Communication Over A Noisy Channel

Mário S. Alvim
(msalvim@dcc.ufmg.br)

Information Theory

DCC-UFMG
(2017/02)

Communication Over A Noisy Channel - Introduction

- In the lectures about data compression we have implicitly assumed that the data generated by the compressor arrived perfectly at the decompressor.

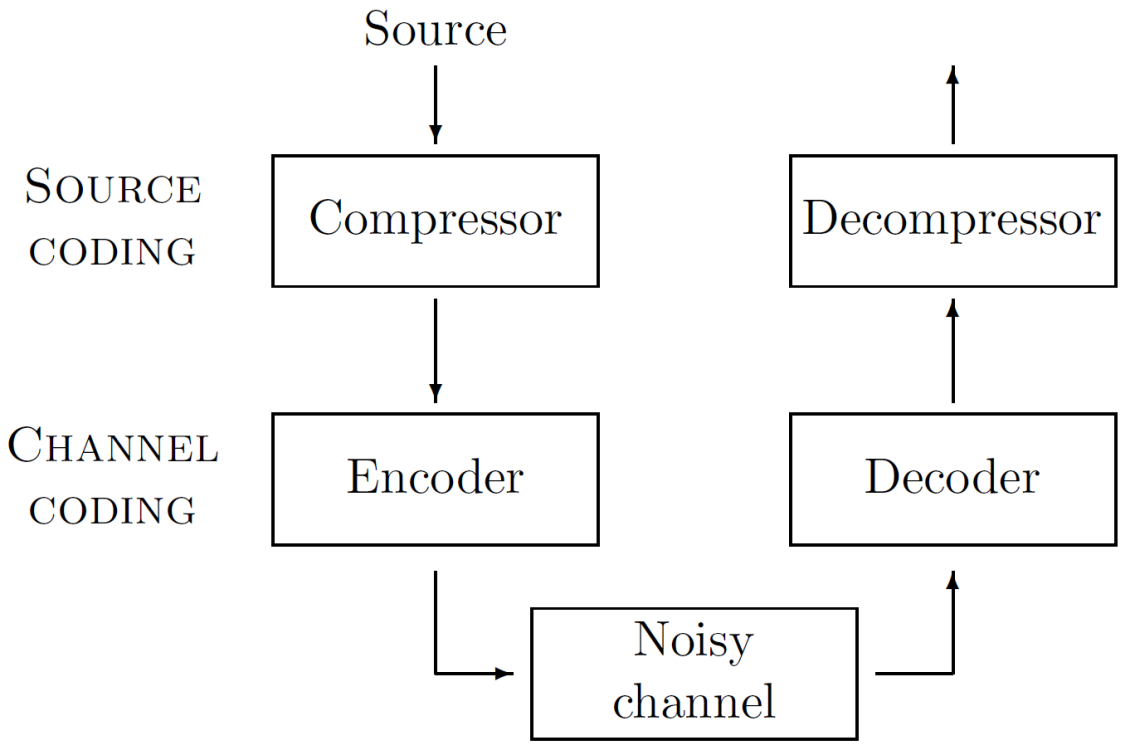
  In other words, we have assumed that the channel between compressor and decompressor was noiseless.

- Real-life channels are not perfect; they are noisy.

  The aim of channel-coding is to make a noisy channel to behave as a noiseless channel and transmit information reliably.

Communication Over A Noisy Channel - Introduction

- Channel-coding introduces some redundancy into the data so to preserve information from being corrupted.



- Combined wisely, the redundancy-removal process of data compression and the redundancy-addition process of channel coding allow the reliable and efficient communication methods.

Communication Over A Noisy Channel - Introduction

- The following two examples show the subtleties involved in measuring the information transmitted through a noisy channel.

- Example 1 Suppose we have a binary source in which:

  a) $p_0 = p_1 = 1/2$, and

  b) we transmit 1 000 bits from this source per second

  c) over a channel that has probability $f = 0.1$ of flipping a bit.

  Intuitively, the real rate of information transmission is not 1 000 bits per second, since 10% of bits are lost.

  What is the real transmission rate in this case?

Communication Over A Noisy Channel - Introduction

- Example 1 (Continued)

  We might guess that it is 900 bits per second, but this is not correct, because the receiver does not know where the errors occur, so the receiver cannot sort what bits were correctly received from the bits that were corrupted.

  We need a more precise way to measure the information transmitted through a channel!

