08

Communication over a Noisy Channel

Notice

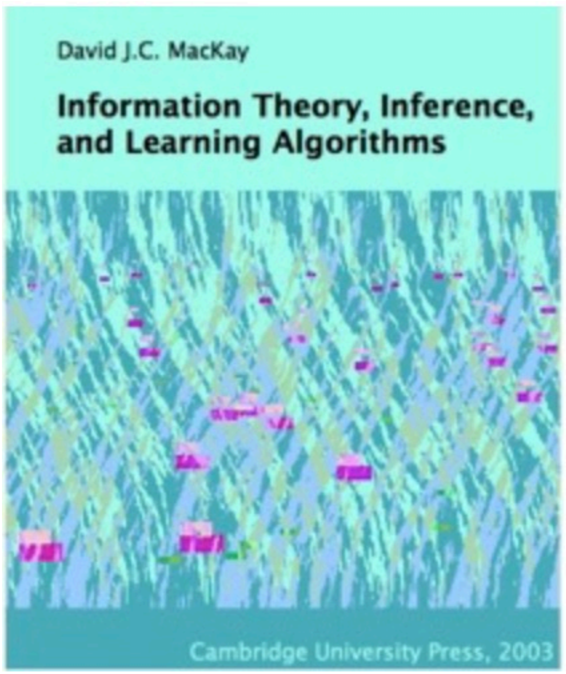
- **Author**
  - **João Moura Pires (jmp@fct.unl.pt)**

- **This material can be freely used for personal or academic purposes without any previous authorization from the author, provided that this notice is maintained/kept.**

- **For commercial purposes the use of any part of this material requires the previous authorization from the author.**

Bibliography

- **Many examples are extracted and adapted from:**



Information Theory, Inference, and Learning Algorithms
David J.C. MacKay
2005, Version 7.2

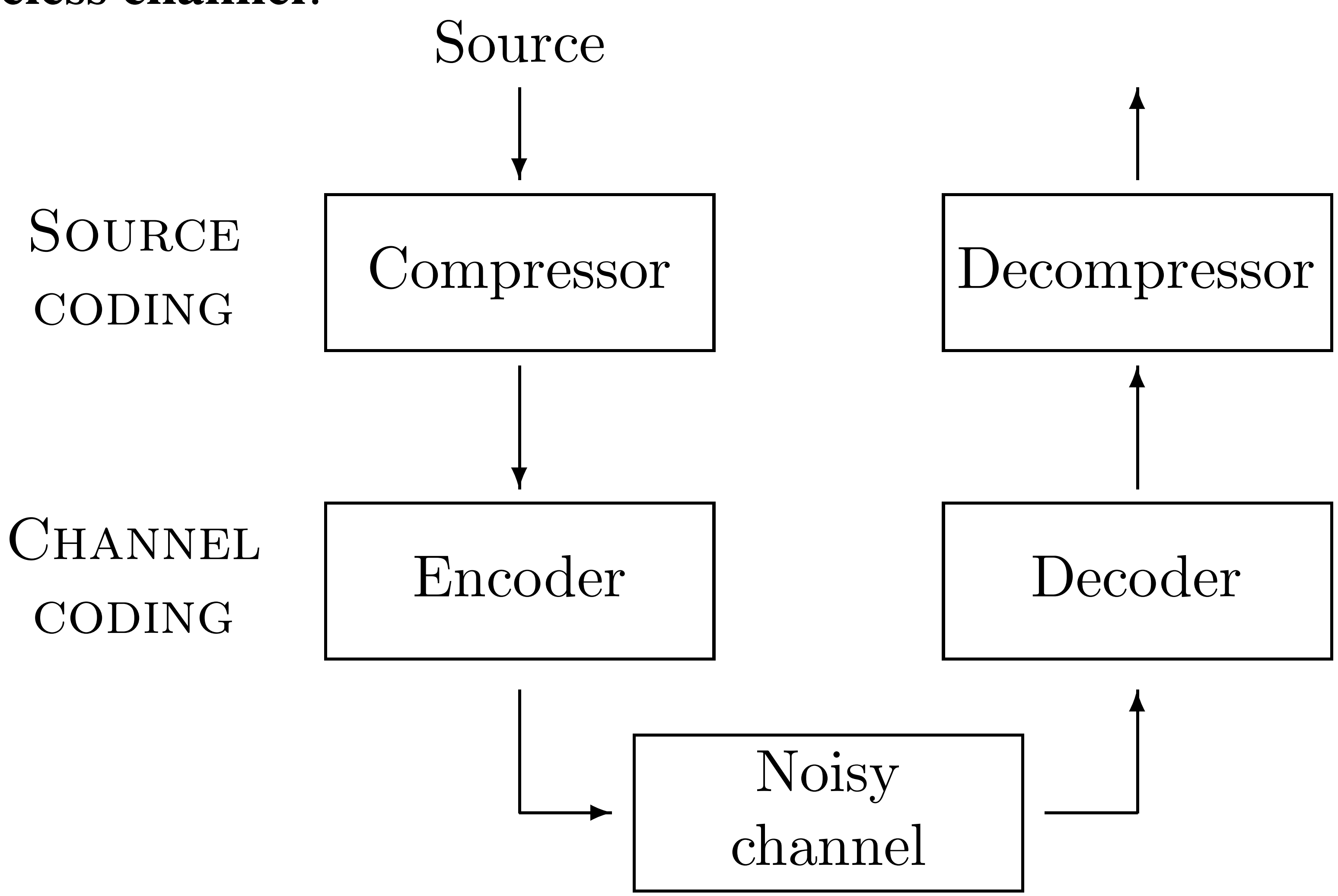
- **And some slides were based on Iain Murray course**
  - **http://www.inf.ed.ac.uk/teaching/courses/it/2014/**

Table of Contents

The big picture

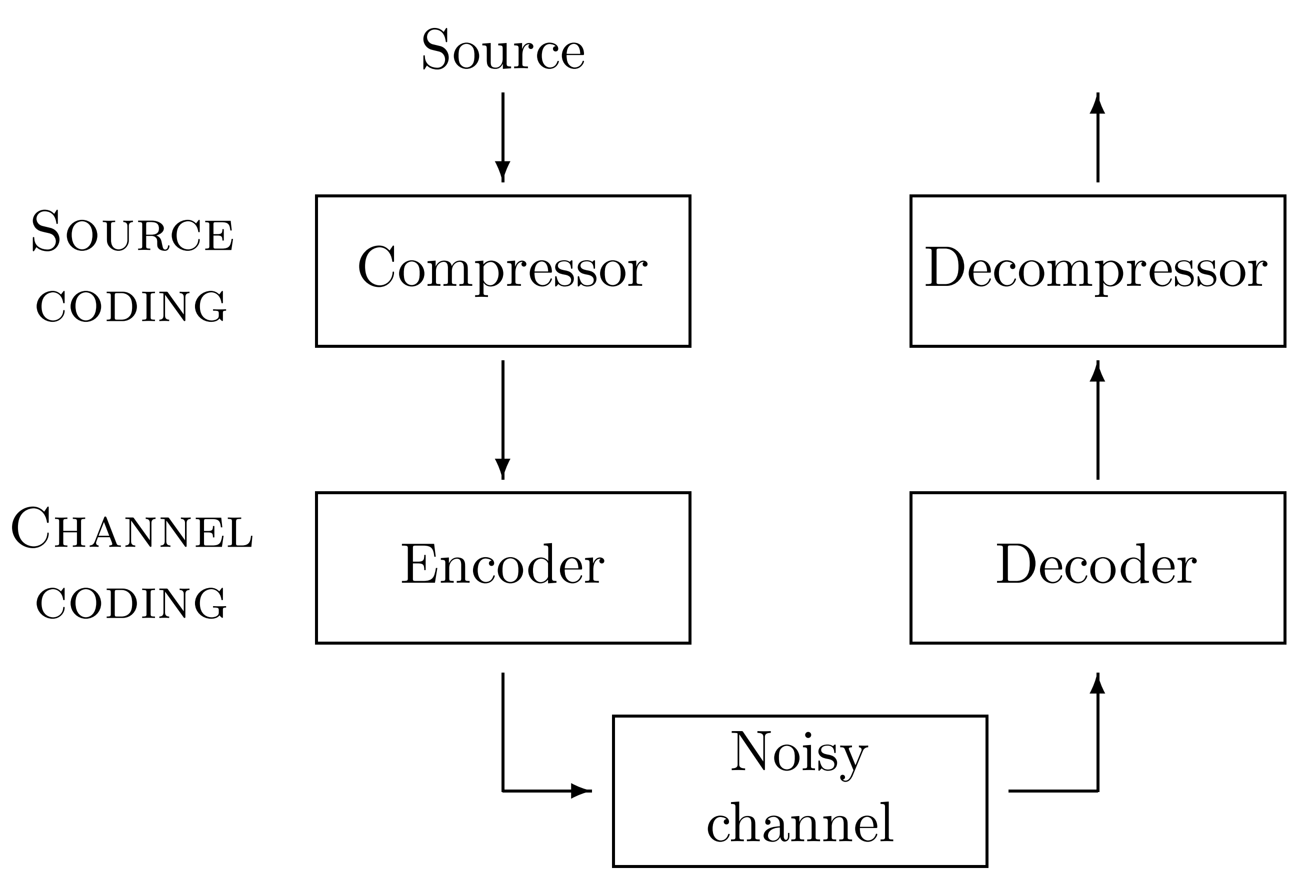
The big picture

- The aim of **source coding** is **data compression**, assuming a noise-free channel.
- Real channel are noisy. The aim of **channel coding** is to make the **noisy channel behave like a noiseless channel**.



The big picture. Channel Coding

- The data to be transmitted has been through a good compressor, so the **bit stream has no obvious redundancy**.
- The channel code, which makes the transmission, will **put back redundancy of a special sort, designed to make the noisy received signal decodable**.



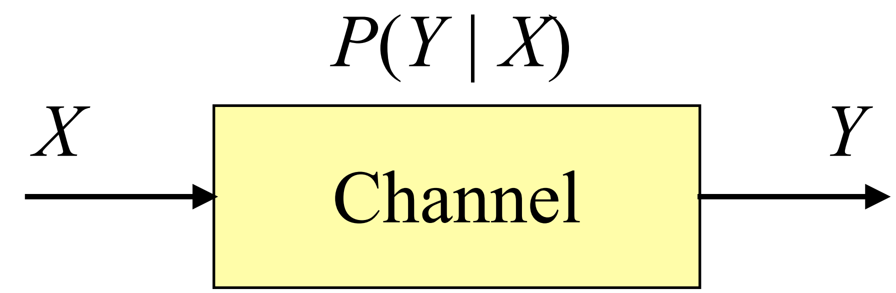
The big picture. Channel Coding

- Suppose we transmit 1000 bits per second with $p_0 = p_1 = 1/2$ over a noisy channel that flips bits with probability $f = 0.1$.
  - What is the **rate of transmission of information**?
    - We might guess that the rate is 900 bits per second by subtracting the expected number of errors per second. But this is not correct! because **the recipient does not know where the errors occurred**.
    - Consider the case where the noise level of $f = 0.5$.
      - Half of the received symbols are correct due to chance alone.
      - But when $f = 0.5$, **no information is transmitted** at all.
- A measure of the **information transmitted** is given by the **mutual information** ***I*(Source; Received)**

Noisy Channels

Discrete memoryless channel

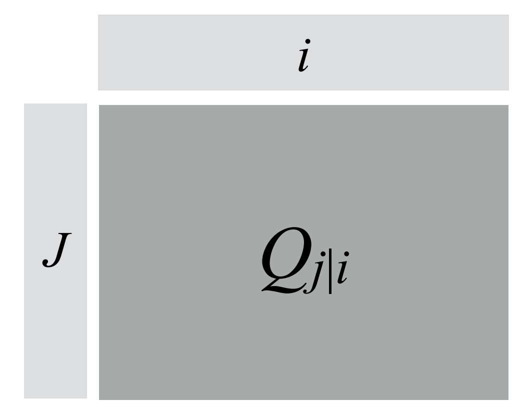
- A **discrete memoryless channel** Q is characterized by
  - an input alphabet A_X,
  - an output alphabet A_Y ,
  - a set of conditional probability distributions $P(y \mid x)$, one for each $x \in A_X$.



- These *transition probabilities* may be written in a *matrix*

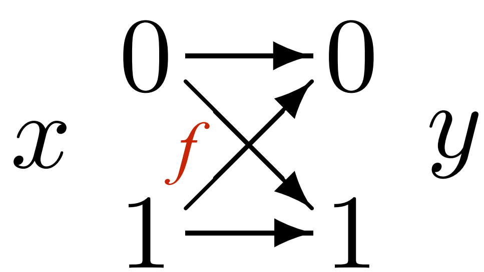
$$Q_{j|i} = P(y = b_j \mid x = a_i)$$

  - The **output variable** j indexing the **rows**
  - The **input variable** i indexing the **columns**
  - **Each column of Q is a probability vector**.
  - $\boldsymbol{p}_{\boldsymbol{y}} = Q\boldsymbol{p}_{\boldsymbol{x}}$



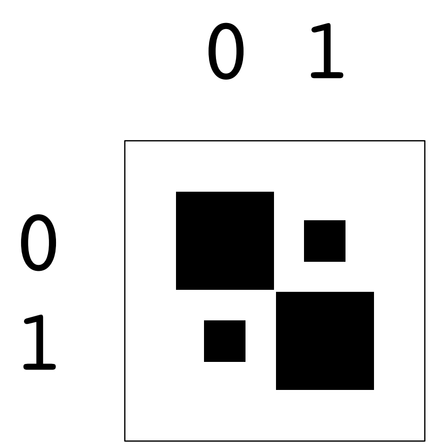
Binary Symmetric Channel

- $A_X = \{0, 1\}$; $A_Y = \{0, 1\}$.



