is a length-n i.i.d. $\mathrm{Ber}(1/2)$ sequence.

Since $\boldsymbol{X}(w), \boldsymbol{X}(\tilde{w}), \boldsymbol{Z}$ are independent, $\boldsymbol{X}(w) \oplus \boldsymbol{Z} \oplus \boldsymbol{X}(\tilde{w})$ is also a length-n i.i.d. $\mathrm{Ber}(1/2)$ sequence.

As a result, $\forall\, \tilde{w} \neq w$,

$$\begin{aligned} &\mathsf{P}_{\mathbf{C},\boldsymbol{Z}}\left\{\mathrm{w}(\boldsymbol{X}(w) \oplus \boldsymbol{Z} \oplus \boldsymbol{X}(\tilde{w})) \leq n(p+\varepsilon)\right\} \\ &= \mathsf{Pr}\left\{\mathrm{Binom}(n, \tfrac{1}{2}) \leq n(p+\varepsilon)\right\} \leq 2^{-n \mathsf{d_b}\left(p+\varepsilon \,\middle\|\, \frac{1}{2}\right)} \qquad \text{(Chernoff Bound)} \end{aligned}$$

Wrapping-up: achievability

Putting everything together, we show that $\exists$ a codebook $\mathcal{C}$ of size 2^k such that

$$\mathsf{P}^{(n)}_{\mathsf{e},\mathrm{ML}} \le \epsilon/2 + (2^k - 1)2^{-n\mathsf{d_b}\left(p+\varepsilon \,\middle\|\, \frac{1}{2}\right)}.$$

A sufficient condition for the second term to vanish is $k < n\mathsf{d_b}\left(p + \varepsilon \,\middle\|\, \frac{1}{2}\right)$.

Choosing $\varepsilon > 0$ judiciously close to 0, we can show that $\forall\, \delta > 0$ and $\forall\, \epsilon \in (0,1)$, there exists a codebook $\mathcal{C}$ of size 2^k such that

$$k > n\left(\mathsf{d_b}\left(p \,\middle\|\, \tfrac{1}{2}\right) - \delta\right) \quad \text{and} \quad \mathsf{P}^{(n)}_{\mathsf{e},\mathrm{ML}} \le \epsilon.$$

Hence,

$$\liminf_{n\to\infty} \frac{1}{n} k^*(n,\epsilon) \ge \mathsf{d_b}\left(p \,\middle\|\, \tfrac{1}{2}\right),$$

establishing the achievability part of the coding theorem for $\mathsf{BSC}(p)$.

Optimality

Is it true that $\lim_{n\to\infty} \frac{1}{n} k^*(n,\epsilon) = \mathsf{d_b}\left(p \,\middle\|\, \frac{1}{2}\right) = \mathrm{D}\left(\mathrm{Ber}(p) \,\middle\|\, \mathrm{Ber}(\frac{1}{2})\right)$?

The answer is **yes**. The divergence reminds us of binary hypothesis testing, and the proof below is built on the idea of *testing between two channel laws*.

For a given encoding/decoding pair $(\mathrm{enc}, \mathrm{dec})$ with $\mathsf{P}_\mathsf{e}^{(n)} \le \epsilon$, consider a binary hypothesis testing problem

$$\mathcal{H}_0: \quad W \to \boxed{\mathrm{enc}} \to \boldsymbol{X} \to \boxed{\mathsf{BSC}(p)^{\otimes n}} \to \boldsymbol{Y} \to \boxed{\mathrm{dec}} \to \hat{W}$$

$$\mathcal{H}_1: \quad W \to \boxed{\mathrm{enc}} \to \boldsymbol{X} \to \boxed{\mathsf{BSC}(1/2)^{\otimes n}} \to \boldsymbol{Y} \to \boxed{\mathrm{dec}} \to \hat{W}$$

with observation tuple $(W, \boldsymbol{X}, \boldsymbol{Y}, \hat{W})$ following P_0 and P_1 respectively.