◁

Communication Over A Noisy Channel - Introduction

- Example 2 Suppose we have the same binary source in which $p_0 = p_1 = 1/2$ and we transmit 1 000 bits from this source per second over a channel that has probability $f = 0.5$ of flipping a bit.

  What is the real transmission rate in this case?

  In this case no transmission of information occurs: every observed bit by the receiver has an equal chance of being correct or have been flipped, so there is no way to infer what the transmitted message was.

  ◁

Review of probability and information

- As the previous examples show, measuring the information transmitted through a channel is not as straightforward as one may believe at first.
- To measure the information precisely, we will use the concepts of entropy and mutual information.

  The following example reviews some concepts of probabilities, entropy and mutual information.

Review of probability and information

- Example 3 Consider the following joint ensemble XY.

$P(x,y)$		x				$P(y)$
		1	2	3	4	
	1	$1/8$	$1/16$	$1/32$	$1/32$	$1/4$
y	2	$1/16$	$1/8$	$1/32$	$1/32$	$1/4$
	3	$1/16$	$1/16$	$1/16$	$1/16$	$1/4$
	4	$1/4$	0	0	0	$1/4$
$P(x)$		$1/2$	$1/4$	$1/8$	$1/8$	

The joint entropy is $H(X,Y) = 27/8$ bits.

The marginal entropies are $H(X) = 7/4$ bits and $H(Y) = 2$ bits.

Review of probability and information

- Example 3 (Continued)

We can compute the conditional probabilities of x for each value of y:

$P(x \mid y)$		x				$H(X \mid y)$/bits
		1	2	3	4	
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0
						$H(X \mid Y) = 11/8$

Note that:

a) $H(X \mid y = 4) < H(X) < H(X \mid y = 3)$, so particular values of y can increase or decrease the uncertainty about X.

Review of probability and information

- Example 3 (Continued)

We can compute the conditional probabilities of x for each value of y:

$P(x \mid y)$		x				$H(X \mid y)$/bits
		1	2	3	4	
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0
						$H(X \mid Y) = 11/8$

Note that:

b) $p(x \mid y = 2) \neq p(x)$ but $H(X \mid y = 2) = H(X)$, so particular values of y may not alter the uncertainty about X (even if they alter the distribution on x).

Review of probability and information

- Example 3 (Continued)

We can compute the conditional probabilities of x for each value of y:

$P(x \mid y)$		x				$H(X \mid y)$/bits
		1	2	3	4	
y	1	$1/2$	$1/4$	$1/8$	$1/8$	$7/4$
	2	$1/4$	$1/2$	$1/8$	$1/8$	$7/4$
	3	$1/4$	$1/4$	$1/4$	$1/4$	2
	4	1	0	0	0	0

$$H(X \mid Y) = 11/8$$

Note that:

c) On average we have always that $H(X \mid Y) \leq H(X)$. In other words, learning the value of y conveys information about x.

The mutual information is

$$I(X;Y) = H(X) - H(X \mid Y) = 7/4 - 11/8 = 3/8 \text{ bits.}$$

Noisy Channels

Noisy Channels

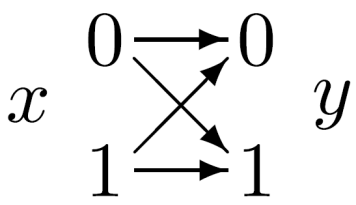
- A **discrete memoryless channel (DMC) Q** is a triple $(\mathcal{A}_X, \mathcal{A}_Y, \mathcal{P}_{Y|X})$ where
  - $\mathcal{A}_X$ is the **input alphabet** for the channel,
  - $\mathcal{A}_Y$ is the **output alphabet** for the channel, and
  - $\mathcal{P}_{Y|X}$ is a set of conditional probability distributions $p(y \mid x)$ indicating the probability of the channel producing output $y \in \mathcal{A}_Y$ when the input is $x \in \mathcal{A}_X$.
- For a channel Q, the set of distributions $\mathcal{P}_{Y|X}$ is called the **transition distributions** or the **channel matrix**.

