09

The Noisy-Channel Coding Theorem

Notice

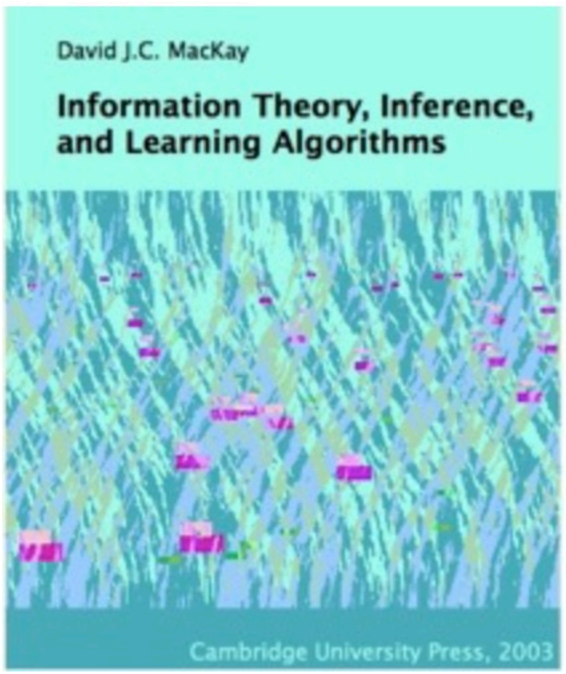
- **Author**
  - **João Moura Pires (jmp@fct.unl.pt)**