Key observation 1: the second channel $\mathsf{BSC}(1/2)$ completely breaks the dependency between $\boldsymbol{Y}$ and $\boldsymbol{X}$, and hence the decoding performance is the same as pure random guess, that is, $\mathsf{P}_1\{W = \hat{W}\} = \frac{1}{2^k}$.

For the above HT problem, one can use a test that accept $\mathcal{H}_0$ iff $W = \hat{W}$, and this test has

$$\pi_{1|0}^{(n)} = \mathsf{P}_0\{W \neq \hat{W}\} \equiv \mathsf{P}_{\mathsf{e}}^{(n)} \leq \epsilon$$
$$\pi_{0|1}^{(n)} = \mathsf{P}_1\{W = \hat{W}\} = \tfrac{1}{2^k}.$$

With the language of binary HT, we immediately have $\frac{1}{2^k} \geq \varpi_{0|1}^{*}(n,\epsilon)$.

Key observation 2: the above binary HT problem is equivalent to

$$\mathcal{H}_0: \quad Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(p),\ i = 1, 2, \ldots, n$$
$$\mathcal{H}_1: \quad Z_i \overset{\text{i.i.d.}}{\sim} \mathrm{Ber}(\tfrac{1}{2}),\ i = 1, 2, \ldots, n.$$

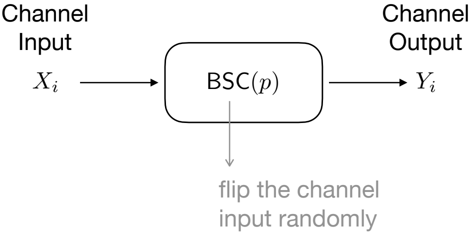
Hence, by the Chernoff-Stein Lemma, $\lim_{n\to\infty} -\frac{1}{n}\log \varpi_{0|1}^{*}(n,\epsilon) = \mathsf{d_b}\left(p \,\middle\|\, \tfrac{1}{2}\right)$.

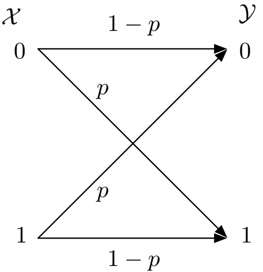
Combining everything together, we show that

$$\limsup_{n\to\infty} \frac{1}{n} k^*(n,\epsilon) \leq \lim_{n\to\infty} -\frac{1}{n}\log \varpi_{0|1}^{*}(n,\epsilon) = \mathsf{d_b}\left(p \,\middle\|\, \tfrac{1}{2}\right),$$

and the optimal rate for communication over the BSC is characterized.

Summary





For the binary symmetric channel with flip probability p,

$$\lim_{n\to\infty} \frac{1}{n} k^*(n,\epsilon) = \mathsf{d_b}\left(p \,\middle\|\, \tfrac{1}{2}\right) = 1 - \mathsf{H_b}(p) \quad \forall\, \epsilon \in (0,1).$$

where $k^*(n,\epsilon)$ is the smallest k such that there exists channel codes of codebook size 2^k and error probability $\le \epsilon$.

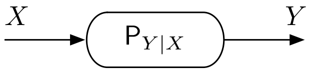
How to extend the result to general channels?

Outline

In this lecture, we will show that (for discrete memoryless channels) the fundamental limit is the channel capacity, a quantity that can be computed by maximizing the "mutual information" between the input and the output of the channel, when we want to reconstruct the bits with vanishing error probability.

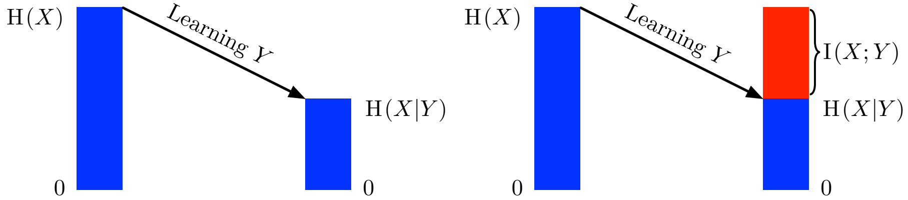
1. We begin with an intuitive motivation to introduce mutual information.
   - Motivation: single use of a channel.
   - Data processing inequality, chain rule, convexity.
2. Next we prove Shannon's noisy channel coding theorem.
   - Use Fano's inequality to prove the converse part.
   - Use typicality arguments to prove the achievability part.
   - Source-channel separation.