  We denote by $Q_{j|i}$ or $Q(i,j)$ the probability

$$Q(i,j) = Q_{j|i} = p(y = b_j \mid x = a_i).$$

Noisy Channels

- Example 4 **(Binary symmetric channel.)**



Channel input alphabet: $\mathcal{A}_X = \{0, 1\}$

Channel output alphabet: $\mathcal{A}_Y = \{0, 1\}$

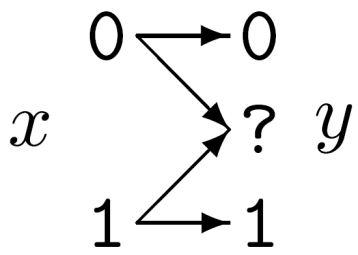
Channel matrix:

$p(y \mid x)$	$y = 0$	$y = 1$
$x = 0$	$1 - f$	f
$x = 1$	f	$1 - f$

◁

Noisy Channels

- Example 5 **(Binary erasure channel.)**



Channel input alphabet: $\mathcal{A}_X = \{0, 1\}$

Channel output alphabet: $\mathcal{A}_Y = \{0, 1, ?\}$

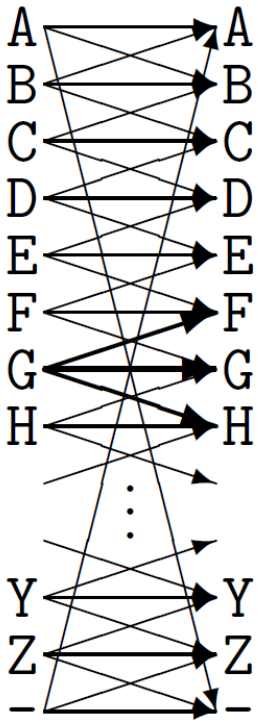
Channel matrix:

$p(y \mid x)$	$y = 0$	$y = 1$	$y = ?$
$x = 0$	$1 - f$	0	f
$x = 1$	0	$1 - f$	f

◁

Noisy Channels

- Example 6 **(Noisy typewriter.)**



Channel input alphabet: $\mathcal{A}_X = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \ldots, \mathtt{Z}, \mathtt{-}\}$

Channel output alphabet: $\mathcal{A}_Y = \{\mathtt{A}, \mathtt{B}, \mathtt{C}, \ldots, \mathtt{Z}, \mathtt{-}\}$

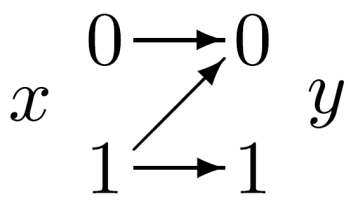
Channel matrix:

$p(y \mid x)$	$y = \mathtt{A}$	$y = \mathtt{B}$	$y = \mathtt{C}$	$\ldots$	$y = \mathtt{Y}$	$y = \mathtt{Z}$	$y = \mathtt{-}$
$x = \mathtt{A}$	1/3	1/3	0	$\ldots$	0	0	1/3
$x = \mathtt{B}$	1/3	1/3	1/3	$\ldots$	0	0	0
$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$	$\vdots$
$x = \mathtt{Z}$	0	0	0	$\ldots$	1/3	1/3	1/3
$x = \mathtt{-}$	1/3	0	0	$\ldots$	0	1/3	1/3

◁

Noisy Channels

- Example 7 **(Z channel.)**



Channel input alphabet: $\mathcal{A}_X = \{0, 1\}$

Channel output alphabet: $\mathcal{A}_Y = \{0, 1\}$

Channel matrix:

$p(y \mid x)$	$y = 0$	$y = 1$
$x = 0$	1	0
$x = 1$	f	$1 - f$

◁

Information Conveyed by a Channel

Inferring the input given the output

- If we are given:

  1. a channel $Q = (\mathcal{A}_X, \mathcal{A}_Y, \mathcal{P}_{Y|X})$ from X to Y, and
  2. an ensemble $X = (x, \mathcal{A}_X, \mathcal{P}_X)$ representing the input to the channel,

  we can derive the distribution on the joint ensemble XY:

$$p(x, y) = p(x)p(y \mid x).$$

- If the channel produces a particular symbol $y \in \mathcal{A}_Y$, we can calculate the probability that a particular input x was fed to the channel using Bayes' Theorem:

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

  and we can infer the input by picking the x that maximizes $p(x \mid y)$.

Inferring the input given the output

- **Example 8** Consider a binary symmetric channel with probability of error $f = 0.15$.

  Let the input ensemble be $\mathcal{P}_X : \{p_0 = 0.9, p_1 = 0.1\}$.