- **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

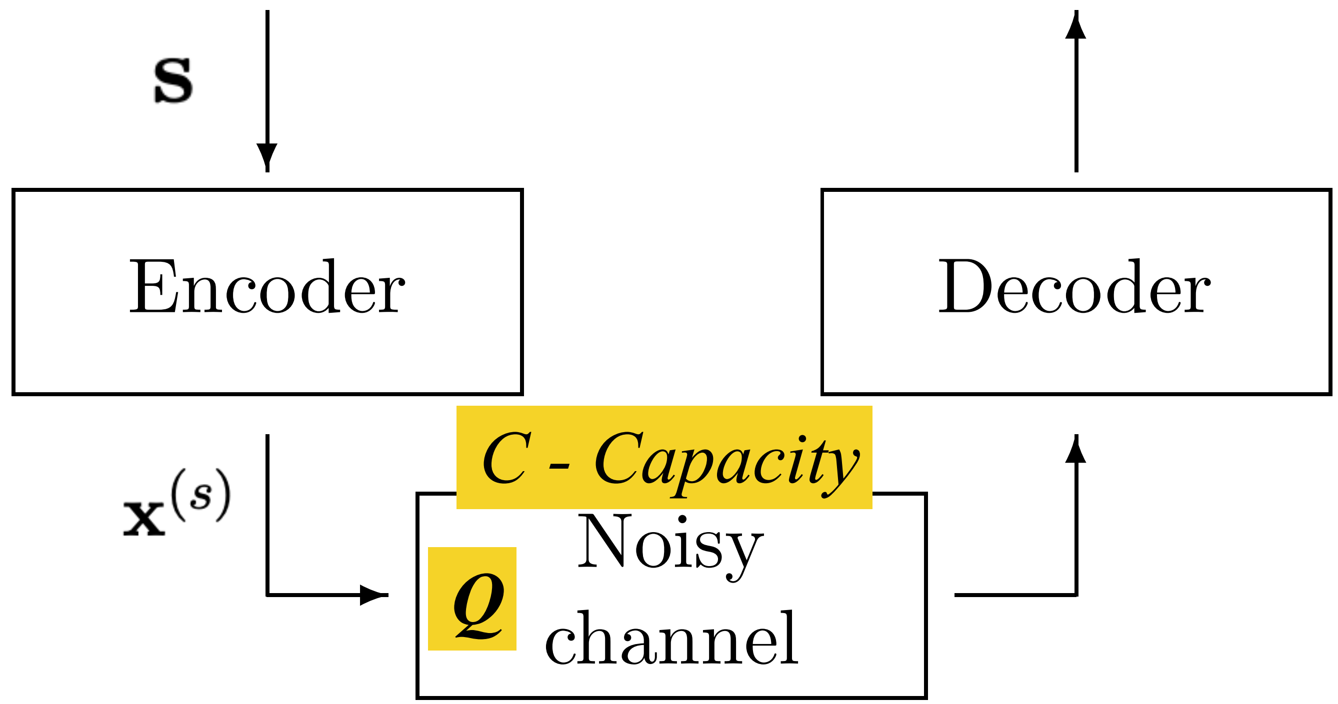
Notation

Notation

$K = \log_2 S$ - the number of bits conveyed by the choice of one codeword from S

a binary representation of the number s (length K)

CHANNEL CODING



$\hat{\mathbf{s}}$ the decoder's guess of s

X^N - an ensemble used to create a random code $\mathcal{C}$

N - the length of the codewords

$\mathbf{x}^{(s)}$ - a codeword, the sth in the code

s - the number of a chosen codeword

$R = K/N$ - the rate of the code, in bits per channel use

$S = 2^K$ - the total number of codewords

Notation

Q	the noisy channel
C	the capacity of the channel
X^N	an ensemble used to create a random code
$\mathcal{C}$	a random code
N	the length of the codewords
$\mathbf{x}^{(s)}$	a codeword, the sth in the code
s	the number of a chosen codeword (mnemonic: the *source* selects s)
$S = 2^K$	the total number of codewords in the code
$K = \log_2 S$	the number of bits conveyed by the choice of one codeword from S, assuming it is chosen with uniform probability
$\mathbf{s}$	a binary representation of the number s
$R = K/N$	the rate of the code, in bits per channel use (sometimes called R' instead)
$\hat{s}$	the decoder's guess of s

The theorem

The theorem

Part 1 - Positive

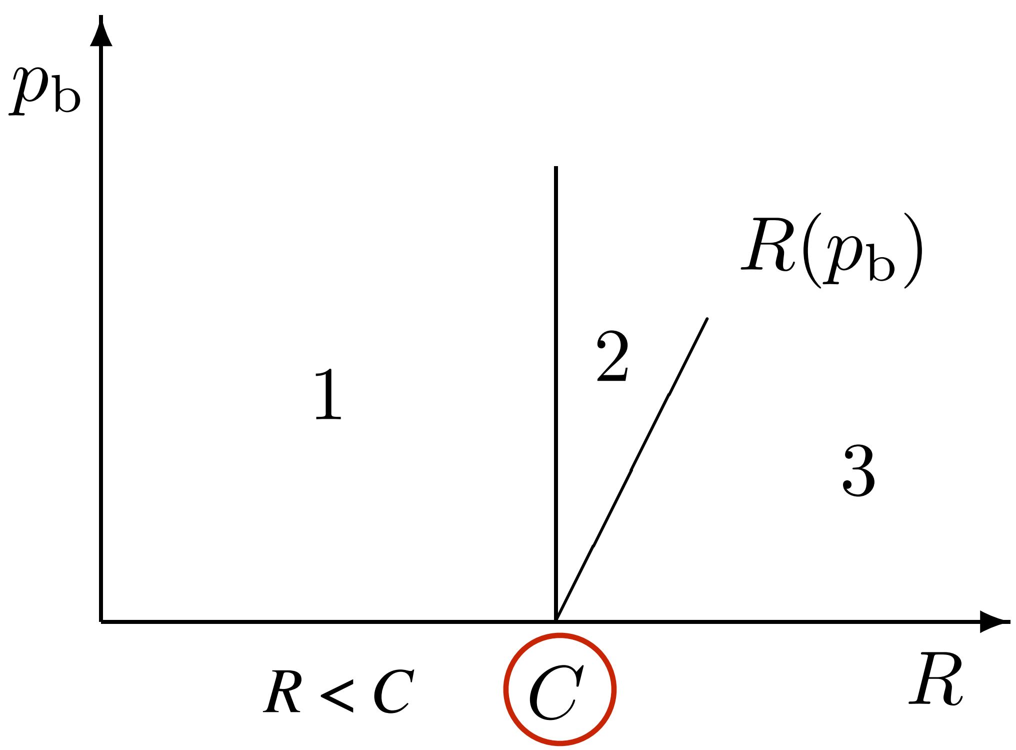
- For every discrete memoryless channel, the channel capacity

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

has the following property.