$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= f; \\ P(y=1 \mid x=0) &= f; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$

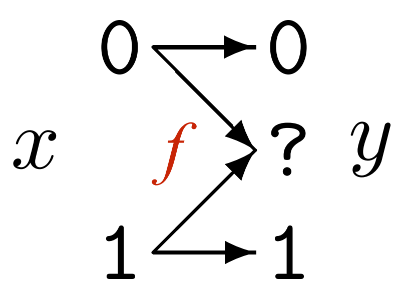
- f is the probability of flipping a bit.
  - So we assume that $f < 0.5$



FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

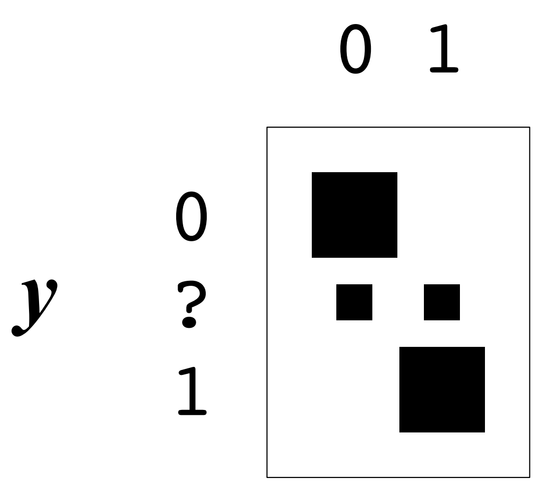
Binary erasure channel

- $A_X = \{0, 1\}$; $A_Y = \{0, \mathbf{?}, 1\}$.



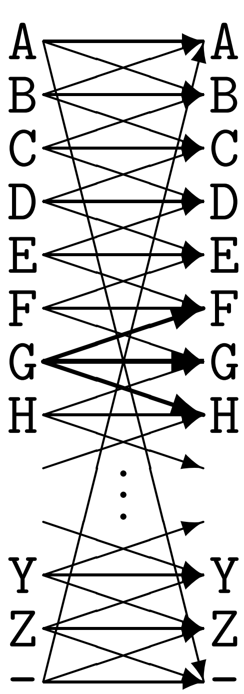
$$
\begin{aligned}
P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= 0; \\
P(y=? \mid x=0) &= f; & P(y=? \mid x=1) &= f; \\
P(y=1 \mid x=0) &= 0; & P(y=1 \mid x=1) &= 1-f.
\end{aligned}
$$

- f is the probability of erasing a bit.
  - So we assume that $f < 0.5$

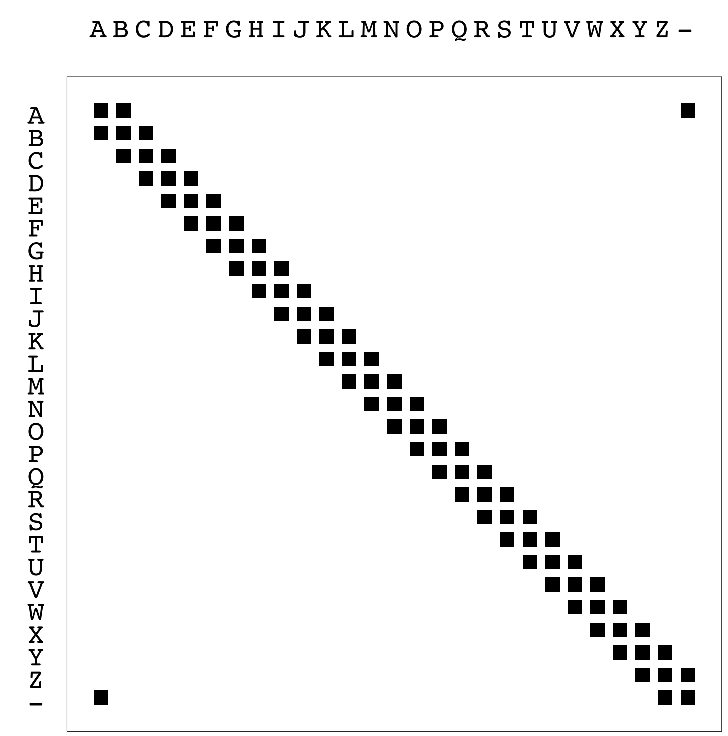


Noisy typewriter

- $A_X = A_Y =$ the 27 letters {A, B, . . . , Z, -}.
- The letters are **arranged in a circle.**
  - When the typist attempts to type B, what comes out is either A, B or C, with probability 1/3 each;
  - When the input is C, the output is B, C or D;
  - and so forth, with the final letter ‘-’ adjacent to the first letter A.

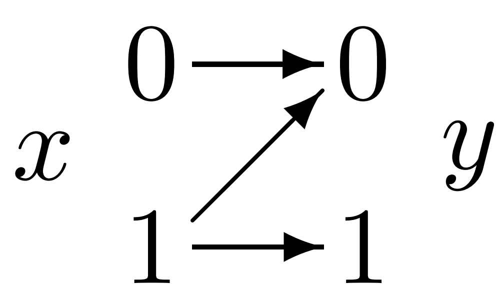


$$
\begin{aligned}
&\vdots \\
P(y = \mathtt{F} \mid x = \mathtt{G}) &= 1/3; \\
P(y = \mathtt{G} \mid x = \mathtt{G}) &= 1/3; \\
P(y = \mathtt{H} \mid x = \mathtt{G}) &= 1/3; \\
&\vdots
\end{aligned}
$$



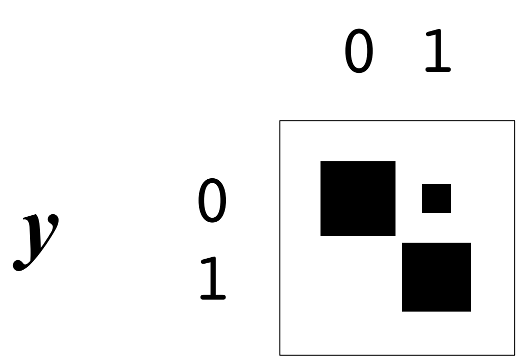
Z channel

- $\mathcal{A}_X = \{0, 1\}$; $\mathcal{A}_Y = \{0, 1\}$.



$$\begin{array}{rclrcl} P(y=0 \mid x=0) & = & 1; & P(y=0 \mid x=1) & = & f; \\ P(y=1 \mid x=0) & = & 0; & P(y=1 \mid x=1) & = & 1-f. \end{array}$$

- f is the probability of flipping a one.
  - So we assume that $f < 0.5$



Inferring the input given the output

Inferring the input given the output

- If we assume that the **input *x*** to a **channel** comes from an ensemble X,
  - We obtain a joint ensemble XY in

$$P(x,y) = P(y \mid x)P(x)$$

- If we receive a particular symbol $\underline{y}$, what was the input symbol x?
  - Typically we won't know for certain
  - The posterior distribution of the input using Bayes' theorem

$$P(x \mid y) = \frac{P(y \mid x)P(x)}{P(y)} = \frac{P(y \mid x)P(x)}{\sum_{x'} P(y \mid x')P(x')}$$

Example - binary symmetric channel

- Consider a **binary symmetric channel** with probability of error $f = 0.15$.
- Let the input ensemble be $P_X : \{p_0 = 0.9, p_1 = 0.1\}$.
- Assume **we observe y = 1.**

$P(y = 1 / x = 1) = 0.85$

$$
\begin{aligned}
P(x=1 \mid y=1) &= \frac{P(y=1 \mid x=1)P(x=1)}{\sum_{x'} P(y \mid x')P(x')} \\
&= \frac{0.85 \times 0.1}{0.85 \times 0.1 + 0.15 \times 0.9} \\
&= \frac{0.085}{0.22} = 0.39.
\end{aligned}
$$

$$P(x=0 \mid y=1) = 0.61$$

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Example - binary symmetric channel

- Consider a **binary symmetric channel** with probability of error $f = 0.15$.
- Let the input ensemble be $P_X : \{p_0 = 0.9, p_1 = 0.1\}$.
- Assume **we observe $y = 0$**.

$$P(x=1 \mid y=0) = \frac{P(y=0 \mid x=1)P(x=1)}{\sum_{x'} P(y \mid x')P(x')}$$

$$P(y=0 / x=1) = 0.15$$

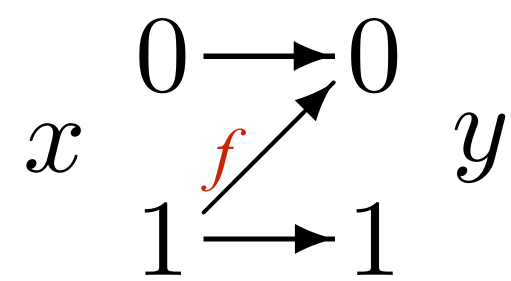
$$= \frac{0.15 \times 0.1}{0.15 \times 0.1 + 0.85 \times 0.9}$$

$$= \frac{0.015}{0.78} = 0.019.$$

$$P(x=0 \mid y=0) = 0.981$$

Example - Z channel

- Consider a **Z channel** with probability of error $f = 0.15$.
- Let the input ensemble be $P_X : \{p_0 = 0.9, p_1 = 0.1\}$.
- Assume we observe $y = 1$.



$$P(x=1 \mid y=1) = \frac{P(y=1 \mid x=1)P(x=1)}{\sum_{x'} P(y \mid x')P(x')}$$

$$P(y=1/x=1) = 0.85$$

$$= \frac{0.85 \times 0.1}{0.85 \times 0.1 + 0 \times 0.9}$$

$$= \frac{0.085}{0.085} = 1.0.$$

So given the output $y = 1$ we become certain of the input

Information conveyed by a channel

Information conveyed by a channel

- We now consider **how much information can be communicated through a channel**.
  - We are interested in finding ways of using the channel **such that all the bits that are communicated are recovered with negligible probability of error**
- Assuming a particular input ensemble X, we can measure **how much information the output conveys about the input** by the mutual information $I(X;Y)$

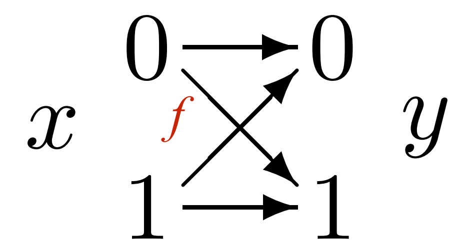
$$I(X;Y) = H(X) - H(X|Y)$$

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

Example with a **BSC**

- Consider the **binary symmetric channel** again, with $f = 0.15$ and $P_X : \{p_0 = 0.9, p_1 = 0.1\}$.

- We need $P(y)$ and $P(y \mid x)$ for



$$I(X;Y) = H(Y) - H(Y \mid X)$$

- Compute $P(y)$

$$P(y) = \sum_x P(x,y) = \sum_x P(y \mid x)P(x) = P(y \mid x=0)P(x=0) + P(y \mid x=1)P(x=1)$$

$$P(y=0) = P(y=0 \mid x=0)P(x=0) + P(y=0 \mid x=1)P(x=1) = 0.85 \times 0.9 + 0.15 \times 0.1 = 0.78$$

$$P(y=1) = P(y=1 \mid x=0)P(x=0) + P(y=1 \mid x=1)P(x=1) = 0.15 \times 0.9 + 0.85 \times 0.1 = 0.22$$

- $P(y \mid x)$ is defined by the channel

$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= f; \\ P(y=1 \mid x=0) &= f; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$

$$f = 0.15$$

Example with a BSC

$$\boxed{I(X;Y) = H(Y) - H(Y \mid X)}$$

- Consider the **binary symmetric channel** again, with $f = 0.15$ and $P_X : \{p_0 = 0.9, p_1 = 0.1\}$.