  The general behavior of this channel with this prior can be derived by first computing the joint distribution $\mathcal{P}_{X,Y}$:

$\mathcal{P}_X$	
$x = 0$	0.9
$x = 1$	0.1

prior probability $p(x)$

$\cdot$

$\mathcal{P}_{Y\|X}$	$y = 0$	$y = 1$
$x = 0$	0.85	0.15
$x = 1$	0.15	0.85

channel $p(y \mid x)$

$\Rightarrow$

$\mathcal{P}_{X,Y}$	$y = 0$	$y = 1$
$x = 0$	0.765	0.135
$x = 1$	0.015	0.085

joint probability $p(x, y)$,

and from that deriving the a posteriori distributions $\mathcal{P}_{X|Y}$ and the marginal distribution $\mathcal{P}_Y$:

$\mathcal{P}_Y$	$y = 0$	$y = 1$
	0.78	0.22

probability $p(y)$

$\mathcal{P}_{X\|Y}$	$y = 0$	$y = 1$
$x = 0$	0.981	0.614
$x = 1$	0.019	0.386

posterior probability $p(x \mid y)$.

Inferring the input given the output

- Example 8 (Continued)

From the behavior of the channel under this prior we can infer the following.

$\mathcal{P}_Y$	$y=0$	$y=1$
	0.78	0.22

probability $p(y)$

$\mathcal{P}_{X\|Y}$	$y=0$	$y=1$
$x=0$	0.981	0.614
$x=1$	0.019	0.386

posterior probability $p(x \mid y)$.

1. Output $y = 0$ will be produced with probability 0.78, and when that happen the distribution on x will be $\mathcal{P}X \mid y = 0 = (0.981, 0.019)$.

   If $y = 0$ is observed, the best inference is that x was 0.

2. output $y = 1$ will be produced with probability 0.22, and when that happen the distribution on x will be $\mathcal{P}X \mid y = 0 = (0.614, 0.386)$.

   If $y = 1$ is observed, the best inference is that x was 0.

◁

Inferring the input given the output

- **Example 9** Consider a Z channel with probability of error $f = 0.15$.

  Let the input ensemble be $\mathcal{P}_X : \{p_0 = 0.9, p_1 = 0.1\}$.

  The general behavior of this channel with this prior can be derived by first computing the joint distribution $\mathcal{P}_{X,Y}$:

$\mathcal{P}_X$	
$x = 0$	0.9
$x = 1$	0.1

prior probability $p(x)$

$\cdot$

$\mathcal{P}_{Y\|X}$	$y = 0$	$y = 1$
$x = 0$	1	0
$x = 1$	0.15	0.85

channel $p(y \mid x)$

$\Rightarrow$

$\mathcal{P}_{X,Y}$	$y = 0$	$y = 1$
$x = 0$	0.9	0
$x = 1$	0.015	0.085

joint probability $p(x, y)$,

and from that deriving the a posteriori distributions $\mathcal{P}_{X|Y}$ and the marginal distribution $\mathcal{P}_Y$:

$\mathcal{P}_Y$	$y = 0$	$y = 1$
	0.915	0.085

probability $p(y)$

$\mathcal{P}_{X\|Y}$	$y = 0$	$y = 1$
$x = 0$	0.984	0
$x = 1$	0.016	1

posterior probability $p(x \mid y)$.

Inferring the input given the output

- Example 9 (Continued)

From the behavior of the channel under this prior we can infer the following.

$\mathcal{P}_Y$	$y=0$	$y=1$
	0.915	0.085

probability $p(y)$

$\mathcal{P}_{X\|Y}$	$y=0$	$y=1$
$x=0$	0.984	0
$x=1$	0.016	1

posterior probability $p(x \mid y)$.

1. Output $y = 0$ will be produced with probability 0.915, and when that happen the distribution on x will be $\mathcal{P}X \mid y = 0 = (0.984, 0.016)$.

   If $y = 0$ is observed, the best inference is that x was 0.

2. output $y = 1$ will be produced with probability 0.085, and when that happen the distribution on x will be $\mathcal{P}X \mid y = 0 = (0, 1)$.

   If $y = 1$ is observed, the best inference is that x was 1.

◁

Information conveyed by a channel

- Let us consider how much information is communicated through a channel.
- For a particular input ensemble X you have an uncertainty $H(X)$.

  After you observe the output Y of the channel, you have an uncertainty $H(X \mid Y)$ about X.