Consider a single use of the channel $\mathsf{P}_{Y|X}$:



$\mathrm{H}(X)$ quantifies the amount of uncertainty of the input X.

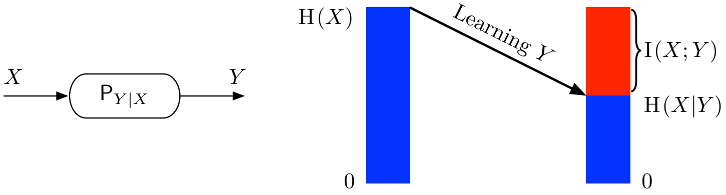
$\mathrm{H}(X|Y)$ quantifies the amount of uncertainty of X given the output Y.



Question: How much information does Y tell about X through the channel?

Answer: $\mathrm{H}(X) - \mathrm{H}(X|Y)$.

Mutual information



Definition 1 (Mutual Information)

For a pair of jointly distributed (X, Y), the mutual information between X and Y is defined as

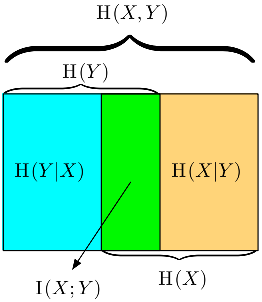
$$\mathrm{I}(X;Y) := \mathrm{H}(X) - \mathrm{H}(X|Y).$$

Interpretation: Infer information about the channel input X from output Y.

Proposition 1 (An Identity)

$$\mathrm{I}(X;Y) = \mathrm{H}(X) - \mathrm{H}(X|Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X)$$
$$= \mathrm{H}(X) + \mathrm{H}(Y) - \mathrm{H}(X,Y).$$

pf: By chain rule: $\mathrm{H}(X|Y) = \mathrm{H}(X,Y) - \mathrm{H}(Y)$. □



Note: Mutual information is symmetric, that is,

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

The mutual information between X and itself is equal to its entropy:

$$\mathrm{I}(X;X) = \mathrm{H}(X) \text{ since } \mathrm{H}(X|X) = 0.$$

Hence, the entropy is also called "self information" in some literatures.

Mutual information and information divergence

Proposition 2

For $(X, Y) \sim \mathsf{P}_{X,Y} = \mathsf{P}_X \mathsf{P}_{Y|X} = \mathsf{P}_Y \mathsf{P}_{X|Y}$*,*

$$\mathrm{I}(X;Y) = \mathsf{E}_{(X,Y)\sim\mathsf{P}_{X,Y}}\left[\log \frac{\mathsf{P}_{X,Y}(X,Y)}{\mathsf{P}_X(X)\mathsf{P}_Y(Y)}\right] = \mathrm{D}(\mathsf{P}_{X,Y} \| \mathsf{P}_X \times \mathsf{P}_Y),$$

where $\mathsf{P}_X \times \mathsf{P}_Y$ *is a product distribution of the two marginals* P_X *and* P_Y*.*

pf: A simple corollary of the identity in Proposition 1. □

Corollary 1 (Extremal Values of Mutual Information)

1. $\mathrm{I}(X;Y) \geq 0$*, with equality iff* X, Y *are independent.*
2. $\mathrm{I}(X;Y) \leq \mathrm{H}(X)$*, with equality iff* X *is a deterministic function of* Y*.*

pf: The proof of the first one is due to the fact that conditioning reduces entropy. The proof of the second one is due to $\mathrm{H}(X|Y) \geq 0$.

Interpretation: the mutual information between X and Y, $\mathrm{I}(X;Y)$ can also be viewed as a measure of the dependency between X and Y.

It is the divergence of the actual distribution $\mathsf{P}_{X,Y}$ from the independent distribution $\mathsf{P}_X \times \mathsf{P}_Y$.