- $P(y = 0) = 0.78$; $P(y = 1) = 0.22$

$$H(Y) = H_2(0.22) = 0.76 \text{ bits}$$

$$H_2(p) = p \log_2 \tfrac{1}{p} + (1-p) \log \tfrac{1}{1-p}$$

- $H(Y \mid X)$

$$H(Y \mid X) = P(x=0) H(Y \mid x=0) + P(x=1) H(Y \mid x=1)$$

$$H(Y \mid x=0) = H_2(f) = H_2(0.15) = 0.61 \text{ bits}$$

$$H(Y \mid x=1) = H_2(f) = H_2(0.15) = 0.61 \text{ bits}$$

$$H(Y \mid X) = 0.9 H_2(f) + 0.1 H_2(f) = H_2(f) = H_2(0.15) = 0.61 \text{ bits}$$

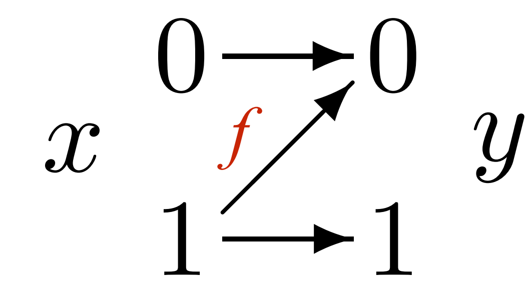
$$I(X;Y) = H(Y) - H(Y \mid X)$$

$$I(X;Y) = H_2(0.22) - H_2(0.15) = 0.76 - 0.61 = 0.15 \text{ bits}$$

$$H(X) = H_2(0.1) = 0.47 \text{ bits}$$

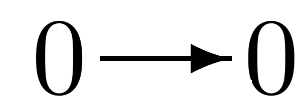
Example with a Z channel

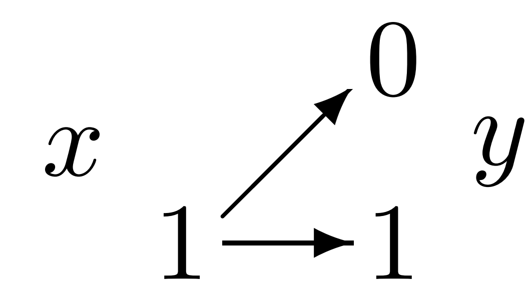
- Consider the **Z channel** again, with $f = 0.15$ and $P_X : \{p_0 = 0.9, p_1 = 0.1\}$.
- $1 - f = 0.85$
- Compute $I(X;Y)$



$$I(X;Y) = H(Y) - H(Y \mid X)$$

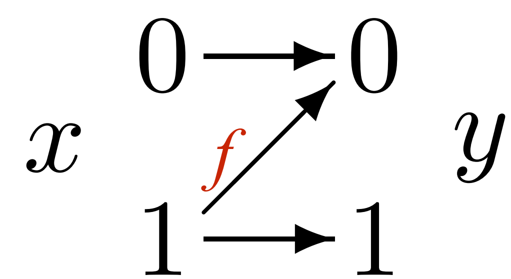
$$= H_2(0.085) - [0.9 H_2(0) + 0.1 H_2(0.15)]$$





Example - Z channel

- Consider a **Z channel** with probability of error $f = 0.15$.
- Let the input ensemble be $P_X : \{p_0 = 0.9, p_1 = 0.1\}$.
- Assume we observe $y = 1$.



$P(y = 1 / x = 1) = 0.85$

$$P(x=1 \mid y=1) = \frac{P(y=1 \mid x=1)P(x=1)}{\sum_{x'} P(y \mid x')P(x')}$$

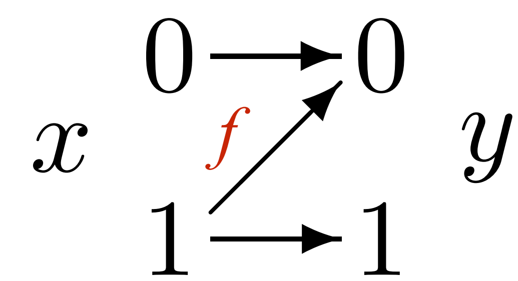
$$= \frac{0.85 \times 0.1}{0.85 \times 0.1 + 0 \times 0.9}$$

$$= \frac{0.085}{0.085} = 1.0.$$

So given the output $y = 1$ we become certain of the input

Example with a Z channel

- Consider the **Z channel** again, with $f = 0.15$ and $P_X : \{p_0 = 0.9, p_1 = 0.1\}$.
- $P(y = 1) = 0.085$
- Compute $I(X;Y)$



$$
\begin{aligned}
I(X;Y) &= H(Y) - H(Y \mid X) \\
&= H_2(0.085) - [0.9 H_2(0) + 0.1 H_2(0.15)] \\
&= 0.42 - 0.1 \times 0.61 = 0.36 \ \texttt{bits}
\end{aligned}
$$

- **BSC** - $I(X; Y) = 0.15$ bits
- **Z** Channel: $I(X; Y) = 0.36$ bits
- The Z channel is a **more reliable channel** (for the same f)

Maximizing the mutual information

- The **mutual information** between the input and the output **depends on the chosen input ensemble !**
- To maximize the mutual information conveyed by the channel by choosing the best possible input ensemble. We define the **capacity of the channel** to be its **maximum mutual information.**
- **The capacity** of a channel Q is:

$$C(Q) = \max_{P_X} I(X;Y)$$

  - The distribution P_X that achieves the maximum is called the **optimal input distribution**, denoted by P^*_X .
  - There may be **multiple optimal input distributions** achieving the same value of $I(X ; Y)$

Capacity - Example for BSC

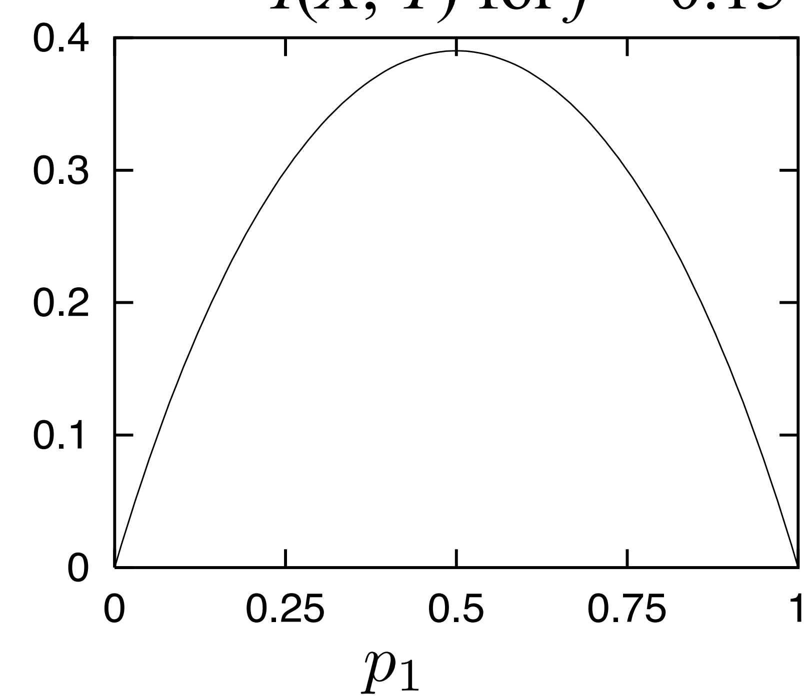
- Consider the **binary symmetric channel** with $f = 0.15$.
- **With $P_X = \{p_0 = 0.9, p_1 = 0.1\}$**, we have **$I(X; Y) = 0.15$ bits**

$$I(X;Y) = H_2(0.22) - H_2(0.15) = 0.76 - 0.61 = 0.15 \text{ bits}$$

- What is the **maximum of $I(X; Y)$**? For which P_X?
- By symmetry, the **optimal input distribution** is $\{0.5, 0.5\}$ and the capacity is 0.39 bits.

$$C(Q_{BSC}) = H_2(0.5) - H_2(0.15) = 1 - 0.61 = 0.39 \text{ bits}$$

$I(X; Y)$ for $f = 0.15$

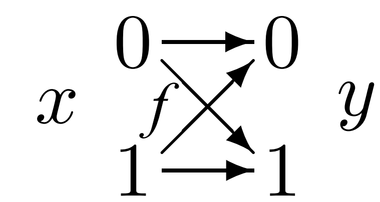


- Note the mutual information $I(X; Y)$

$$I(X;Y) = H_2((1-f)p_1 + (1-p_1)f) - H_2(f)$$

Capacity - Example for BSC

- Consider the **binary symmetric channel** with f.



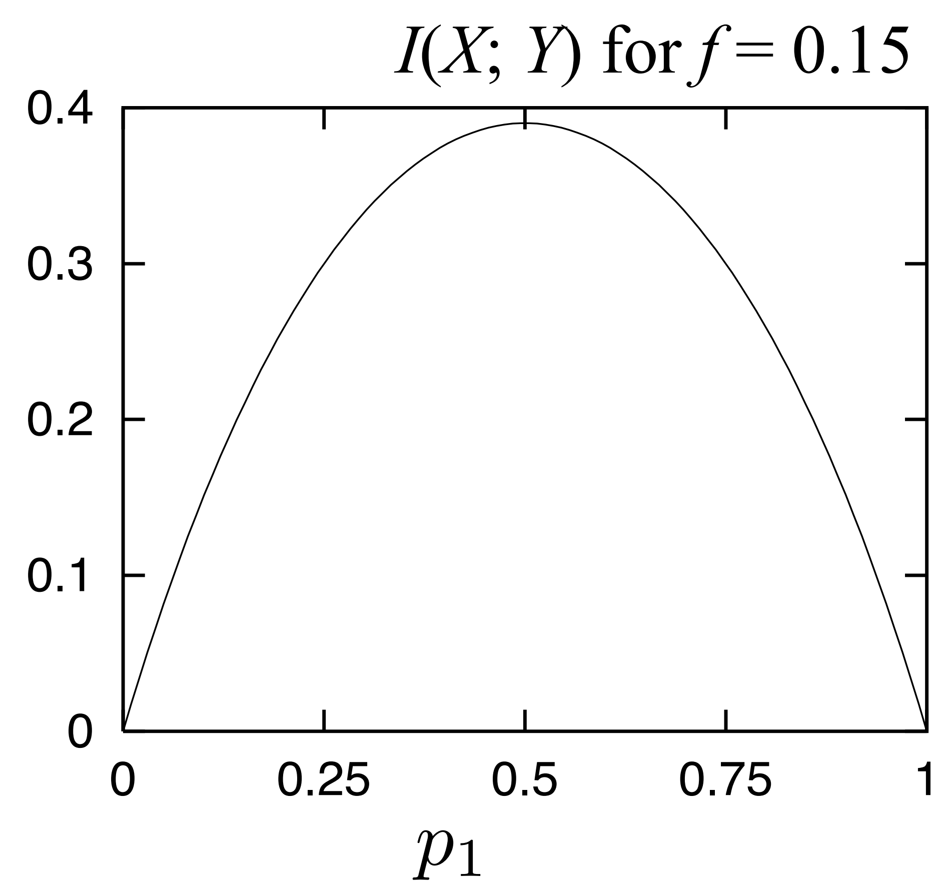
$$I(X;Y) = H(Y) - H(Y\,|\,X)$$

$$H(Y) = H_2((1-f)p_1 + (1-p_1)f)$$

$$P(y=1) = p_1(1-f) + (1-p_1)f$$

$$H(Y\,|\,X) = p_1 H_2(f) + (1-p_1)H_2(f) = H_2(f)$$

$$I(X;Y) = H_2((1-f)p_1 + (1-p_1)f) - H_2(f)$$



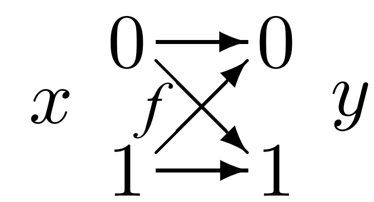
$$p_1 = 0.5 \rightarrow (1-f)p_1 + (1-p_1)f = 0.5$$

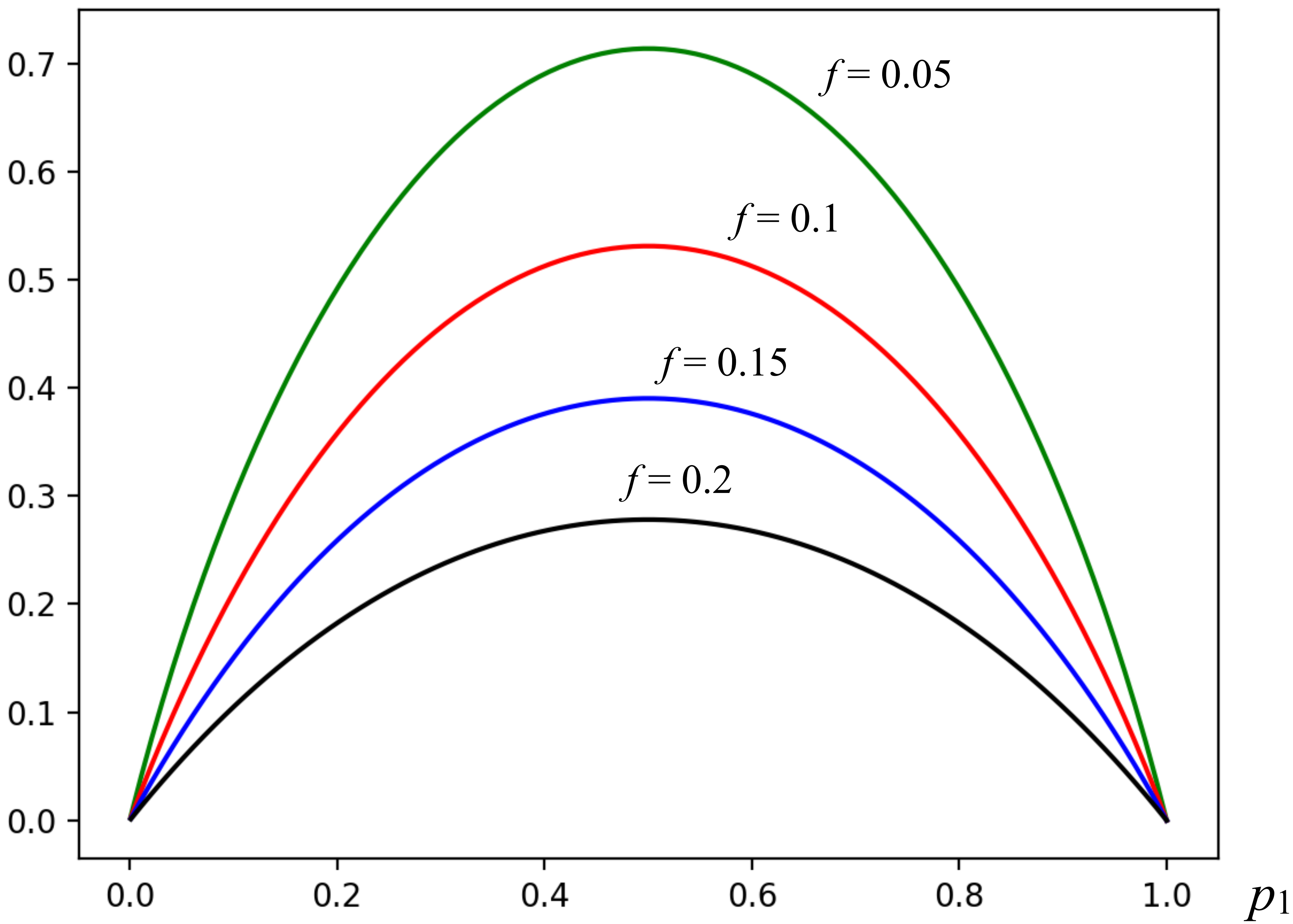
$$C(Q_{BSC}) = H_2(0.5) - H_2(0.15) = 1 - 0.61 = 0.39 \text{ bits}$$

Capacity - Example for BSC

$$\boxed{I(X;Y) = H_2((1-f)p_1 + (1-p_1)f) - H_2(f)}$$

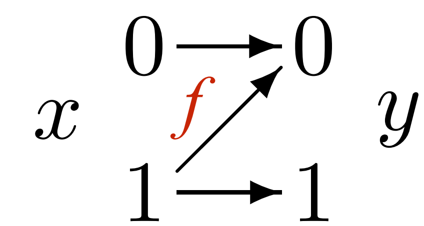
- Consider the **binary symmetric channel** with f.