  Intuitively, the information the output Y conveys about X is the difference in uncertainty before and after the chain was used:

$$\overbrace{I(X;Y)}^{\left(\begin{array}{c}\text{information } Y\\ \text{conveys about } X\end{array}\right)} = \overbrace{H(X)}^{\left(\begin{array}{c}\text{prior uncertainty}\\ \text{about } X\end{array}\right)} - \overbrace{H(X \mid Y)}^{\left(\begin{array}{c}\text{uncertainty about } X\\ \text{once } Y \text{ is konwn}\end{array}\right)}$$

- We say that the value $I(X;Y)$ is the amount of bits transmitted per use of the channel, since each use of the channel is expected to decrease the uncertainty about the input by exactly $I(X;Y)$ bits.

Information conveyed by a channel

- Example 10 Consider the binary symmetric channel again, with $f = 0.15$ and $\mathcal{P}_X = \{p_0 = 0.9, p_1 = 0.1\}$.

  Compute the mutual information $I(X;Y)$.

  Solution.

  In a previous example we have already computed the behavior of the binary symmetric channel with this prior, finding:

$\mathcal{P}_Y$	$y=0$	$y=1$
	0.78	0.22

} probability $p(y)$

$\mathcal{P}_{X\|Y}$	$y=0$	$y=1$
$x=0$	0.981	0.614
$x=1$	0.019	0.386

} posterior probability $p(x \mid y)$.

Information conveyed by a channel

- Example 10 (Continued)

Now we can compute the mutual information $I(X;Y)$.

First:

$$H(X) = H(0.9,\, 0.1) = 0.469.$$

Then:

$$\begin{aligned} H(X \mid Y) &= p(y=0)H(X \mid y=0) + p(y=1)H(X \mid y=1) \\ &= 0.78\, H(0.981,\, 0.019) + 0.22\, H(0.614,\, 0.386) \\ &= 0.319. \end{aligned}$$

Hence

$$\begin{aligned} I(X;Y) &= H(X) - H(X \mid Y) \\ &= 0.469 - 0.319 \\ &= 0.150. \end{aligned}$$

Information conveyed by a channel

- Example 11 Consider again the channels of Example 1 and Example 2, together with the given input ensembles

  Find how many bits of information can actually be transmitted per second using each channel, for the given input ensembles.

  Solution.

  Homework!

  ◁

Channel capacity

- The mutual information between input and output in a channel depends on the probability distribution of the input ensemble.

  It is a natural question to ask what is the maximum mutual information between input and output the channel can provide.

- The **capacity of channel** Q is defined as

$$C(Q) = \max_{\mathcal{P}_X} I(X;Y).$$

The distribution $\mathcal{P}_X$ that achieves the maximum is called the **optimal input distribution**, and denoted by $\mathcal{P}_X^*$.

There may be multiple input distributions that are optimal.

Channel capacity

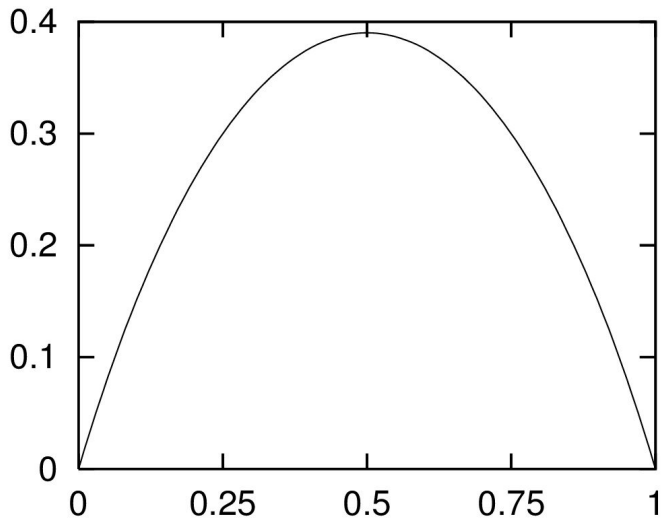
- Example 12 Consider the binary symmetric channel again, with $f = 0.15$.

  In a previous example we calculated that for input ensemble $\mathcal{P}_X = \{p_0 = 0.9, p_1 = 0.1\}$ the information flowing through the channel is $I(X;Y) = 0.15$ bits per use of the channel.

  If we are free to pick another input ensemble, how much better could we do?

  Plotting the mutual information $I(X;Y)$ as a function of the ensemble $\mathcal{P}_X = \{1 - p_1, p_1\}$, we get the following graph.

$I(X;Y)$



From this graph we conclude that capacity is achieved for the input ensemble $\mathcal{P}_X = \{1/2, 1/2\}$, with $I(X;Y) = 0.39$ bits.

