Lecture 6: Shannon Theory

Claude Shannon (1948)

1. How much can a message be compressed?

2. At what rate can one communicate reliably over a noisy channel?

We will model the message as a random variable

$$X := \{ x, p(x) \}$$

letter   probability

$$x \in \{0, 1, \ldots, d-1\}$$

alphabet

$p(x) \in [0, 1]$

$$\sum_x p(x) = 1$$

n - letter message

$x_1 \ x_2 \ \cdots \ x_n$

has probability $\prod_{i=1}^{n} p(x_i)$

if it is independently and identically distributed (iid).

We can represent this as a random variable

$$X^n = \{ \underline{x}, p(\underline{x}) \}$$

$(x_1, x_2, \ldots, x_n)$ $\quad$ $p(x_1)\,p(x_2)\ldots p(x_n)$

Consider a long message

i.e. $n \gg 1$.

Can we compress such a message?

Consider a binary alphabet $x \in \{0, 1\}$

$p(0) = 1 - p \qquad p \in [0,1]$

$p(1) = p$

Law of large numbers:

Typical strings have $n(1-p)$ 0's

and np 1's.

# of possible strings

$$\binom{n}{np} = \frac{n!}{(np)!\,(n(1-p))!} = \binom{n}{n(1-p)}$$

Recall Stirling's approximation

$$\log n! = n \log n - n + O(\log n)$$

$$\log \binom{n}{np}$$

$$= \log \left(\frac{n!}{(np)!\,(n(1-p))!} \right)$$

$$= \log n! - \log (np)! - \log (n(1-p))!$$

$$\approx n \log n - \cancel{n} - np \log np + \cancel{np}$$

$$- n(1-p) \log (n(1-p)) + n(\cancel{1}-\cancel{p})$$

$$= n \left(\cancel{\log} n - p \cancel{\log} n - p \log p \right.$$

$$\left. - (\cancel{1} - \cancel{p}) \log n - (1-p) \log (1-p) \right)$$

$$= n \left(-p \log p - (1-p) \log (1-p) \right)$$

$$= n H(p)$$

where $H(p)$ is the binary entropy

Stirling's approximation is for the natural logarithm but we will use log base 2 (convenient for binary alphabet).

Making this change just scales the log by a constant, which is not significant.

The # of typical strings is therefore

$$\boxed{2^{nH(p)}}$$

We can represent the typical messages using only

$nH(p) + \delta$ bits rather than n

As we will see, the probability that the message is atypical is negligible in the limit $n \to \infty$.

Note that $H(p) \in [0, 1]$

$H(p) = 1$ for $p = 1/2$ only

So for any distribution other than the uniform distribution, we can compress the message

Let's make this more precise

Consider $X = \{x, p(x)\}$ Random Variable (RV)

$x \in \{0, 1, \ldots, d-1\}$

d-letter alphabet again

$$\mathbb{E}_X[f(x)] = \sum_x f(x)\, p(x)$$

Expectation value of $f(x)$

$$\mu[X] = \mathbb{E}_X[x] = \sum_x x\, p(x)$$

Strong law of large numbers

For any $\varepsilon, \delta > 0$ $\exists N$ s.t.

$$\left| \frac{1}{n} \sum_{i=1}^{n} x_i - \mu[X] \right| \leq \delta$$

w/ probability at least $1-\varepsilon$

for all $n \geq N$.

Def : Shannon entropy of an RV

$$H(X) = \mathbb{E}_X\left[\log_2 \frac{1}{p(x)}\right]$$

$$= -\sum_x p(x) \log_2 p(x)$$

Def : a sequence of n letters is **δ-typical** if

$$H(X) - \delta \leq -\frac{1}{n}\log_2 p(x_1 \dots x_n) \leq H(X) + \delta$$

$$-(H(X)-\delta) \geq \frac{1}{n}\log_2 p(-) \geq -(H(X)+\delta)$$

Lemma

For any $\varepsilon, \delta > 0$ $\exists N$ s.t. all sequences of $n \geq N$ letters are δ-typical w/ probability $1-\varepsilon$.