Capacity - Example for Z Channel

- Consider the **Z channel** with $f = 0.15$.



$$I(X;Y) = H(Y) - H(Y \mid X)$$

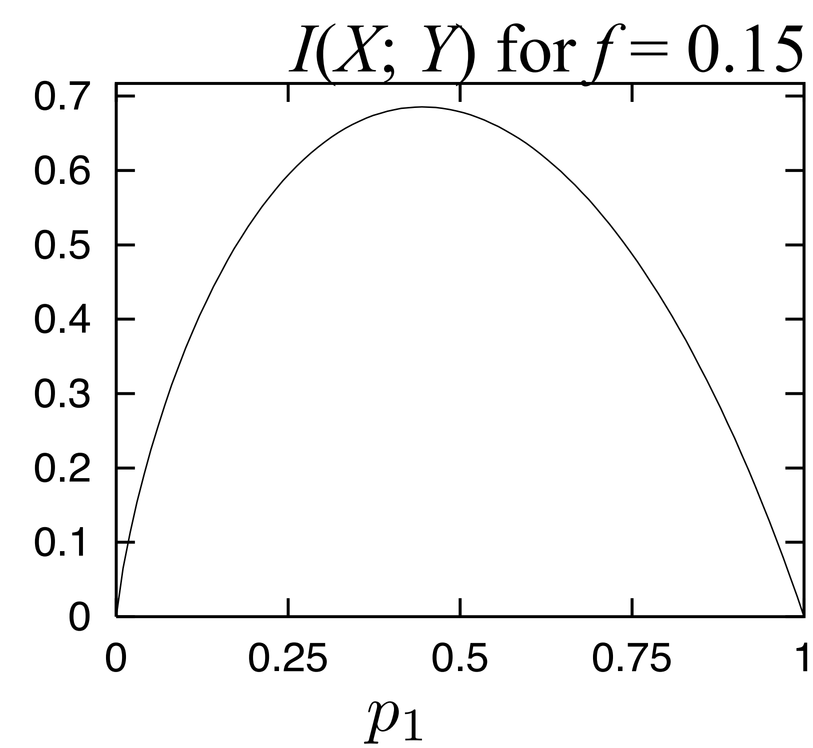
$$H(Y) = H_2((1-f)p_1) \qquad P(y=1) = p_1(1-f)$$

$$H(Y \mid X) = p_1 H_2(f) + p_0 H_2(0) = p_1 H_2(f)$$

$$\boxed{I(X;Y) = H_2((1-f)p_1) - p_1 H_2(f)}$$

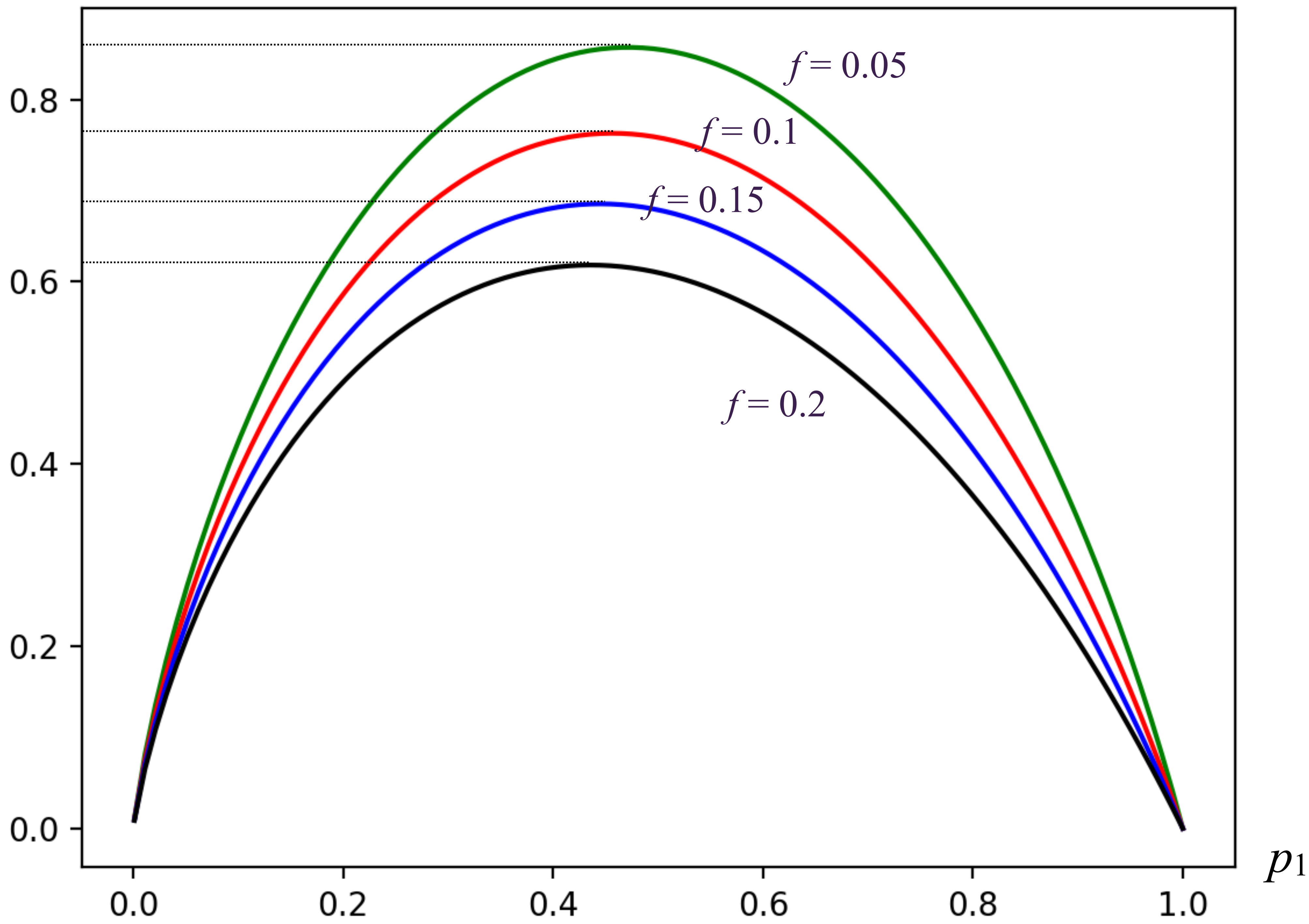
- It is maximized for $f = 0.15$ by $p_1^* = 0.445$

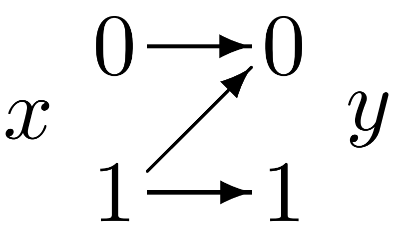
- We find $C(Q_Z) = 0.685$.



Capacity - Example for Z Channel

$$I(X;Y) = H_2((1-f)p_1) - p_1 H_2(f)$$





Capacity - Example the noisy typewriter

- $A_X = A_Y$ = the 27 letters {A, B, . . . , Z, -}.
- When the typist attempts to type B, what comes out is either A, B or C, with probability 1/3 each;
- The **optimal input distribution is a uniform distribution over *x***.
  - The output distribution is is also a uniform distribution over *y*.

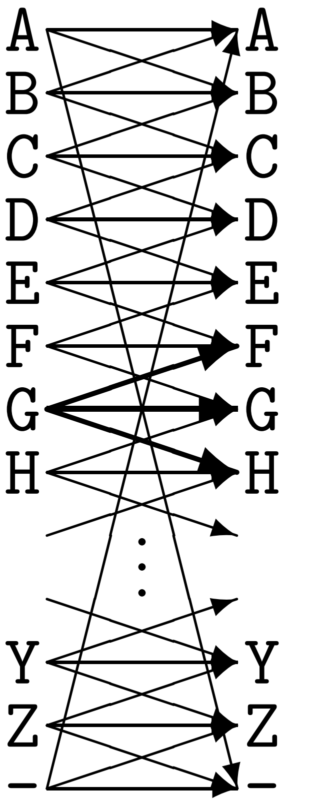
$$H(Y) = \log_2 27 = \log_2 3^3 \text{ bits}$$

  - For each *x*, $P(y \mid x)$ = 1/3 for 3 letters and zero for the others.

$$H(Y \mid X) = (3\frac{1}{3}\log_2 3) = \log_2 3 \text{ bits}$$

$$I(X;Y) = 3\log_2 3 - \log_2 3 = 2\log_2 3 = \log_2 9 \text{ bits}$$

$$C_{TypeWritter} = \log_2 9 \text{ bits}$$



The noisy-channel coding theorem

The noisy-channel coding theorem

- It seems plausible that the '**capacity**' we have defined may be a **measure of information conveyed by a channel**.

- What is not obvious, is that the **capacity** indeed **measures** the **rate at which blocks of data can be communicated** **over the channel with *arbitrarily small probability of error*.**

(N, K) block code

- An (N, K) block code for a channel Q is a list of $S = 2^K$ codewords

$$\left\{\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(2^K)}\right\}, \quad \mathbf{x}^{(s)} \in A_X^N \quad ,$$

  each of length N.

- Using this code we can encode a signal $s \in \{1, 2, 3, \ldots, 2^K\}$ as $\mathbf{x}^{(s)}$
- The ***rate*** of the code is $R = K/N$ bits per channel use.
  - This definition of the rate for any channel, not only channels with binary inputs
  - It is sometimes conventional to define the rate of a code for a channel with q input symbols to be $K/(N \log q)$.

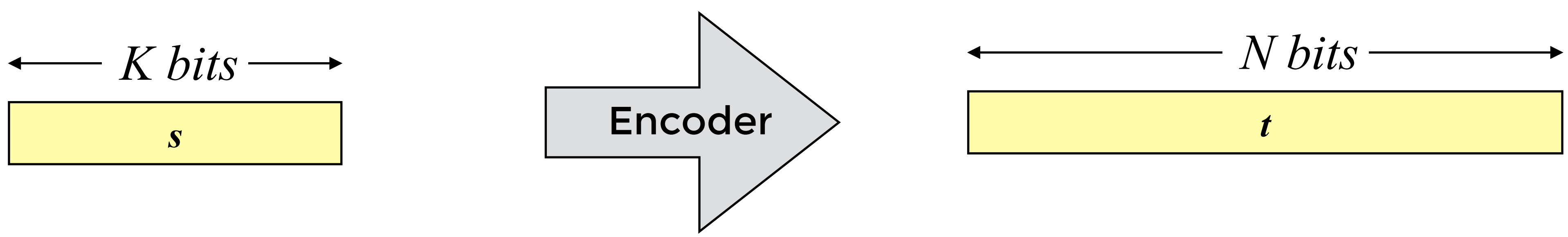
Block Codes (N, K)

$R = K/N$ bits per channel use

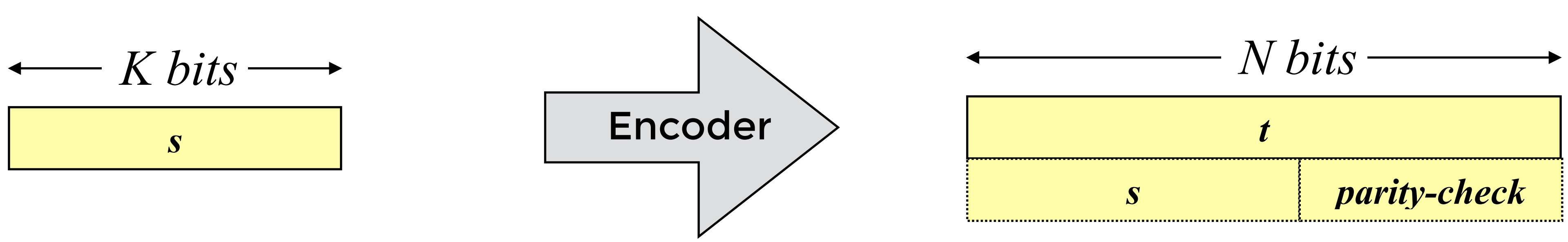
- Add redundancy to **blocks of data** instead of encoding one bit at a time
- A **block code** is a rule for converting a **sequence of source bits** s, of length ***K***, say, into a transmitted sequence t of length N bits.