◁

Channel capacity

- Example 13 For the noisy typewriter, the optimal distribution on the input ensemble is the uniform distribution.

  (You can find this via calculus, using derivatives).

  This means that the entropy of the optimal input is $H(X) = \log_2 27$.

  We can calculate the posterior entropy for the optimal input to be $H(X \mid Y) = \log_2 3$.

  Hence, capacity is $\log_2 27 - \log_2 3 = \log_2 9$ bits per use of the channel.

◁

The Noisy-Channel Coding Theorem

The Noisy-Channel Coding Theorem

- The **mathematical definition of channel capacity** is the maximum achievable mutual information between input and output in a given channel:

$$C(Q) = \max_{\mathcal{P}_X} I(X;Y)$$

- In practical scenarios, we are interested in finding ways to use the channel so that all bits transmitted through it can be reliably recovered.

  The **operational definition of channel capacity** is the maximum transmission rate you can achieve through a noisy channel if you want the transmission to be reliable (i.e., if you want the probability of incorrectly inferring the input from the output to be negligibly small).

The Noisy-Channel Coding Theorem

- The Noisy-Channel Coding Theorem provides the connection between the mathematical definition of capacity and the operational interpretation of capacity.
- Intuitively, the **Noisy-Channel Coding Theorem** states that:

  "The operational capacity of a channel is exactly its mathematical capacity."

  In other words, the theorem states that for any channel from X to Y, the maximum rate at which it is possible to transmit information through this channel with a negligible probability of error is $C(Q) = \max_{\mathcal{P}_X} I(X;Y)$.
- The Noisy-Channel Coding Theorem ensures there is a always a way of encoding data to transmit it through a noisy channel at a rate that achieves capacity.

  Finding what this encoding is, however, is a hard problem in the general case.

The Noisy-Channel Coding Theorem

- To better understand the Noisy-Channel Coding Theorem, let us formalize some concepts.
- An (N, K) **block code** for a channel Q is a list of $S = 2^K$ codewords

$$x^{(1)}, x^{(2)}, \ldots, x^{(2^K)}, \quad \text{for } X^{(s)} \in \mathcal{A}_X^N,$$

each of length N.

Using an (N, K) block code we can encode a signal $s \in \{1, 2, 3, \ldots, 2^K\}$ as $x^{(s)}$.

The **rate of the code** is

$$R = \frac{K}{N} \quad \text{bits per channel use.}$$

The Noisy-Channel Coding Theorem

- A **decoder** for an (N, K) block code is a mapping from the set of length-N strings of channel outputs,

$$\mathcal{A}_Y^N$$

to a codeword label

$$\hat{s} \in \{0, 1, 2, \ldots, 2^K\}.$$

The extra symbol $\hat{s} = 0$ can be used to indicate "failure".

The Noisy-Channel Coding Theorem

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

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

  where s_{in} is the input signal to the channel and s_{out} is the decoded output signal.

- The **maximum probability of block error** is

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

The Noisy-Channel Coding Theorem

- The **optimal decoder** for a channel code is the one that minimizes the probability of block error.

  It decodes an output y as the input s that has maximum posterior probability $p(s \mid y)$.

$$p(s \mid y) = \frac{p(s)p(y \mid s)}{\sum_{s'} p(s')p(y \mid s')}$$

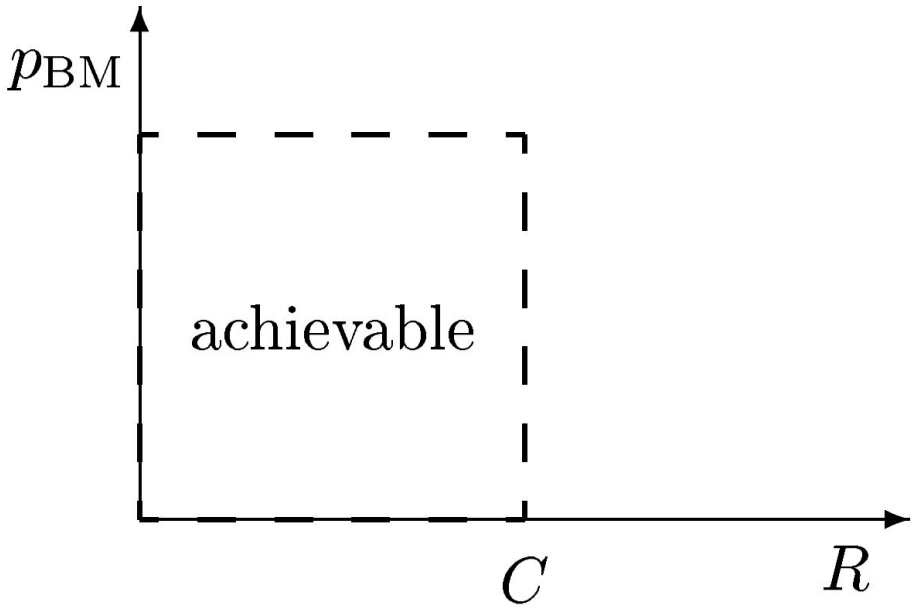
$$\hat{s}_{optimal} = \text{argmax}\, p(s \mid y)$$