Proof

Define RV $Y = \{\underset{y}{\log_2(1/p(x))}, \underset{p(y)}{p(x)}\}$

Strong law of large numbers $\exists N$ s.t.

$$\left| \frac{1}{n}\sum_{i=1}^{n} y_i - \mu[Y] \right| \leq \delta$$

w/ probability $1-\varepsilon$ $\forall n \geq N$

$$\mu[Y]$$

$$= \mathbb{E}_Y[y]$$

$$= \sum_y y\, p(y)$$

$$= \sum_x \log_2(1/p(x))\; p(x)$$

$$= \mathbb{E}_X[\log_2(1/p(x))] = H(X)$$

$$\left| \frac{1}{n} \sum_{i=1}^{n} \log(1/p(x_i)) - H(X) \right| \leq \delta$$

$$\frac{1}{n} \sum_{i=1}^{n} \log(1/p(x_i)) - H(X) \leq \delta$$

$$-\frac{1}{n} \sum_{i=1}^{n} \log(p(x_i)) \leq H(X) + \delta$$

$$-\frac{1}{n} \log p(x_1 x_2 \dots x_n) \leq H(X) + \delta$$

For the lower bound the manipulation is analogous. $\square$

This lemma tells us that each typical n-letter sequence $\underline{x} = (x_1, x_2, \ldots, x_n)$ occurs w/ probability $p(\underline{x})$ satisfying

$$p_{min} = 2^{-n(H(X)+\delta)}$$

$$\leq p(\underline{x})$$

$$\leq 2^{-n(H(X)-\delta)} = p_{max}$$

(Multiply δ-typical inequality by -1) and raise to power 2

The number of typical sequences

$N_{typ}(\varepsilon, \delta, n)$ is bounded

$$N_{typ} \, p_{min} \leq \sum_{\substack{\text{typical} \\ \underline{x}}} p(\underline{x}) \leq 1$$

$$N_{typ} \leq \frac{1}{p_{min}} = 2^{n(H(X)+\delta)}$$

$$N_{typ} \, p_{max} \geq \sum_{\substack{\text{typical} \\ \underline{x}}} p(\underline{x}) \geq 1 - \varepsilon$$

$$N_{typ} \geq \frac{1}{p_{max}}(1-\varepsilon) = (1-\varepsilon) 2^{n(H(X)-\delta)}$$

$$2^{n(H(X)+\delta)} \geq N_{typ}(\varepsilon, \delta, n)$$

$$\geq (1-\varepsilon) 2^{n(H(X)-\delta)}$$

Therefore we can encode all typical sequences using just $n(H(X)+\delta)$ bits, or equivalently $H(X)+\delta$ bits per letter.

This will only fail for atypical sequences, which occur w/ probability ε. Therefore the success probability is $1-\varepsilon$.

Suppose we try instead to use only $H(X)-\delta'$ bits per letter.

$\delta' > \delta$

Probability that we encounter a typical sequence we can encode is upper bounded by

$$\frac{\#\text{ of sequences we encode}}{(1-\varepsilon)\, 2^{n(H(X)+\delta)}}$$

$\uparrow$ lower bound on N_{typ}

$$= \frac{2^{n(H(X)-\delta')}}{(1-\varepsilon)\, 2^{n(H(X)-\delta)}}$$

$$= \frac{2^{-n(\delta'-\delta)}}{(1-\varepsilon)}$$

$\delta' - \delta > 0$

$\Rightarrow$ Exponentially small in n

Compression rate

$$R = \frac{m}{n}$$

$\leftarrow$ bits encoded per letter

Theorem : Source coding (Shannon)

Compression rate

$R = H(X) + o(1)$ is achievable

$R = H(X) - \Omega(1)$ is not achievable

Aside : Big O notation

Intuition

$f(x) = O(g(x))$

$f(x) \lesssim g(x)$ asymptotically

$f(x) = o(g(x))$

$f(x) < g(x)$ asymptotically

$f(x) = \Omega(g(x))$

$f(x) \geq g(x)$ asymptotically

$f(x) = \omega(g(x))$

$f(x) > g(x)$ asymtotically

$f(x) = O(g(x))$

& $f(x) = \Omega(g(x))$

then $f(x) = \Theta(g(x))$

<u>Formally</u>

$f(x) = O(g(x))$

$\Rightarrow \exists\, x_0 \;,\; a > 0$ s.t.

$\forall\, x \geq x_0 \quad f(x) \leq a\, g(x)$ etc.

$$f(x) = o(1)$$

$$\forall \varepsilon > 0 \quad \exists x_0$$

$$f(x) \leq \varepsilon \quad \forall x \geq x_0$$

$$f(x) = \Omega(1)$$

$$\exists x_0, \quad a > 0 \quad \text{s.t.}$$

$$\forall x \geq x_0 \quad f(x) \geq a$$

Note that we did not discuss how to do the compression, we just argued that it is possible to achieve a certain compression rate. Studying this is a entire topic unto itself and beyond the scope of this course.