$R_{H(7,4)} = 0.57$

  - To add redundancy, $N > K$



- In a **linear block code**, the extra $N - K$ bits are **linear functions** of the original K bits

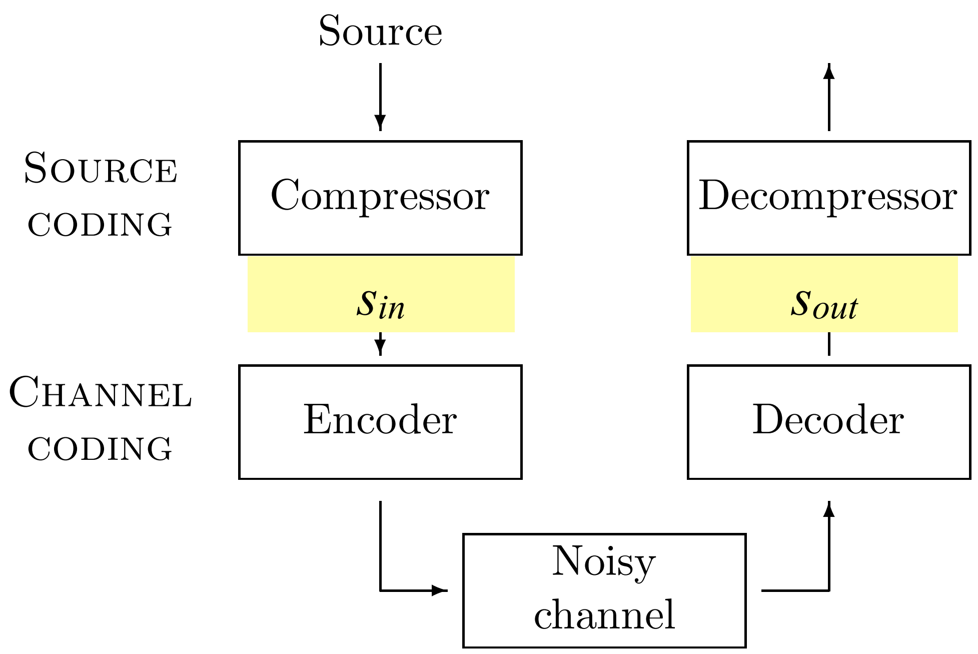


Decoding and the probability of block error

- A **decoder** for an (N, K) block code is a **mapping** from the set of **length-N strings** of channel outputs, A_Y^N, to a **codeword label** $\hat{s} \in \{0, 1, 2, \ldots, 2^K\}$.
  - The extra symbol $\hat{s} = 0$ can be used to indicate a 'failure'.

- **The** **probability of block error** of a code and decoder, for a **given channel**, and for **a given probability distribution** over the encoded signal $P(s_{in})$, is:

$$p_B = \sum_{s_{in}} P(s_{in}) P(s_{out} \neq s_{in} \mid s_{in})$$



The maximal probability of block error and optimal decoder

- **The maximal probability of block error** is:

$$p_{BM} = \max_{s_{in}} P(s_{out} \neq s_{in} \mid s_{in})$$

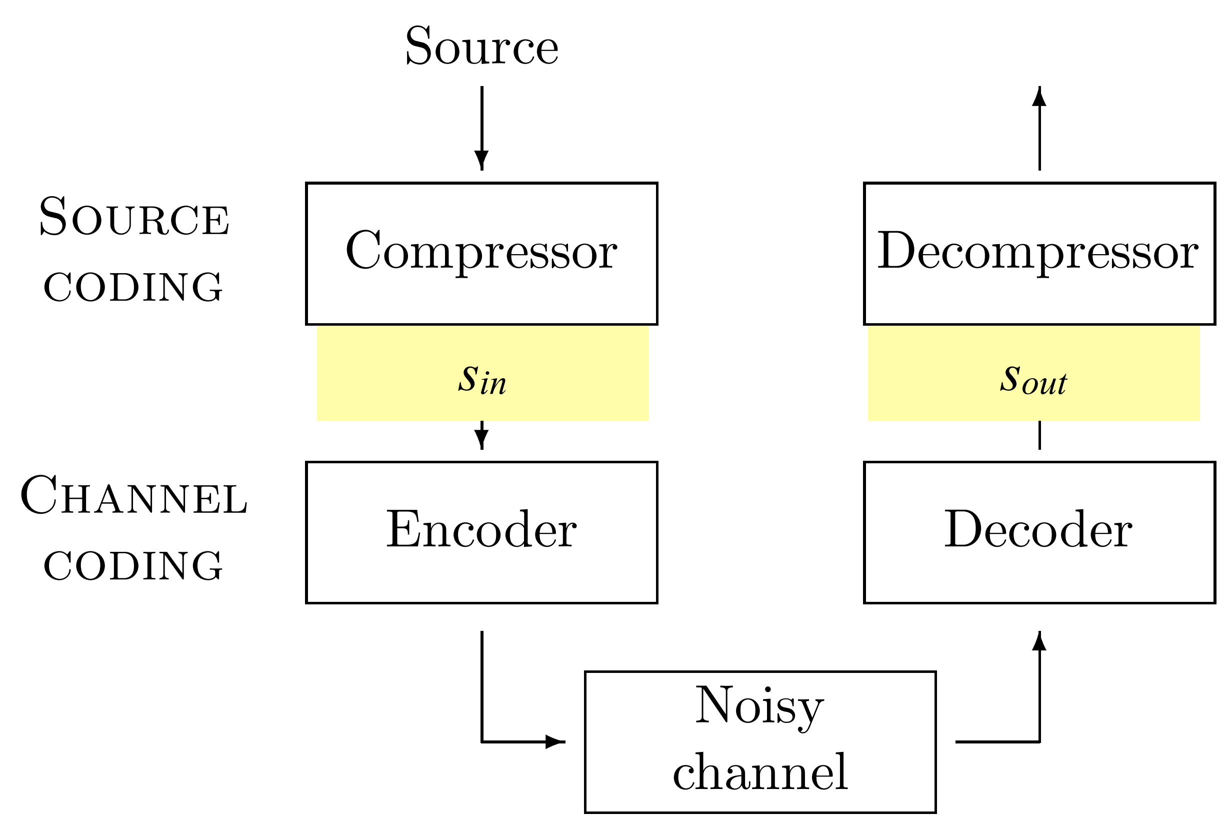
- The **optimal decoder** for a channel code is the one that **minimizes the probability of block error**.
  - It decodes an output $\mathbf{y}$ as the input s that has maximum posterior probability $P(s \mid \mathbf{y})$.

$$P(s \mid \mathbf{y}) = \frac{P(\mathbf{y} \mid s) P(s)}{\sum_{s'} P(\mathbf{y} \mid s') P(s')} \qquad \hat{s}_{\text{optimal}} = \text{argmax}\, P(s \mid \mathbf{y})$$

  - A **uniform prior distribution on** s **is usually assumed**, in which case the **optimal decoder is also the maximum likelihood decoder**, i.e., the decoder that maps an output $\mathbf{y}$ to the input s that has maximum likelihood $P(\mathbf{y} \mid s)$.

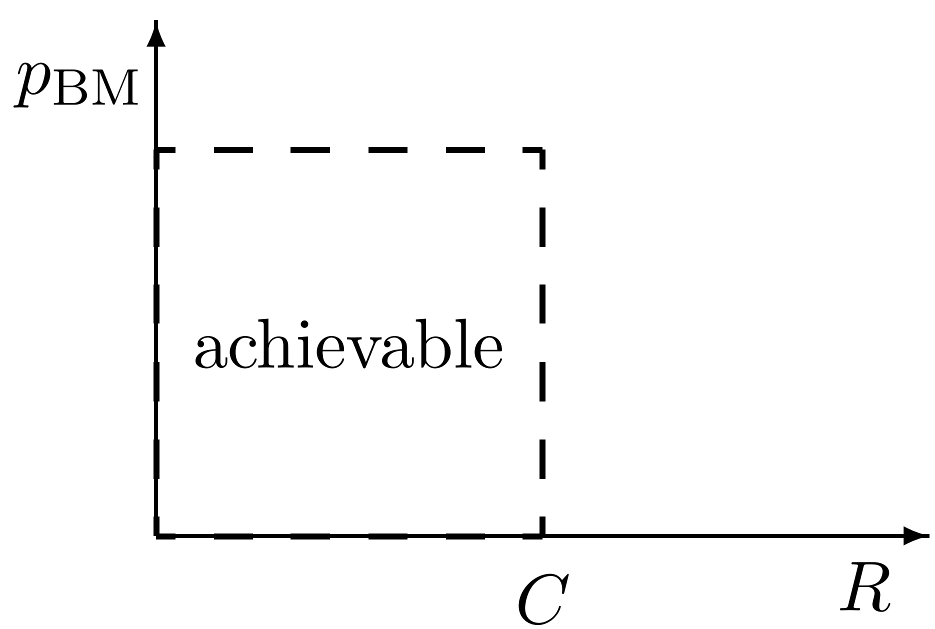
The probability of bit error - p_b

- **The probability of bit error** p_b
  - Assuming that the codeword number s is represented by a binary vector $\mathbf{s}$ of length K bits.
  - It is the average probability that a bit of s_{out} is not equal to the corresponding bit of s_{in} (averaging over all K bits).



Shannon's noisy-channel coding theorem (part one)

- Associated with each discrete memoryless channel, there is a non-negative number C (called the **channel capacity**) with the following property:
- For any $\varepsilon > 0$ and $R < C$, for large enough N,
  - there **exists a block code of length** N and rate $\geq$ R
  - and a decoding algorithm, **such that the maximal probability of block error is $< \varepsilon$**.



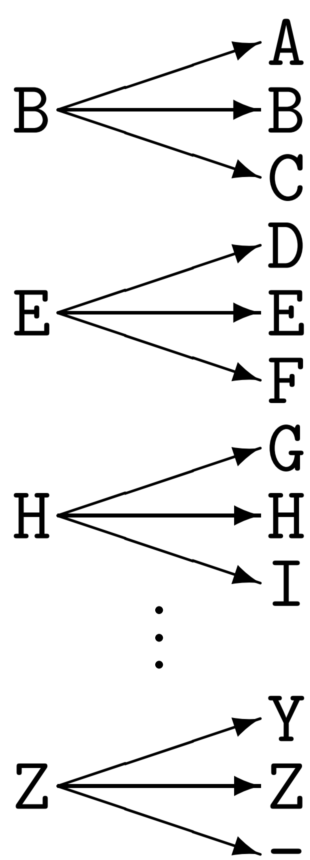
Block code (N, K)

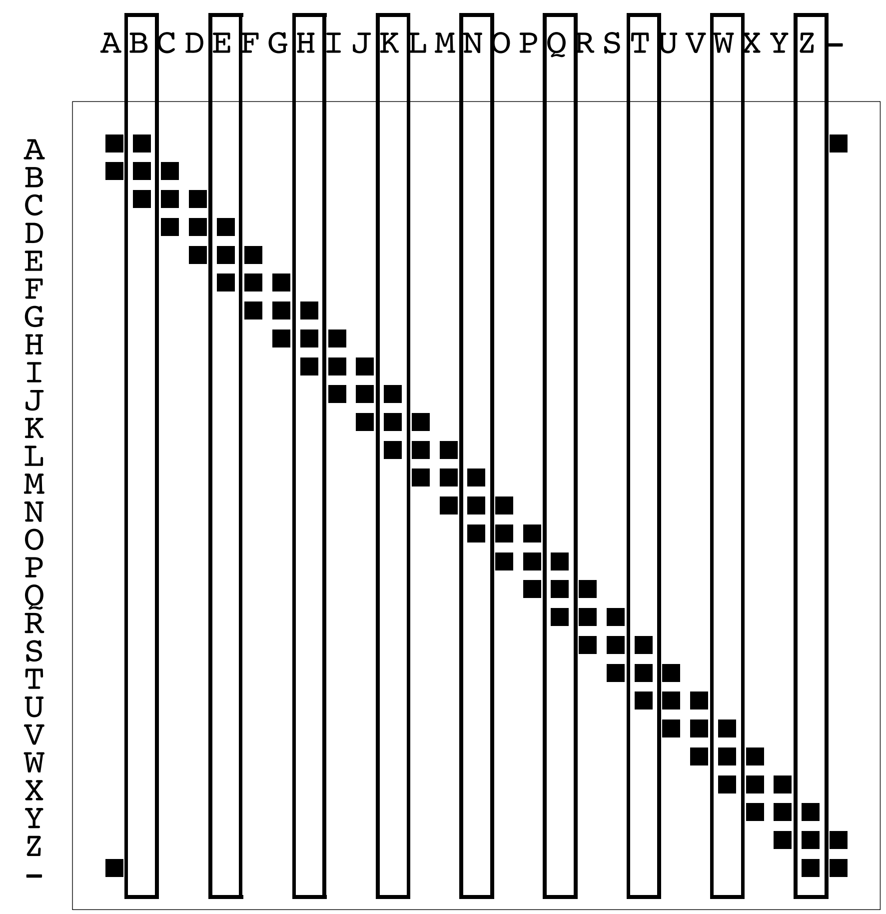
Rate $R = K / N$

Confirmation of the theorem for the noisy typewriter channel

- In the case of the noisy typewriter, we can easily confirm the theorem, because we can create a **completely error-free communication strategy** using a block code of length $N = 1$.