The Noisy-Channel Coding Theorem

- **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 $\epsilon > 0$ and $R \leq C$, 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 $< \epsilon$.



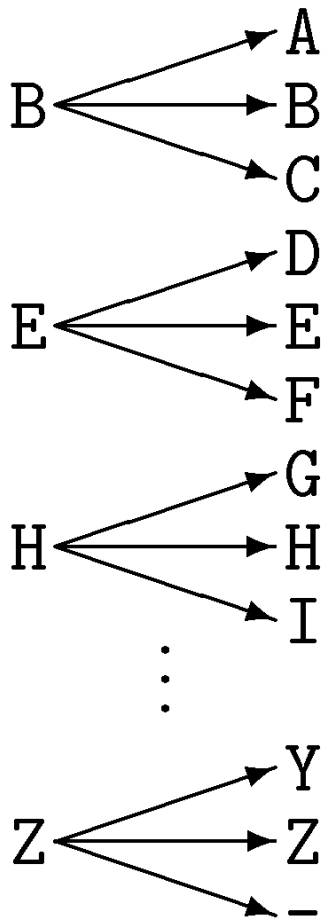
The Noisy-Channel Coding Theorem

- Example 14 Capacity of the Noisy typewriter.

We can get a completely error-free communication strategy using a block code of length $N = 1$: we use every third letter from the alphabet $\mathtt{B}, \mathtt{E}, \mathtt{H}, \ldots, \mathtt{Z}$.

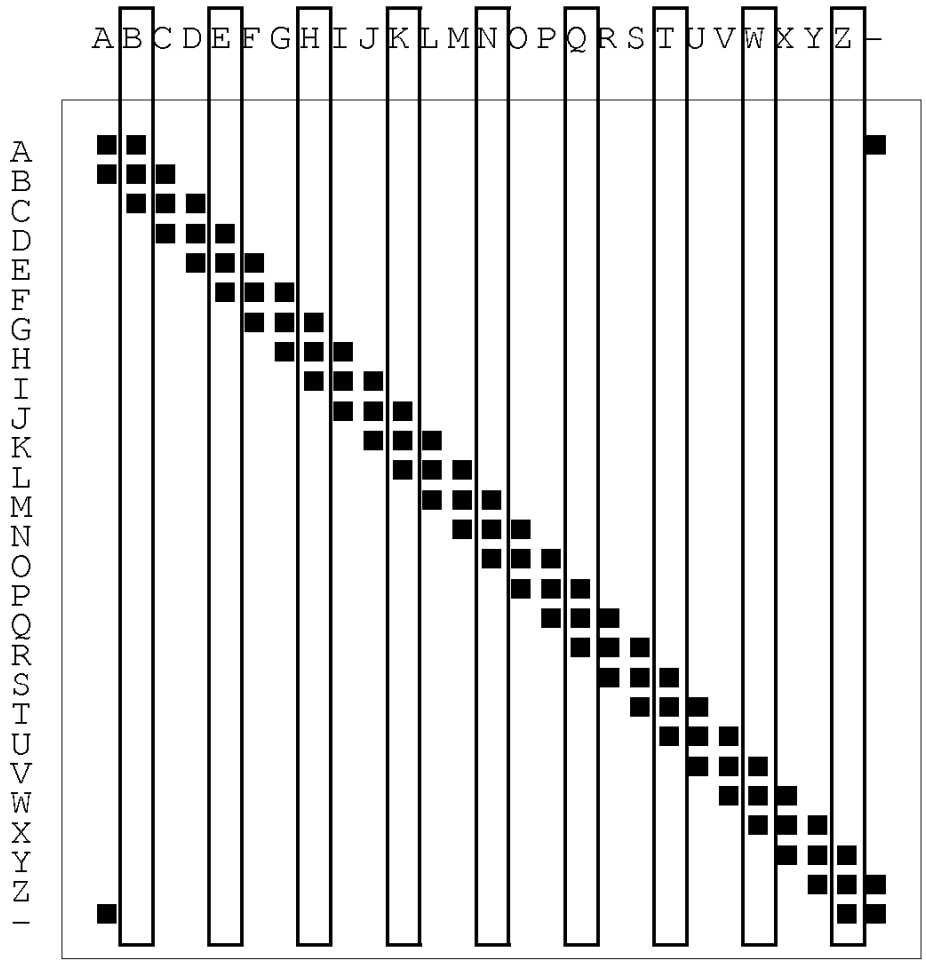
These letters from a **non-confusable subset** of the input alphabet.