Noisy channel coding

Suppose Alice wants to send information to Bob over a noisy communication channel.

What is the maximal communication rate that she can achieve?

Binary symmetric channel

$$p(0 \mid 0) = 1 - p = p(1 \mid 1)$$

B gets A sends

$$p(0 \mid 1) = p = p(1 \mid 0)$$

Alice uses the channel n times to send a message to Bob.

She chooses 2^k codeword

strings from the possible 2^n strings of length n.

The encoding rate $R = \frac{k}{n}$

How to ensure successful transmission?

Choose codewords whose Hamming distance from each other is large.

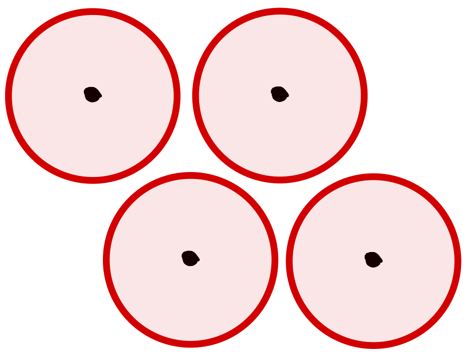
Hamming distance between $\underline{x}$ & $\underline{y}$ is the number of bits we need to flip to turn $\underline{x}$ into $\underline{y}$.

e.g. $\underline{x} = 01101$ $\quad$ Hamming distance $= 2$

$\underline{y} = 11100$

Expected # of bit-flips is np

For a given input codeword the output is one of $2^{nH(p)}$ typical strings w/ high probability.



Choose codewords • such that each error sphere ○ contains around $2^{nH(p)}$ and error spheres for different codewords are distinct.

# of codewords $2^k = 2^{nR}$

volume of error sphere $2^{nH(p)}$

To construct the code we need

$$2^{nR}\, 2^{nH(p)} \leq 2^n$$

$$R \leq 1 - H(p) := C(p)$$

channel capacity

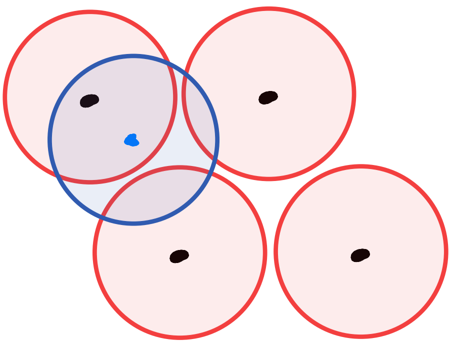
Is this rate achievable?

Yes using random codes.

Suppose that Z is the uniformly random distribution (for a single bit).

Sample from Z^n a total of 2^{nR} times to generate 2^{nR} random codewords.

To send a message, Alice chooses one of these codewords and sends it to Bob using the channel n times. To decode Bob draws a Hamming sphere with radius $np + \delta$ around his received string. If the sphere contains a unique codeword he decodes accordingly. Otherwise he picks one of the options at random.



For any $\delta > 0$ Bob's Hamming sphere contains Alice's codeword w/ high probability. What is the probability that it also contains another codeword?

possible strings 2^n

strings in Bob's Hamming sphere $2^{n(H(p)+\delta)}$

Prob. that a given string is in Bob's Hamming sphere

$$\frac{2^{n(H(p)+\delta)}}{2^n} = 2^{-n(C(p)-\delta)}$$

codewords 2^{nR}

Codewords are uniformly random, so the probability that Bob's Hamming sphere contains another codeword is upper bounded by

$$2^{nR}\, 2^{-n(C(p)-\delta)}$$

$$= 2^{-n(C(p)-R-\delta)}$$

Choose $R = C(p) - \text{const.}$

to make this arbitrarily small as $n \to \infty$.

So far we have shown that for a random code, a randomly chosen codeword will be decoded successfully with high probability when sent over the channel.

$$\frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} p_i^{\text{error}} \leq \varepsilon$$

Let $N_{2\varepsilon}$ = # of codewords w/ $p_i^{\text{error}} \geq 2\varepsilon$

$$\frac{1}{2^{nR}} N_{2\varepsilon}\, 2\cancel{\varepsilon} \leq \cancel{\varepsilon}$$

$$N_{2\varepsilon} \leq 2^{nR-1}$$

So if we throw away half the codewords then we are guaranteed to have error less than 2ε for all remaining codewords.