- If X is determined by Y (highly dependent), $\mathrm{I}(X;Y)$ is maximized.
- If X is independent of Y (no dependency), $\mathrm{I}(X;Y) = 0$.

Exercise 1

Prove the following identity:

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

$\forall\, \mathsf{Q}_Y$ such that $\mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y) < \infty$. Furthermore,

$$\mathrm{I}(X;Y) = \min_{\mathsf{Q}_Y : \mathrm{D}(\mathsf{P}_Y \| \mathsf{Q}_Y) < \infty} \mathrm{D}\left(\mathsf{P}_{Y|X} \middle\| \mathsf{Q}_Y \middle| \mathsf{P}_X\right).$$

Example 1

Compute $\mathrm{I}(X_1; X_2)$ for $X_1, X_2 \in \{0, 1\}$ with joint PMF

(x_1, x_2)	$(0,0)$	$(0,1)$	$(1,0)$	$(1,1)$
$\mathsf{P}(x_1, x_2)$	$\frac{1}{6}$	$\frac{1}{3}$	$\frac{1}{3}$	$\frac{1}{6}$

sol: From the previous examples, we have

$$\mathrm{H}(X_1, X_2) = \log 3 + \tfrac{1}{3},\ \mathrm{H}(X_1) = \mathrm{H}(X_2) = 1,$$
$$\mathrm{H}(X_1|X_2) = \mathrm{H}(X_2|X_1) = \log 3 - \tfrac{2}{3}.$$

Hence, $\mathrm{I}(X_1; X_2) = \mathrm{H}(X_1) - \mathrm{H}(X_1|X_2) = \frac{5}{3} - \log 3$.

Conditional mutual information

Definition 2 (Conditional Mutual Information)

For a tuple of jointly distributed r.v.'s (X, Y, Z), the mutual information between X and Y given Z is defined as

$$\mathrm{I}(X;Y|Z) := \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z)\,.$$

Similar to the previous identity (Proposition 1), we have

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

Similar to Proposition **??**, we have

1. $\mathrm{I}(X;Y|Z) \geq 0$, with equality iff X, Y are independent given Z, that is, $X - Z - Y$ forms a Markov chain.
2. $\mathrm{I}(X;Y|Z) \leq \mathrm{H}(X|Z)$, with equality iff X is a (deterministic) function of Y, Z.

Chain rule for mutual information

Theorem 1 (Chain Rule for Mutual Information)

$$\mathrm{I}(X;Y^n) = \sum_{i=1}^{n} \mathrm{I}\left(X;Y_i \middle| Y^{i-1}\right).$$

pf: Proved by definition and the chain rule for entropy. $\square$

Exercise 2

Show that

$$\mathrm{I}(X;Z) \le \mathrm{I}(X;Y,Z) \quad \text{and} \quad \mathrm{I}(X;Y|Z) \le \mathrm{I}(X;Y,Z)\,.$$

Data processing inequality

Theorem 2 (Data Processing Inequality)

For a Markov chain $X - Y - Z$, that is, $\mathsf{P}_{X,Y,Z} = \mathsf{P}_X \mathsf{P}_{Y|X} \mathsf{P}_{Z|Y}$, we have

$$\mathrm{I}(X;Y) \geq \mathrm{I}(X;Z)\,.$$

Interpretation: $X - Y - Z$ says that the information of X that Z can provide is contained in Y. Hence, the amount of information of X that can be inferred by Z is not greater than that can be inferred by Y.

pf: Since $X - Y - Z$, we have $\mathrm{I}(X;Z|Y) = 0$. Hence,

$$\begin{aligned}
&\mathrm{I}(X;Y,Z) = \mathrm{I}(X;Y) + \mathrm{I}(X;Z|Y) = \mathrm{I}(X;Y) && (\because \mathrm{I}(X;Z|Y) = 0) \\
&\mathrm{I}(X;Y,Z) = \mathrm{I}(X;Z) + \mathrm{I}(X;Y|Z) && \text{(Chain Rule)} \\
\Longrightarrow\ &\mathrm{I}(X;Y) = \mathrm{I}(X;Z) + \mathrm{I}(X;Y|Z) \geq \mathrm{I}(X;Z)\,.
\end{aligned}$$

□

Markov chains are common in communication systems. For example, in channel coding (without feedback), the message W, the channel input X^n, the channel output Y^n, and the decoded message $\hat{W}$ form a Markov chain $W - X^n - Y^n - \hat{W}$ (see the figure above).