A **B** C D **E** F G **H** I J **K** L M **N** O P **Q** R S **T** U V **W** X Y **Z** _

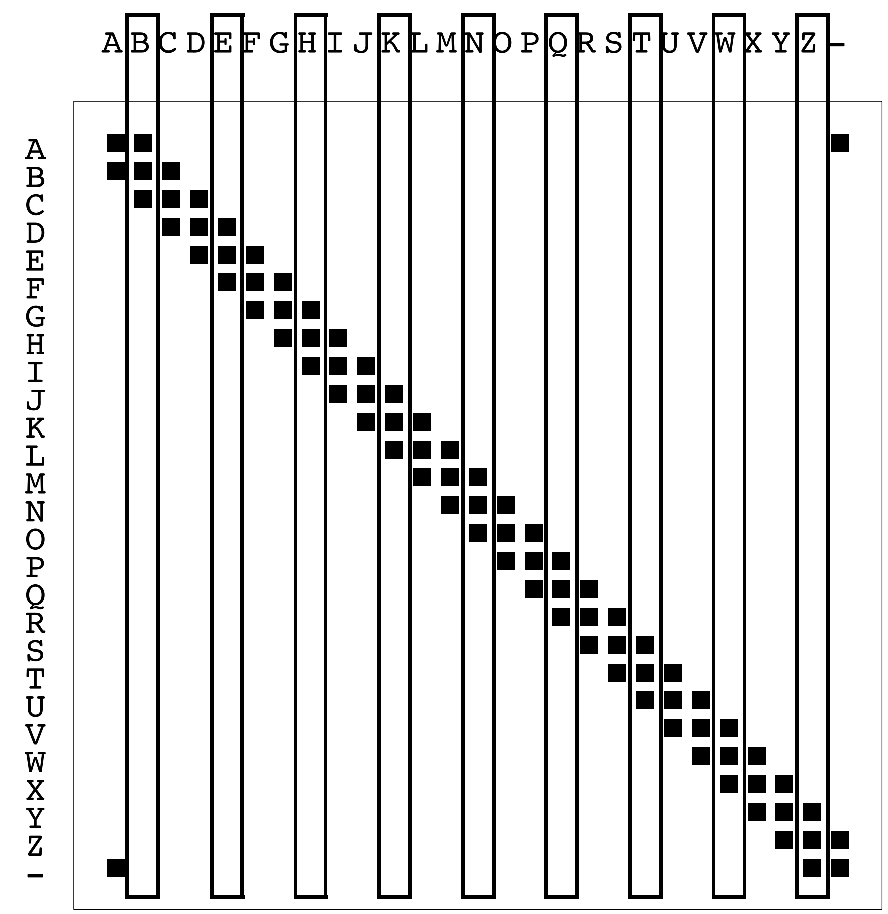




Confirmation of the theorem for the noisy typewriter channel

- In the case of the noisy typewriter, we can easily confirm the theorem, because we can create a **completely error-free communication strategy** using a block code of length $N = 1$.

A **B** C D **E** F G **H** I J **K** L M **N** O P **Q** R S **T** U V **W** X Y **Z** _



- These letters form a **non-confusable subset of the input** alphabet. <u>Any output can be uniquely decoded</u>.
- The **number of inputs** in the **non-confusable subset is 9**, so the error-free information rate of this system is $\log_2 9$ bits, which is equal to the capacity C.

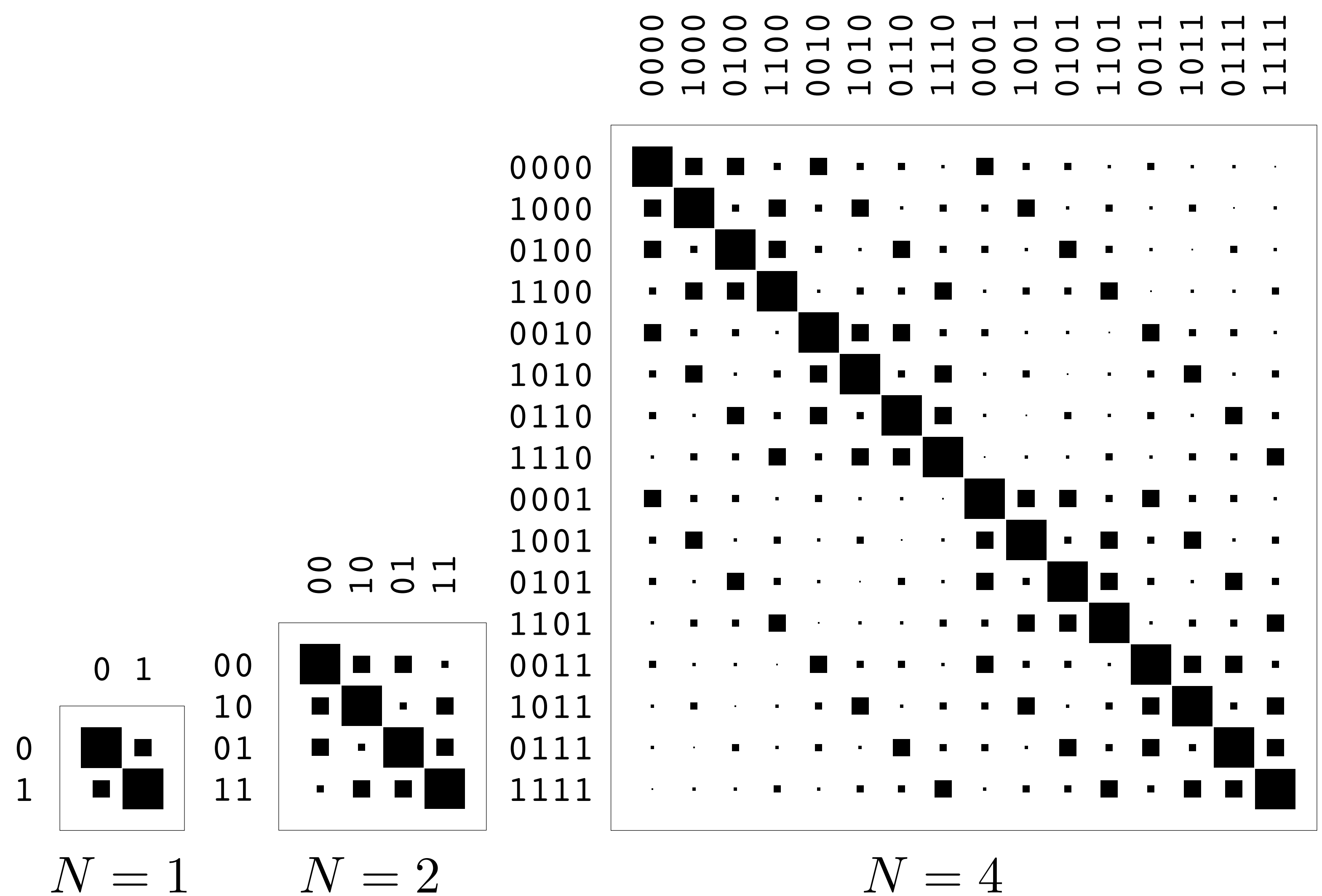
The noisy typewriter channel and the Theorem

The theorem	How it applies to the noisy typewriter
Associated with each discrete memoryless channel, there is a non-negative number C.	The capacity C is $\log_2 9$.
For any $\epsilon > 0$ and $R < C$, for large enough N,	No matter what ϵ and R are, we set the blocklength N to 1.
there exists a block code of length N and rate $\geq R$	The block code is $\{\texttt{B}, \texttt{E}, \ldots, \texttt{Z}\}$. The value of K is given by $2^K = 9$, so $K = \log_2 9$, and this code has rate $\log_2 9$, which is greater than the requested value of R.
and a decoding algorithm,	The decoding algorithm maps the received letter to the nearest letter in the code;
such that the maximal probability of block error is $< \epsilon$.	the maximal probability of block error is zero, which is less than the given ϵ.

Intuitive preview of proof

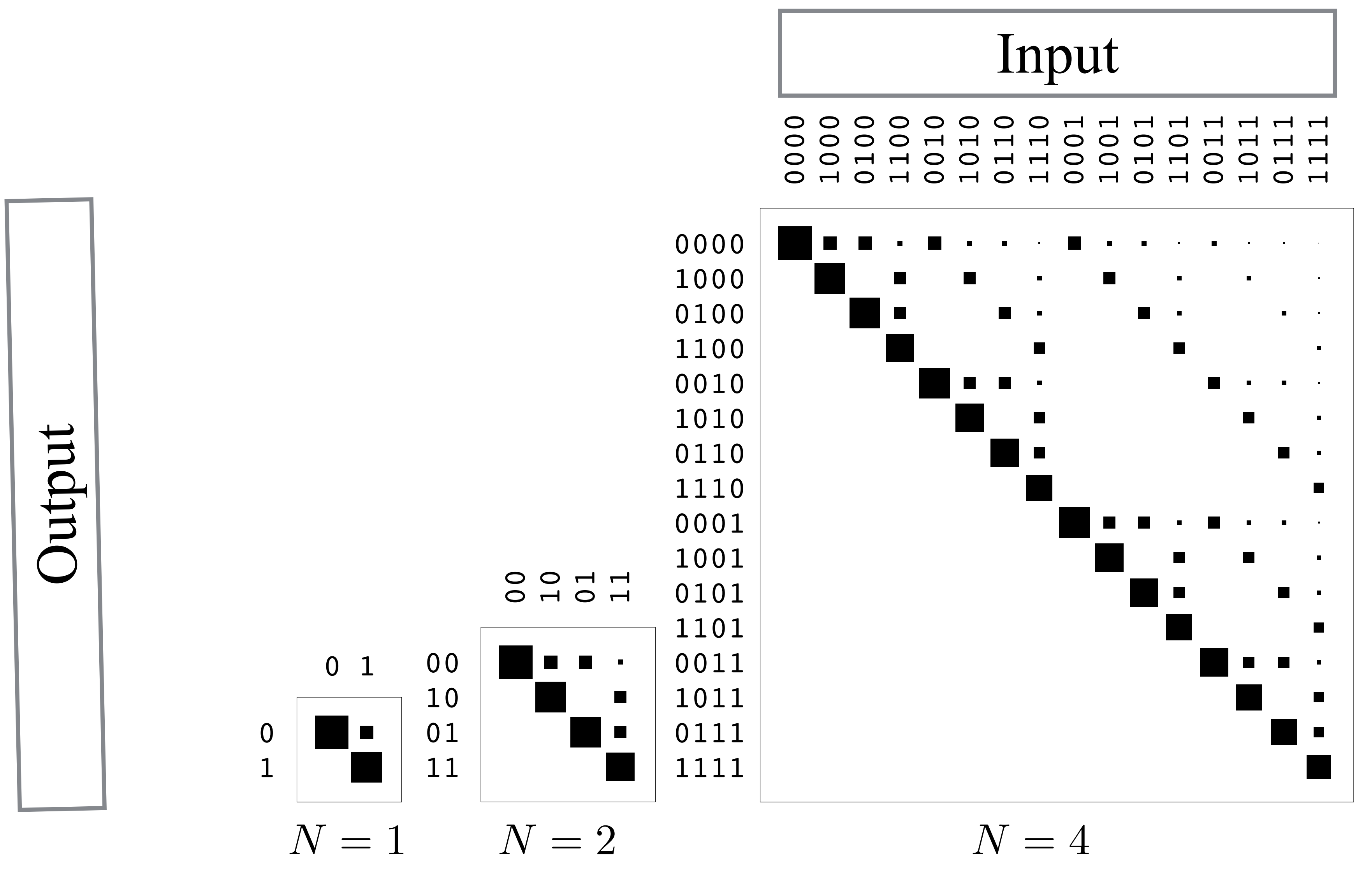
Extended channels

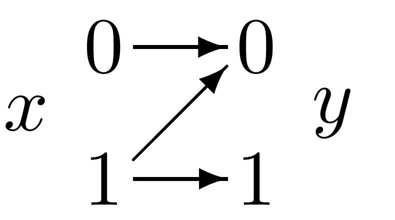
- The **extended channel** corresponding to *N* **uses of the channel**.
- The extended channel has $|A_X|^N$ possible inputs **x** and $|A_Y|^N$ possible outputs.
- Extended channels obtained from a **binary symmetric channel** with $f = 0.15$



Extended channels

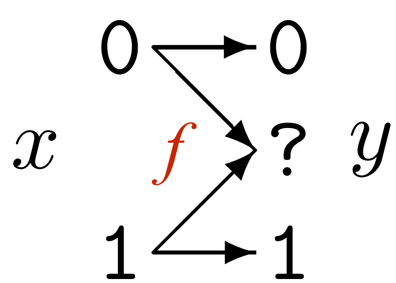
- Extended channels obtained from a **Z channel** with $f = 0.15$





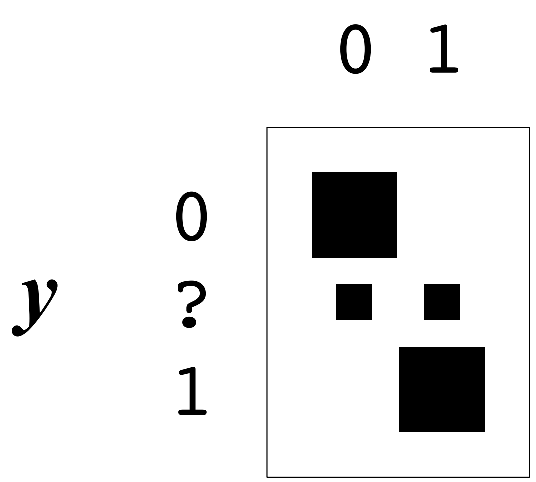
Binary erasure channel

- $A_X = \{0, 1\}$; $A_Y = \{0, \mathbf{?}, 1\}$.