- For any $\varepsilon > 0$ and $R < C$, for large enough N, there exists a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \varepsilon$.

- $\varepsilon > 0$
- large enough N



The theorem

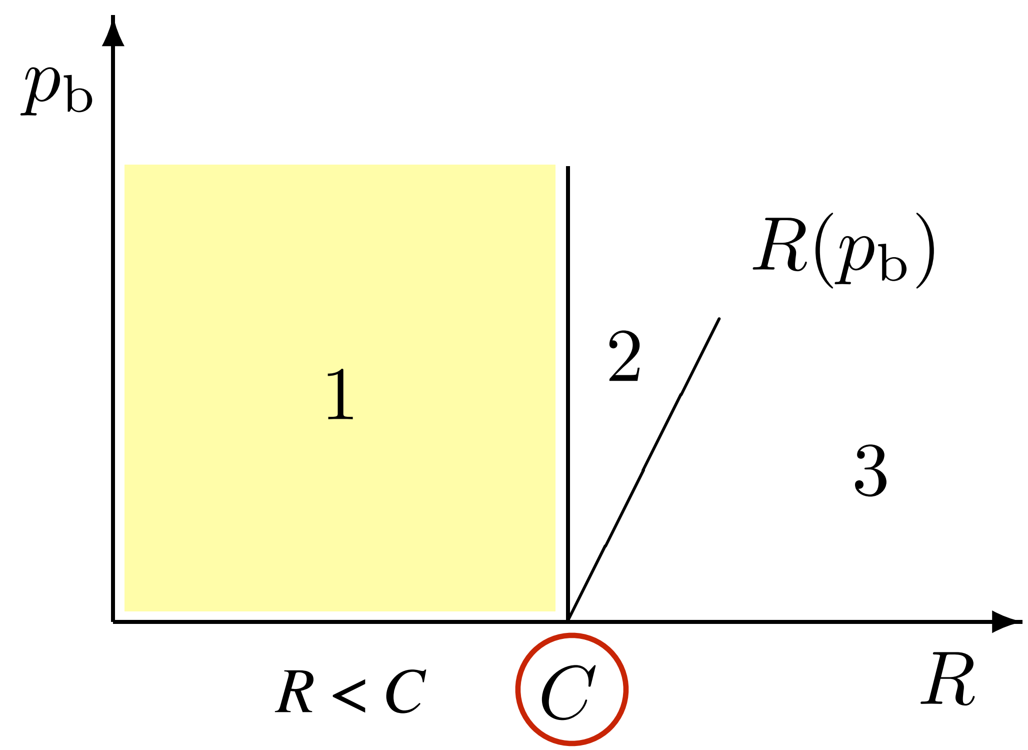
Part 1 - Positive

- For every discrete memoryless channel, the channel capacity

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

has the following property. For any $\varepsilon > 0$ and $R < C$, for large enough N, there **exists a code of length N and rate $\geq R$** and a **decoding algorithm, such that the maximal probability of block error is $< \varepsilon$**.