Data processing inequality turns out to be crucial in obtaining impossibility results in information theory.

Exercise 3 (Functions of R.V.)

For $Z = g(Y)$ being a deterministic function of Y, show that

$$\mathrm{H}(Y) \geq \mathrm{H}(Z) \quad \text{and} \quad \mathrm{I}(X;Y) \geq \mathrm{I}(X;Z).$$

Exercise 4

Show that

$$X_1 - X_2 - X_3 - X_4 \implies \mathrm{I}(X_1;X_4) \leq \mathrm{I}(X_2;X_3).$$

Example 2

Consider two random variables $X_1, X_2 \in \{0,1\}$ with the same joint PMF as that in Example 1. Let $X_3 = X_2 \oplus Z$, where $Z \sim \mathrm{Ber}(p)$ and Z is independent of (X_1, X_2).

1. Compute $\mathrm{I}(X_1; X_3)$ and $\mathrm{I}(X_1; X_2|X_3)$.
2. Show that $X_1 - X_2 - X_3$ forms a Markov chain.
3. Verify the data processing inequality $\mathrm{I}(X_1; X_2) \geq \mathrm{I}(X_1; X_3)$.

sol:

(x_1, x_2, x_3)	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(0,1,1)$
$\mathsf{P}(x_1, x_2, x_3)$	$\frac{1}{6}(1-p)$	$\frac{1}{6}p$	$\frac{1}{3}p$	$\frac{1}{3}(1-p)$
(x_1, x_2, x_3)	$(1,0,0)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$
$\mathsf{P}(x_1, x_2, x_3)$	$\frac{1}{3}(1-p)$	$\frac{1}{3}p$	$\frac{1}{6}p$	$\frac{1}{6}(1-p)$

Then it is straightforward to compute mutual informations and verify the Markov chain $X_1 - X_2 - X_3$.

Conditioning reduces mutual information?

Sometimes yes, sometimes no.

Proposition 3 (Conditioning May Decrease Mutual Information)

For a Markov chain $X - Y - Z$, we have $\mathrm{I}(X;Y) \ge \mathrm{I}(X;Y|Z)$.

pf: The same argument as that of the data processing inequality. $\square$

Example 3 (Conditioning May Increase Mutual Information)

Let X and Y be i.i.d. $\mathrm{Ber}(\frac{1}{2})$ random variables, and $Z = X \oplus Y$. Evaluate $\mathrm{I}(X;Y|Z)$ and show that $\mathrm{I}(X;Y|Z) > \mathrm{I}(X;Y)$.

sol:
$$\begin{aligned}\mathrm{I}(X;Y|Z) &= \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z) = \mathrm{H}(X|Z) - \mathrm{H}(X|Y,X \oplus Y)\\ &= \mathrm{H}(X|Z) - \mathrm{H}(X|Y,X) = \mathrm{H}(X|Z) = \mathrm{H}(X) = 1.\end{aligned}$$

(note that X and Z are independent)

On the other hand, $\mathrm{I}(X;Y) = 0$. Hence, $1 = \mathrm{I}(X;Y|Z) > \mathrm{I}(X;Y) = 0$.

Convexity and concavity of mutual information

Theorem 3

Let $(X,Y) \sim \mathsf{P}_{X,Y} = \mathsf{P}_X \mathsf{P}_{Y|X}$*.*

1. *With* $\mathsf{P}_{Y|X}$ *fixed,* $\mathrm{I}(X;Y)$ *is a concave function of* P_X*.*
2. *With* P_X *fixed,* $\mathrm{I}(X;Y)$ *is a convex function of* $\mathsf{P}_{Y|X}$*.*

pf: For the first property, note that

$$\mathrm{I}(X;Y) = \mathrm{H}(Y) - \mathrm{H}(Y|X) = \mathrm{H}(\mathsf{P}_Y) - \underbrace{\textstyle\sum_{x\in\mathcal{X}} \mathsf{P}_X(x)\mathrm{H}(Y|X=x)}_{\text{linear in } \mathsf{P}_X}.$$

It suffices to show that $\mathrm{H}(\mathsf{P}_Y)$ is concave in P_X.