New code has rate

$$R' = R - \frac{1}{n}$$ $\frac{1}{n} = o(1)$

$$\Rightarrow \boxed{R = C(p) - o(1)}$$

is achievable where

$$C(p) = 1 - H(p).$$

If we pick a sequence of random codes then we will achieve this performance w/ high probability. Then there must exist a particular sequence of codes that achieves the desired performance.

For RV $X = \{x, p(x)\}$

there always exists

x' s.t. $\mathbb{E}_X[x] \leq x'$

& x'' s.t. $x'' \leq \mathbb{E}_X[x]$

We now consider two RVs X e Y that may be correlated. We can write the joint distribution

$$XY = \{(x,y), p(x,y)\}$$

The marginal distribution

$$X = \{x, p(x) = \sum_y p(x,y)\}.$$

Suppose we sample from XY n-times, giving a message

$$(\underline{x}, \underline{y}) = (x_1 x_2 \ldots x_n y_1 y_2 \cdots y_n)$$

$$p(\underline{x}, \underline{y}) = p(x_1,y_1)\, p(x_2,y_2) \cdots p(x_n,y_n)$$

We say that $(\underline{x}, \underline{y})$ is jointly δ-typical if

$$2^{-n(H(X)+\delta)} \leq p(\underline{x}) \leq 2^{-n(H(X)-\delta)}$$

$$2^{-n(H(Y)+\delta)} \leq p(\underline{y}) \leq 2^{-n(H(Y)-\delta)}$$

$$2^{-n(H(XY)+\delta)} \leq p(\underline{x}, \underline{y}) \leq 2^{-n(H(XY)-\delta)}$$

Strong law of large numbers implies for any $\varepsilon, \delta > 0$ $\exists N$ s.t. $\forall n \geq N$ such that $(\underline{x}, \underline{y})$ is jointly δ-typical w/ probability at least $1 - \varepsilon$.

Using Bayes's rule we can derive expressions for the conditional probabilities

$$p(\underline{x} \mid \underline{y}) = \frac{p(\underline{x}, \underline{y})}{p(\underline{y})}$$

$$\geqslant \frac{2^{-n(H(XY)+\delta)}}{2^{-n(H(Y)-\delta)}}$$

$$= 2^{-n(H(X \mid Y) + 2\delta)}$$

$$p(\underline{x} \mid \underline{y}) \leqslant \frac{2^{-n(H(XY)-\delta)}}{2^{-n(H(Y)+\delta)}}$$

$$= 2^{-n(H(X \mid Y) - 2\delta)}$$

where we have introduced the conditional entropy of X given Y.

$$H(X|Y) = H(XY) - H(Y)$$

This quantifies the remaining uncertainty I have about x once I know y.

If $(\underline{x}, \underline{y})$ is jointly δ-typical then $H(X|Y) + o(1)$ bits are needed to specify $\underline{x}$ once $\underline{y}$ is known (with probability $1-\varepsilon$).

The information about x that I gain when I learn y is the

mutual information

$$I(X;Y) = H(X) - H(X|Y)$$

$$= H(X) - (H(XY) - H(Y))$$

$$= H(Y) - H(Y|X)$$

This quantifies how much X & Y are correlated.

Noisy channel coding

General case

Alphabet $\{0, 1, \ldots, d-1\}$

Channel $p(y \mid x)$

Bob gets ↗ ↖ Alice sends

Again, A & B use a random code.

① Choose some distribution

$X = \{x, p(x)\}$

② Generate a codeword by sampling from X n times

③ Repeat ② 2^{nR} times $R = k/n$

How does Bob decode?

- He gets message $\underline{y}$.
- He checks whether a codeword $\underline{x}$ exists such that $\underline{x}$ & $\underline{y}$ are jointly typical.
- If $\underline{x}$ exists & is unique then he outputs $\underline{x}$
- Otherwise he chooses at random.

We now bound the probability of a decoding error p^{error}.

The input distribution and the channel determine the joint distribution XY.

$$X = \{ x, p(x) \}$$

$$Y = \{ y, p(y) = \sum_x p(x,y) \}$$

$$\overset{\text{Bayes}}{=} \sum_x p(y|x)\, p(x)$$

For n uses of the chanel we get the distribution $X^n Y^n$, as the codewords are randomly sampled from X.

By the strong law of large numbers, for $\varepsilon, \delta > 0$ & $n \geq N$ a sequence drawn from $X^n Y^n$

will be jointly δ-typical w/ probability $1-\varepsilon$.