The number of inputs in the non-confusable set is 9, so the error free information rate of this system is $\log_2 9$ bits, which is equal to the capacity of the channel.



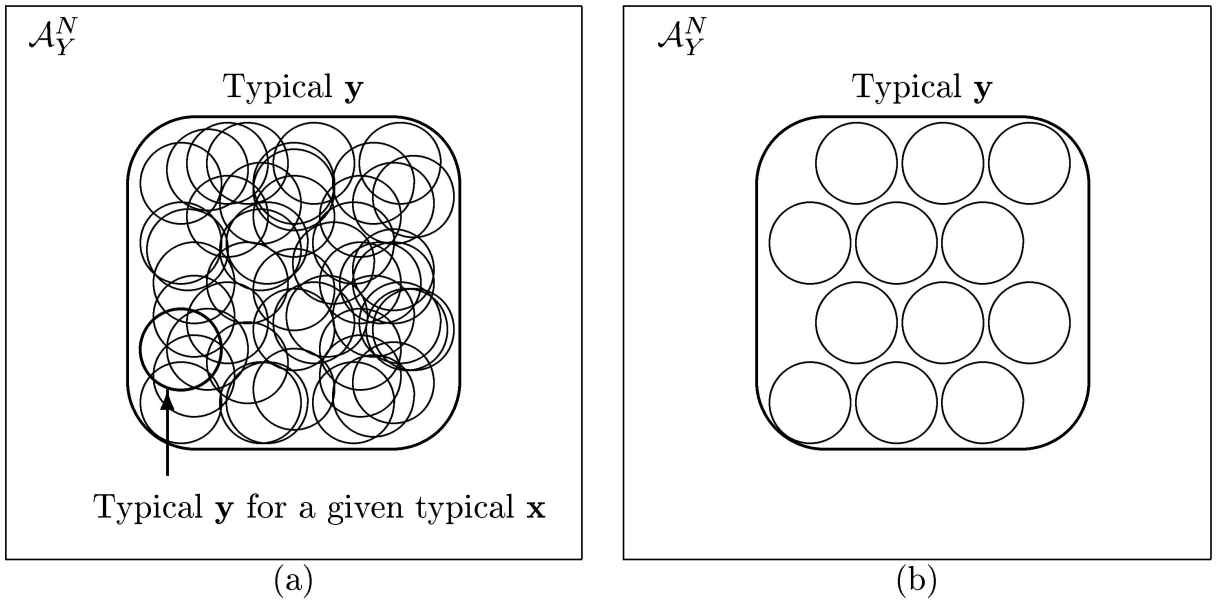
The Noisy-Channel Coding Theorem

- Example 14 (Continued)



The Noisy-Channel Coding Theorem

- The intuition behind the proof of part one of the theorem is to build a block-coding of size N large enough and look for how many non-confusable inputs this encoding generates.



There are $2^{NI(X;Y)}$ non-confusable inputs for a block-code of size N.

If we maximize over all inputs ensembles, we get 2^{NC} non-confusable inputs.

Hence, the maximum rate is $\log_2 2^{NC}/N = C$.

Appendix - Complete statement of Shannon's noisy-channel coding theorem

The Noisy-Channel Coding Theorem

- **Shannon's noisy-channel coding theorem.**

  1. For every discrete memoryless channel, the channel capacity

  $$C = \max_{\mathcal{P}_X} I(X;Y)$$

  has the following property. For any $\epsilon > 0$ and $R \leq C$, 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 $< \epsilon$.

  2. If a probability of bit error p_b is acceptable, rates up to $R(p_b)$ are achievable, where

  $$R_{p_b} = \frac{C}{1 - H_2(p_b)}.$$

  3. For any p_b, rates greater than $R(p_b)$ are not achievable.