- $\varepsilon > 0$
- large enough N

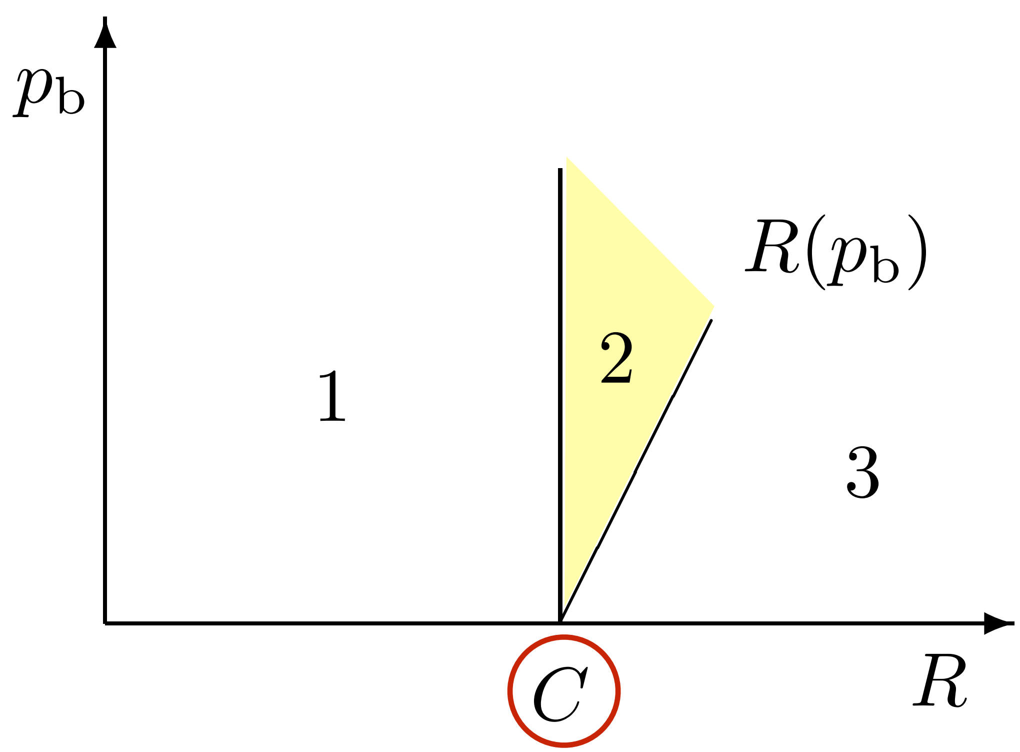


The theorem

Part 2 - Positive

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

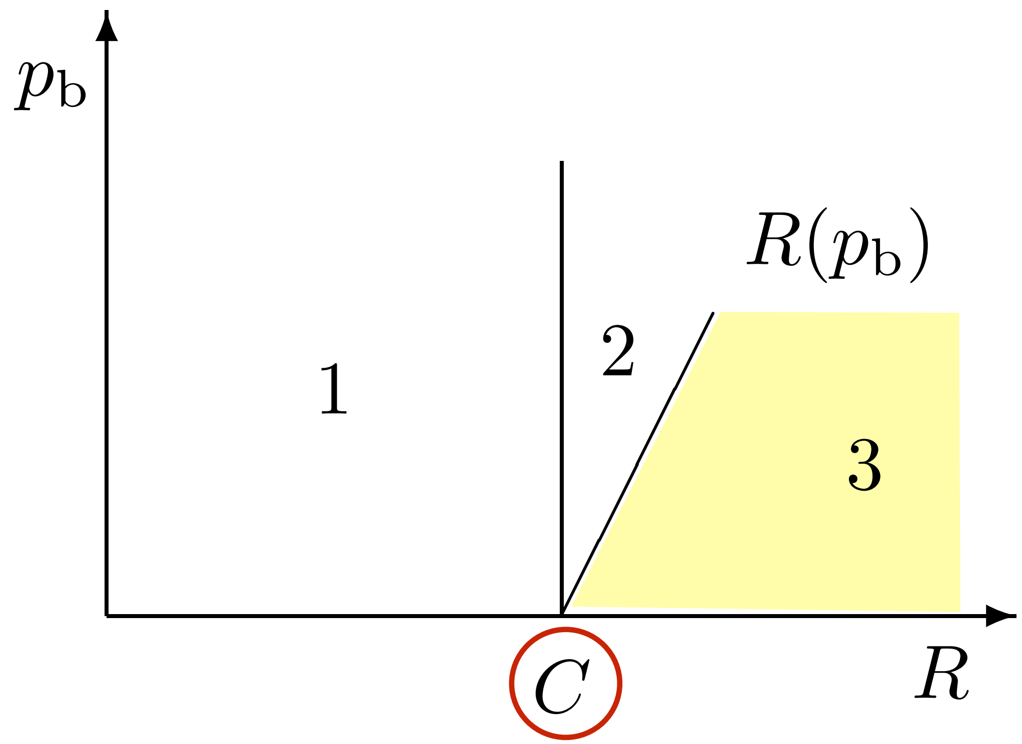
$$R(p_b) = \frac{1}{1 - H_2(p_b)}$$



Communication (with errors) above capacity

The theorem

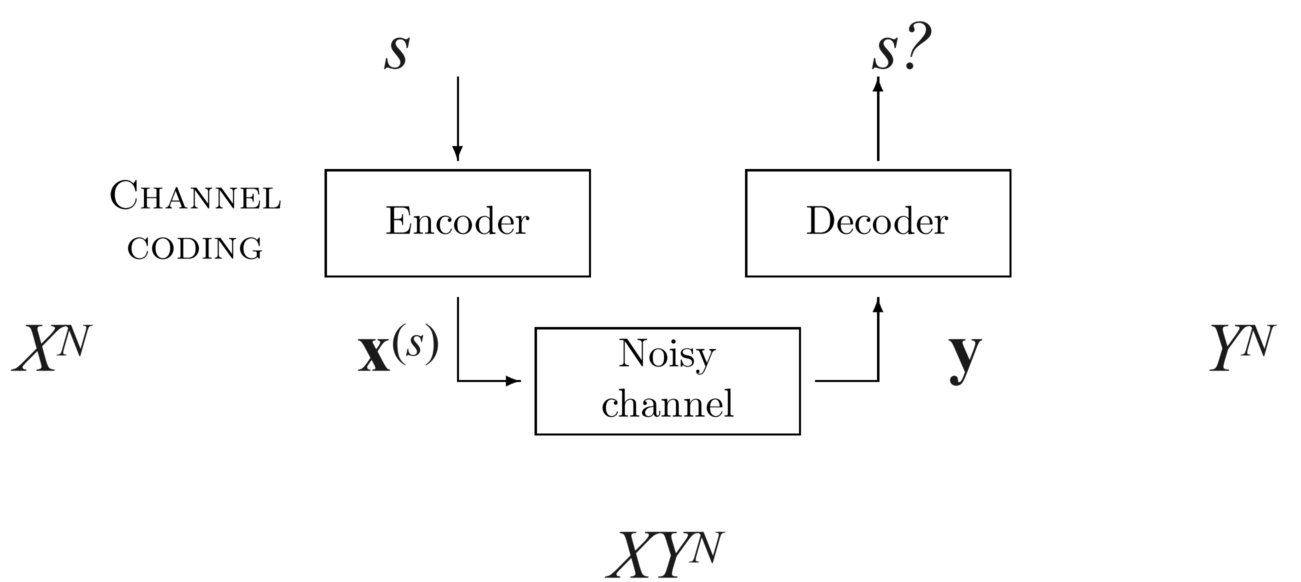
- For any p_b, rates greater than $R(p_b)$ are not achievable.



Jointly-typical sequences

Jointly-typical sequences

- We will define codewords $\mathbf{x}^{(s)}$ as coming from an ensemble X^N
- Consider the random selection of one codeword and a **corresponding channel output y,** thus defining a joint ensemble $(XY)^N$.
- **A typical-set decoder**, decodes a received signal $\mathbf{y}$ as s if $\mathbf{x}^{(s)}$ and $\mathbf{y}$ are **jointly typical**.