So w/ prob. $1-\varepsilon$ Bob's received vector $\underline{y}$ will be jointly δ-typical w/ the codeword $\underline{x}$.

But are there any other codewords that are jointly δ-typical w/ $\underline{y}$?

Let $\underline{x}' \neq \underline{x}$ denote another codeword.

$\underline{x}'$ is sampled independently from $\underline{x}$, so $\underline{x}'$ is independent of $\underline{y}$.

$$p(\underline{x},\underline{y}) \leq 2^{-n(H(XY)-\delta)}$$

$$1 \geq \sum_{\substack{\underline{x},\underline{y} \\ \text{jointly } \delta\text{-typical}}} p(\underline{x},\underline{y}) \geq N_{jt}\, 2^{-n(H(XY)-\delta)}$$

$$N_{jt} \leq 2^{n(H(XY)-\delta)}$$

$$p(\underline{y}) \leq 2^{-n(H(Y)-\delta)}$$

$$p(\underline{x}') \leq 2^{-n(H(X)-\delta)}$$

$$\sum_{\substack{\underline{x}',\underline{y} \\ j.\,\delta\text{-typ.}}} p(\underline{x}',\underline{y}) = \sum_{\substack{\underline{x}',\underline{y} \\ j.\,\delta\text{-typ.}}} p(\underline{x}')\,p(\underline{y})$$

$$\leq N_{jt}\, 2^{-n(H(X)-\delta)}\, 2^{-n(H(Y)-\delta)}$$

$$\leq 2^{n(H(XY)-H(X)-H(Y)-3\delta)}$$

$$= 2^{n(I(X;Y)-3\delta)}$$

The code has $k = nR$ codewords so the probability that any other codeword except $\underline{x}$ is jointly δ-typical w/ $\underline{y}$ is upper bounded by

$$2^{nR} \, 2^{-n(I(X;Y) - 3\delta)}$$

$$= 2^{n(R - I(X;Y) + 3\delta)}$$

rate

Choose $R = I(X;Y) - c - 3\delta$

then the probabity of error is

$$p^{\text{error}} = \underset{\underline{x},\underline{y} \text{ not jt}}{\varepsilon} + (1-\varepsilon)\,2^{-nc} \quad \begin{matrix}\underline{x}, \dots, \underline{x}' \\ \text{jt w/ } \underline{y}\end{matrix}$$

We can make this arbitrarily close to 0 as we increase n.

We have actually bounded the average error probability

$$\frac{1}{2^{nR}} \sum_{i=1}^{2^{nR}} p_i^{\text{error}} \leq \varepsilon + (1-\varepsilon) 2^{-nc}$$

$$= \varepsilon'$$

We can again prune the code. Let $N_{2\varepsilon'}$ denote the # of codewords w/ $p_i^{\text{error}} \geq 2\varepsilon'$

$$\frac{1}{2^{nR}} N_{2\varepsilon'} 2\varepsilon' \leq \varepsilon'$$

$$N_{2\varepsilon'} \leq 2^{nR-1}$$

Discard $1/2$ of the codewords to achieve $p^i_{\text{error}} \leq 2\varepsilon' \ \forall i$.

The new code has rate

$$R' = R - \frac{1}{n}$$

So we can conclude that

$R' = I(X;Y) - o(1)$ is achievable.

We are free to choose X so the channel capacity is

$$\boxed{C := \max_{X} I(X;Y)}$$

This only depends on the probabilities $p(y|x)$ that define the channel.

So we can achieve any $R < C$.

Can we do better?

Consider the uniform distribution over codewords

$$\tilde{X}^n = \{ \underline{\tilde{x}},\ p(\underline{\tilde{x}}) = 2^{-nR} \}$$

codeword

$$H(\tilde{X}^n) = - \sum_{\underline{\tilde{x}}} p(\underline{\tilde{x}}) \log_2 p(\underline{\tilde{x}})$$

$$= nR \sum_{\underline{\tilde{x}}} p(\underline{\tilde{x}}) = nR$$

$$\tilde{Y}^n = \{ \underline{\tilde{y}},\ p(\underline{\tilde{y}}) = \sum_{\underline{\tilde{x}}} p(\underline{\tilde{y}} | \underline{\tilde{x}})\, p(\underline{\tilde{x}}) \}$$

$$= 2^{-nR} \sum_{\underline{\tilde{x}}} p(\underline{\tilde{y}} | \underline{\tilde{x}})$$