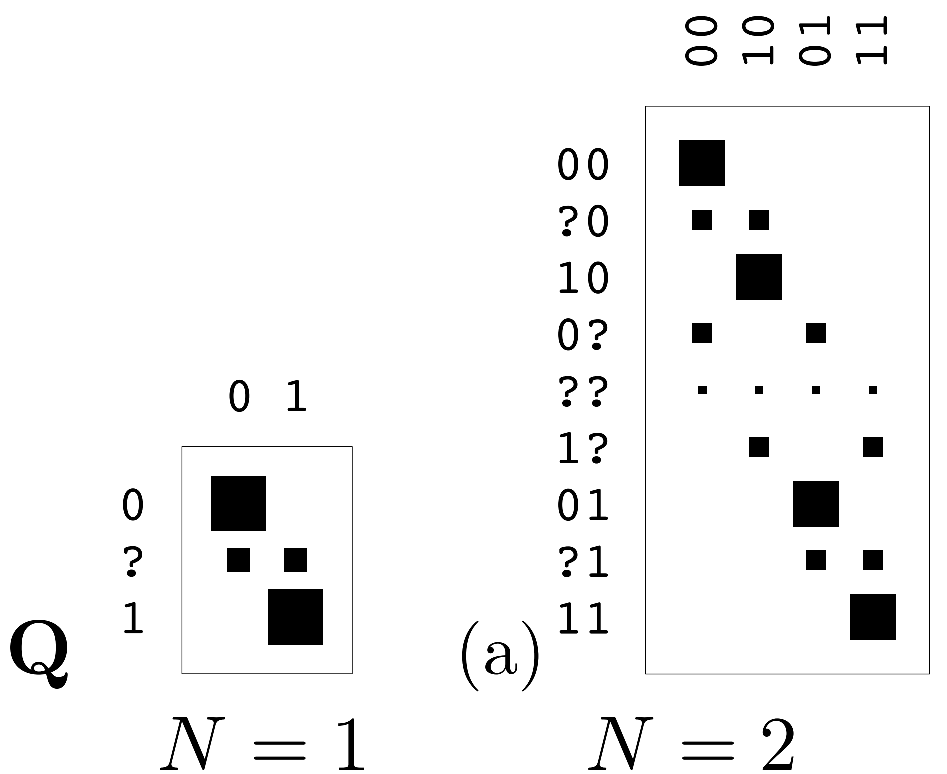
$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= 0; \\ P(y=\text{?} \mid x=0) &= f; & P(y=\text{?} \mid x=1) &= f; \\ P(y=1 \mid x=0) &= 0; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$

- f is the probability of erasing a bit.
  - So we assume that $f < 0.5$



Extended channels for the binary erasure channel

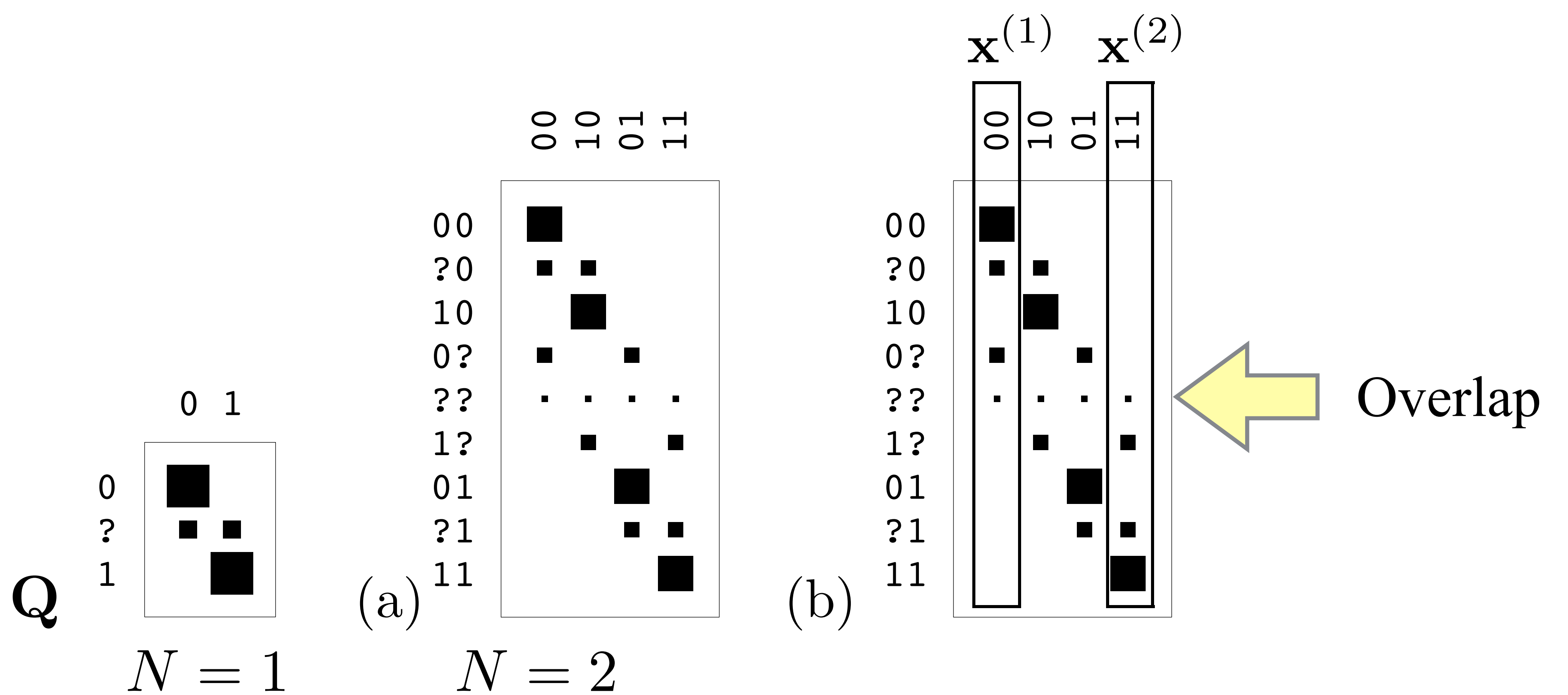
- For $N = 2$



- The best code for this channel with $N = 2$ is obtained by **choosing two columns that have minimal overlap.**

Extended channels for the binary erasure channel

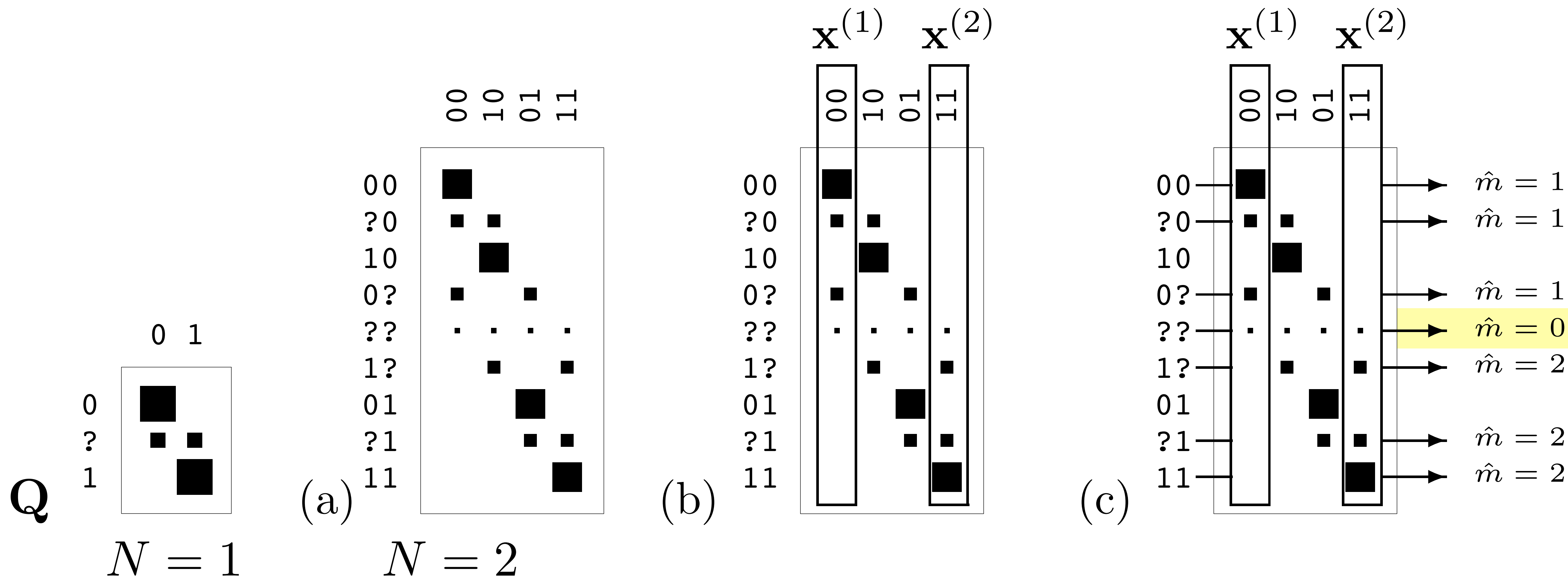
- For $N = 2$



- The best code for this channel with $N = 2$ is obtained by **choosing two columns that have minimal overlap**, for example, columns 00 and 11.

Extended channels for the binary erasure channel

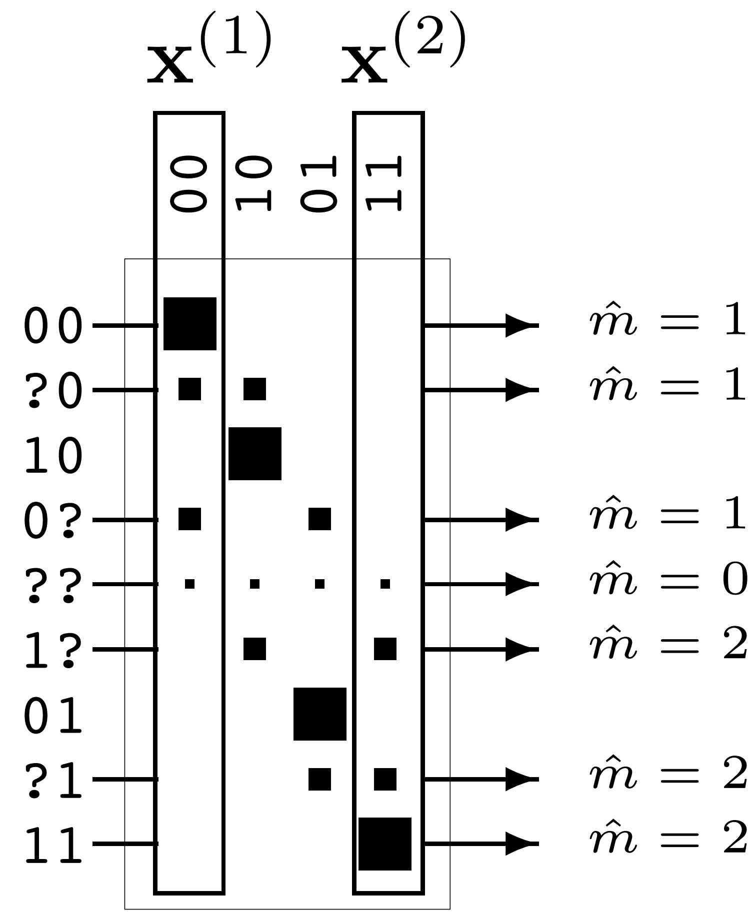
- For $N = 2$



- The decoding algorithm returns '00' if the extended channel output is among the top four and '11' if it's among the bottom four, and **gives up if the output is '??'**.