P_Y is a linear function of P_X and $\mathrm{H}(\mathsf{P}_Y)$ is a concave function of P_Y. As a result, $\mathrm{H}(\mathsf{P}_Y)$ is concave in P_X, which completes the proof of concavity.

For the convexity, let us consider two conditional distributions $\mathsf{P}^{(0)}_{Y|X}, \mathsf{P}^{(1)}_{Y|X}$, and random variables X, Y, Θ such that $\mathsf{P}_{\Theta,X,Y} = \mathsf{P}_X \mathsf{P}_\Theta \mathsf{P}_{Y|X,\Theta}$, $\Theta \sim \mathrm{Ber}(\lambda)$, and $\mathsf{P}_{Y|X,\Theta}(y\,|\,x,\theta) = \mathsf{P}^{(\theta)}_{Y|X}(y\,|\,x)$, $\forall\, y \in \mathcal{Y},\ x \in \mathcal{X},\ \theta \in \{0,1\}$.

Hence, $\Theta \perp\!\!\!\perp X$ and $\mathsf{P}_{Y|X} = (1-\lambda)\mathsf{P}^{(0)}_{Y|X} + \lambda \mathsf{P}^{(1)}_{Y|X}$.

By the basic properties of mutual information and entropy,

$$\mathrm{I}(X;Y) \le \mathrm{I}(X;Y,\Theta) = \mathrm{I}(X;Y|\Theta) + \mathrm{I}(X;\Theta)\,.$$

Note that $\Theta \perp\!\!\!\perp X \implies \mathrm{I}(X;\Theta) = 0$. Also note that

- $\mathrm{I}(X;Y) = \mathrm{I}(X;Y)\big|_{(X,Y)\sim \mathsf{P}_X\left\{(1-\lambda)\mathsf{P}^{(0)}_{Y|X} + \lambda \mathsf{P}^{(1)}_{Y|X}\right\}}$.
- $\mathrm{I}(X;Y|\Theta) = (1-\lambda)\mathrm{I}(X;Y|\Theta=0) + \lambda \mathrm{I}(X;Y|\Theta=1)$ and

$$\mathrm{I}(X;Y|\Theta=\theta) = \mathrm{I}(X;Y)\big|_{(X,Y)\sim \mathsf{P}_X \mathsf{P}^{(\theta)}_{Y|X}},\ \theta \in \{0,1\}.$$

The proof is complete. $\square$

Summary: mutual Information

- $\mathrm{I}(X;Y) := \mathrm{H}(X) - \mathrm{H}(X|Y)$ measures the amount of information of X contained in Y (and vice versa). $\mathrm{I}(X;Y) = \mathrm{D}(\mathsf{P}_{X,Y} \| \mathsf{P}_X \times \mathsf{P}_Y)$ also tells the level of dependency between X and Y.
- $\mathrm{I}(X;Y|Z) := \mathrm{H}(X|Z) - \mathrm{H}(X|Y,Z)$ measures the amount of information of X in Y given Z.
- Nonnegative; Concave function of P_X with fixed $\mathsf{P}_{Y|X}$; Convex function of $\mathsf{P}_{Y|X}$ with fixed P_X.
- Chain rule: $\mathrm{I}(X;Y^n) = \sum_{i=1}^{n} \mathrm{I}\left(X;Y_i \middle| Y^{i-1}\right)$.
- Conditioning may reduce MI: $X - Y - Z \implies \mathrm{I}(X;Y|Z) \le \mathrm{I}(X;Y)$.
- Data processing decreases MI: $X - Y - Z \implies \mathrm{I}(X;Y) \ge \mathrm{I}(X;Z)$.

Next: noisy channel coding theorem.