The channel acts on the lettes of $\underline{\tilde{x}}$ independently so

$$p(\underline{\tilde{y}} \mid \underline{\tilde{x}})$$

$$= p(\tilde{y}_1 \mid \tilde{x}_1) \cdots p(\tilde{y}_n \mid \tilde{x}_n)$$

$$H(\tilde{Y}^n \mid \tilde{X}^n) = \mathbb{E}_{\tilde{X}^n \tilde{Y}^n}\left[-\log_2 p(\underline{\tilde{y}} \mid \underline{\tilde{x}})\right]$$

$$= -\sum_{\underline{\tilde{x}}, \underline{\tilde{y}}} p(\underline{\tilde{x}}, \underline{\tilde{y}}) \log_2 p(\underline{\tilde{y}} \mid \underline{\tilde{x}})$$

$$= -\sum_{\underline{\tilde{x}}, \underline{\tilde{y}}} p(\underline{\tilde{x}}, \underline{\tilde{y}}) \log_2 \prod_i p(\tilde{y}_i \mid \tilde{x}_i)$$

$$= -\sum_{\underline{\tilde{x}}, \underline{\tilde{y}}} p(\underline{\tilde{x}}, \underline{\tilde{y}}) \sum_i \log_2 p(\tilde{y}_i \mid \tilde{x}_i)$$

$$= -\sum_i \sum_{\underline{\tilde{x}}, \underline{\tilde{y}}} p(\underline{\tilde{x}}, \underline{\tilde{y}}) \log_2 p(\tilde{y}_i \mid \tilde{x}_i)$$

Consider specific i

$$\sum_{\tilde{\underline{x}},\, \tilde{\underline{y}}} p(\tilde{\underline{x}}, \hat{\underline{y}}) \log_2 p(\tilde{y}_i \mid \tilde{x}_i)$$

$$\sum_{\tilde{x}_1} \cdots \sum_{\tilde{x}_n} \sum_{\tilde{y}_1} \cdots \sum_{\hat{y}_n} p\big((\tilde{x}_1 \dots \tilde{x}_n), (\tilde{y}_1 \dots \hat{y}_n)\big) \log_2 p(\tilde{y}_i \mid \hat{x}_i)$$

$$= \sum_{\tilde{x}_i,\, \tilde{y}_i} \log_2 p(\hat{y}_i \mid \tilde{x}_i) \times \sum_{\tilde{x}_{-i},\, \tilde{y}_{-i}} p(\underline{x}, \underline{y})$$

$$= \sum_{\tilde{x}_i,\, \tilde{y}_i} p(\hat{x}_i, \tilde{y}_i) \log_2 p(\tilde{y}_i \mid \tilde{x}_i)$$

$$= -H(\tilde{Y}_i \mid \hat{X}_i)$$

$$H(\hat{Y}^n \mid \tilde{X}^n) = \sum_i H(\tilde{Y}_i \mid \tilde{X}_i)$$

Shannon entropy is subadditive

$$H(\tilde{Y}^n) = H(\hat{Y}_1 \dots \hat{Y}_n) \leq \sum_i H(\tilde{Y}_i)$$

$$I(\tilde{Y}^n; \hat{X}^n) = H(\tilde{Y}^n) - H(\tilde{Y}^n | \tilde{X}^n)$$

$$\leq \sum_i H(\tilde{Y}_i) - H(\tilde{Y}_i | \tilde{X}_i)$$

$$= \sum_i I(\tilde{Y}_i; \hat{X}_i) \leq nC$$

$$I(\tilde{Y}^n; \tilde{X}^n) = I(\hat{X}^n; \tilde{Y}^n)$$

$$= H(\hat{X}^n) - H(\hat{X}^n | \bar{Y}^n)$$

$$= nR - H(\hat{X}^n | \tilde{Y}^n) \leq nC$$

If Bob can decode reliably then

$$\lim_{n \to \infty} \frac{1}{n} H(\hat{X}^n | \hat{Y}^n) = 0$$

The received vector determines the sent codeword.

$$\Rightarrow \quad \boxed{R \leq C + o(1)}$$

Two things to note

① The formula for the capacity $C = \max_X I(X;Y)$ is a single-letter formula i.e. it depends only on a single use of the channel but applies to arbitrarily long messages. We can often compute the capacity.

(2) The random codes method is <u>not efficient</u>. Encoding and decoding require an exponentially large code book. Finding efficient codes that achieve the capacity is highly non-trivial. For the BSC this was only achieved in the 90s, ~ 50 years after Shannon's paper.