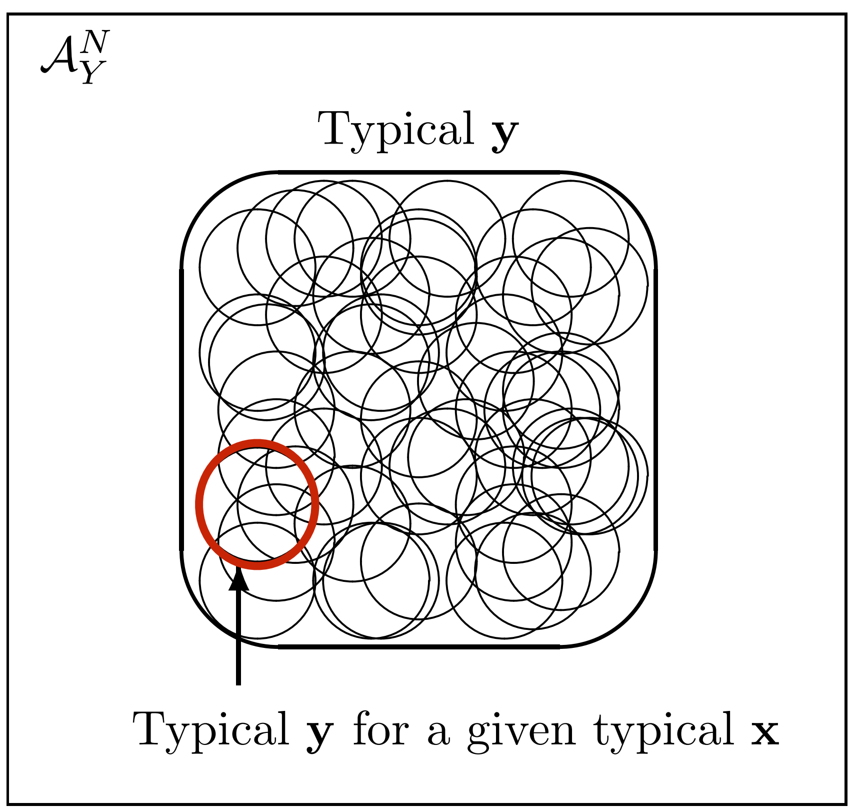
Intuitive preview of proof

- To prove the **noisy-channel coding theorem**, we make **use of large block-lengths** N.
- The intuitive idea is that, if N is large, an extended channel looks a lot like the noisy typewriter.
- Any particular input $\mathbf{x}$ is very likely to produce an output in a **small subspace of the output alphabet** – the typical output set, given that input.
- So we can find a **non-confusable subset of the inputs** that **produce essentially disjoint output sequences**.



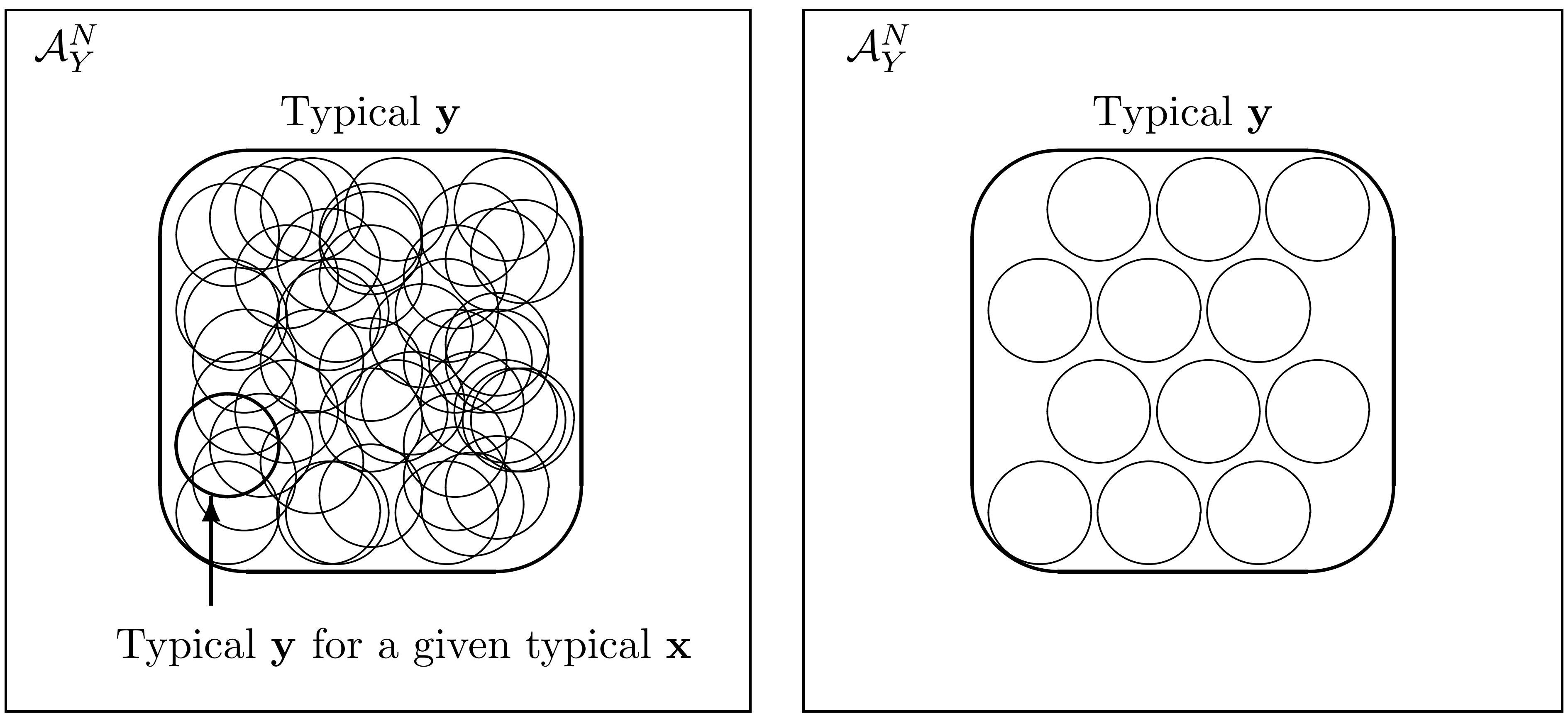
Intuitive preview of proof

- For a given N, how to **generating such a non-confusable subset of the inputs**, and **count up how many distinct inputs it contains**?
- Let $\mathbf{x}$ be an input sequence for the extended channel by drawing it from an ensemble X^N
- The **total number of typical output sequences y** is $2^{NH(Y)}$.
- For any particular typical input sequence $\mathbf{x}$, there are about $2^{NH(Y|X)}$ probable sequences



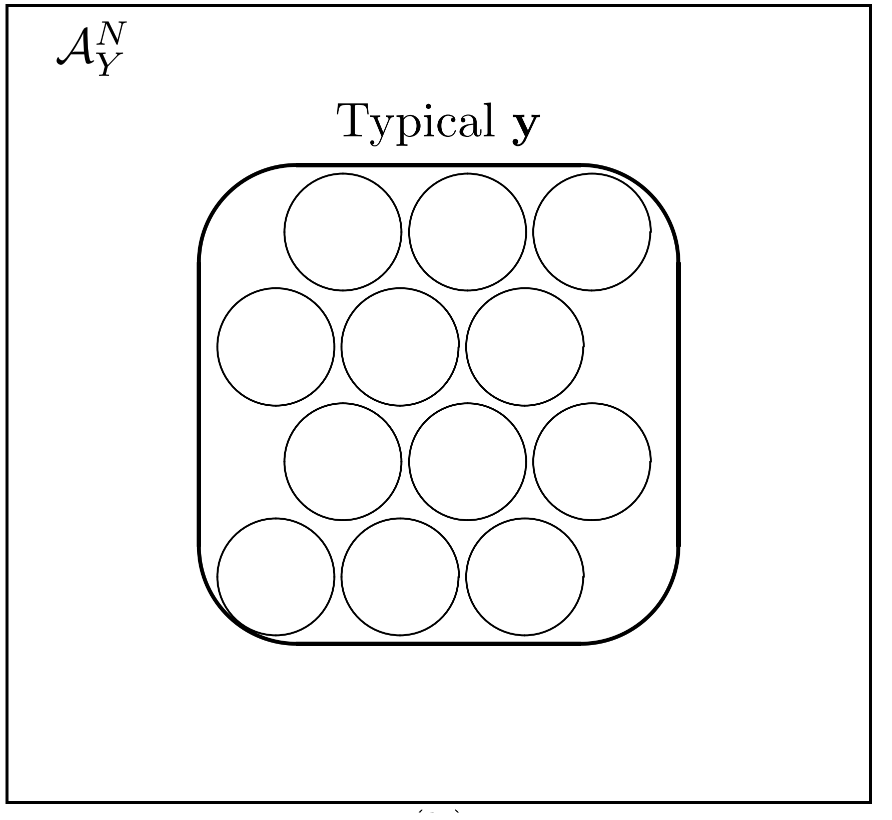
Intuitive preview of proof

- We now imagine **restricting ourselves to a subset of the typical inputs x** such that the corresponding typical output sets do not overlap.
- We can then **bound the number of non-confusable inputs** by dividing the size of the typical **y** set, $2^{NH(Y)}$, by the size of each typical-y given-typical-x set, $2^{NH(Y|X)}$



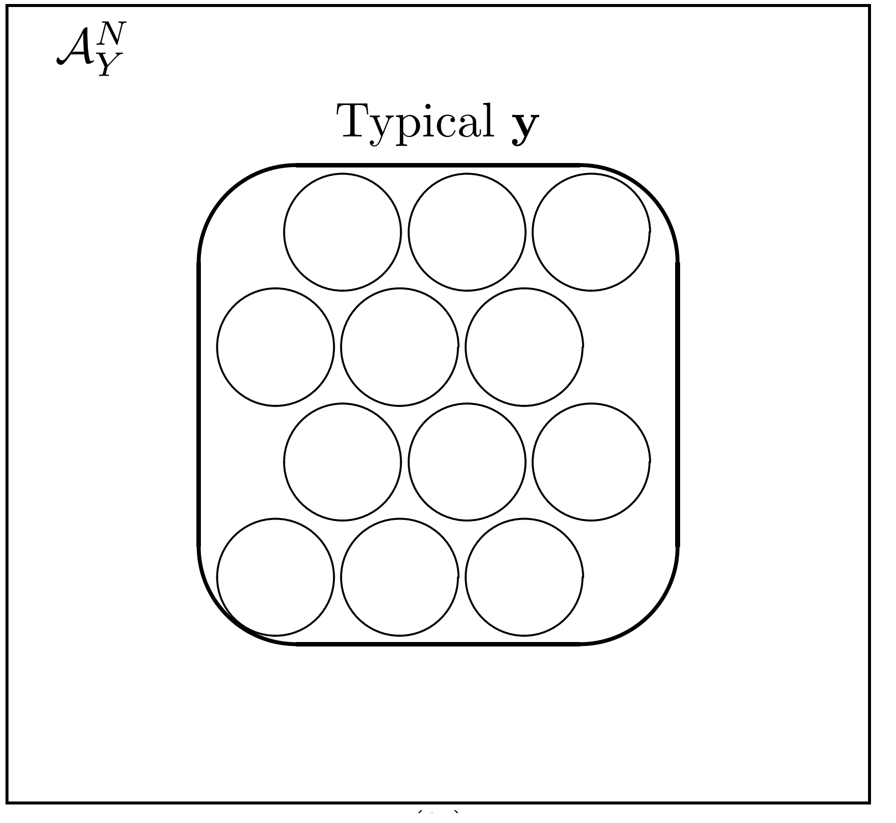
Intuitive preview of proof

- We now imagine restricting ourselves to a subset of the typical inputs x such that the corresponding typical output sets do not overlap.
- We can then **bound the number of non-confusable inputs** by dividing the size of the typical **y** set, $\mathbf{2}^{\boldsymbol{NH(Y)}}$, by the size of each typical-y given-typical-x set, $\mathbf{2}^{\boldsymbol{NH(Y|X)}}$
- So the number of non-confusable inputs, if they are selected from the set of typical inputs $\mathbf{x} \sim X^N$, is $\leq 2^{NH(Y)-NH(Y \mid X)} = 2^{NI(X;\,Y)}$.
- The maximum value of this bound is achieved if X is the ensemble that maximizes $I(X;\,Y)$, in which case the number of non-confusable inputs is $\leq 2^{NC}$



Intuitive preview of proof

- We can then **bound the number of non-confusable inputs** by dividing the size of the typical **y** set, $\mathbf{2^{NH(Y)}}$, by the size of each typical-y given-typical-x set, $\mathbf{2^{NH(Y|X)}}$
- So the number of non-confusable inputs, if they are selected from the set of typical inputs $\mathbf{x} \sim X^N$, is $\leq 2^{NH(Y)-NH(Y\mid X)} = 2^{NI(X;\,Y)}$.
- The maximum value of this bound is achieved if X is the ensemble that maximizes $I(X;\,Y)$, in which case the number of non-confusable inputs is $\leq 2^{NC}cv$
- Thus asymptotically **up to *C* bits per cycle**, and no more, **can be communicated with vanishing error probability.**



Further Reading and Summary

Q&A

Further Reading

- **Recommend Readings**
  - Information Theory, Inference, and Learning Algorithms from David MacKay, 2015, pages 146 - 160.
- **Supplemental readings:**

What you should know

- What is the purpose of source code and the purpose of channel code.
- The idea that the information transmitted depends on the input probability distribution
- Some common channels: BSC, Z, EBC, TypeWitter
- How to infer the input based on the output
- The Channel capacity and the mutual information; the concept of optimal input distribution
- How to compute a channel capacity for some common channels
- The concepts of probability of block error, the maximal probability of block error and The probability of bit error.
- Understanding the Shannon's noisy-channel coding theorem (part one) and the corresponding general strategy for channel coders.

Further Reading and Summary

Q&A