Jointly-typical sequences

- **Joint typicality**. A pair of sequences **x**, **y** of length N are defined to be jointly typical **(to tolerance β)** with respect to the distribution $P(x, y)$, if:

$$
\begin{aligned}
\mathbf{x} \text{ is typical of } P(\mathbf{x}), \quad \text{i.e.,} \quad & \left| \frac{1}{N} \log \frac{1}{P(\mathbf{x})} - H(X) \right| < \beta, \\
\mathbf{y} \text{ is typical of } P(\mathbf{y}), \quad \text{i.e.,} \quad & \left| \frac{1}{N} \log \frac{1}{P(\mathbf{y})} - H(Y) \right| < \beta, \\
\text{and } \mathbf{x}, \mathbf{y} \text{ is typical of } P(\mathbf{x}, \mathbf{y}), \quad \text{i.e.,} \quad & \left| \frac{1}{N} \log \frac{1}{P(\mathbf{x}, \mathbf{y})} - H(X, Y) \right| < \beta.
\end{aligned}
$$

- Example with $N = 100$
  - $P(x, y)$ in which $P(x)$ has $(p_0, p_1) = (0.9, 0.1)$ and $P(y \mid x)$ corresponds to a BSC ($f = 0.2$)

```
x   1111111111000000000000000000000000000000000000000000000000000000000000000000000000000000000000000000
y   0011111111000000000000000000000000000000000000000000000000000000000000000000000000111111111111111111
```

  - **x** has 10 1s, and so is typical of the probability $P(x)$ (at any tolerance β);
  - **y** has 26 1s, so it is typical of $P(y)$ (because $P(y = 1) = 0.26$)
  - **x** and **y** differ in 20 bits, which is the typical number of flips for this channel.

Joint typicality theorem

- **The jointly-typical set** $J_{N\beta}$ is the set of all jointly-typical sequence pairs of length N.
- **Joint typicality theorem.** Let $\mathbf{x}, \mathbf{y}$ be drawn from the ensemble $(XY)^N$ defined by

$$P(\mathbf{x}, \mathbf{y}) = \prod_{n=1}^{N} P(x_n, y_n).$$

then,

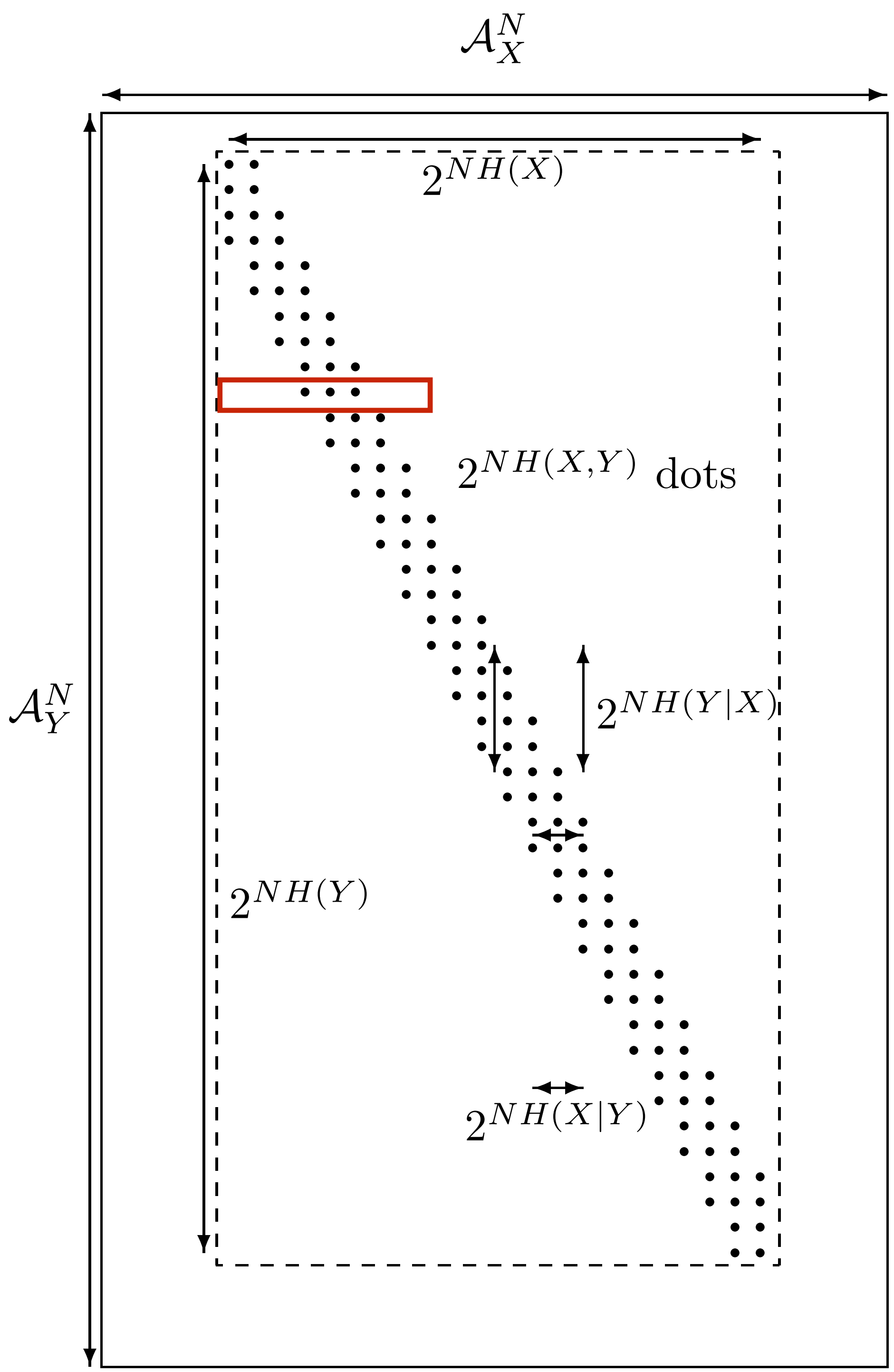
  - The probability that $\mathbf{x}, \mathbf{y}$ are jointly typical (to tolerance β) tends to 1 as $N \rightarrow \infty$
  - The number of jointly-typical sequences $|J_{N\beta}|$ is close to $2^{NH(X,Y)}$

$$\left|J_{N\beta}\right| \leq 2^{N(H(X,Y)+\beta)}$$

  - If $\mathbf{x}' \sim X^N$ and $\mathbf{y}' \sim Y^N$, i.e., $\mathbf{x}'$ and $\mathbf{y}'$ are independent samples with the same marginal distribution as $P(x, y)$, then the probability that $(\mathbf{x}', \mathbf{y}')$ lands in the jointly-typical set is about $2^{-NI(X;Y)}$.

$$P\left[(\mathbf{x}',\mathbf{y}') \in J_{N\beta}\right] \leq 2^{-N(I(X;Y)-3\beta)}$$

Joint typicality theorem



$A^N{}_X$, the set of all input strings of length N.

$A^N{}_Y$, the set of all output strings of length N.

Each dot represents a jointly-typical pair of sequences ($\mathbf{x}$, $\mathbf{y}$).

The *total* number of **independent** typical pairs is the area of the dashed rectangle is $2^{NH(X)}\,2^{NH(Y)}$

The number of **jointly**-typical pairs is roughly $2^{NH(X,Y)}$

The probability of hitting a jointly-typical pair is roughly

$$2^{NH(X,Y)}/2^{NH(X)+NH(Y)} = 2^{-NI(X;Y)}$$

FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

Proof of the noisy-channel coding theorem

Proof of the noisy-channel coding theorem - General idea

- The proof will then centre on determining the probabilities
  - (a) that **the true input codeword** is **not jointly typical** with the output sequence;
  - (b) that a **false input codeword is jointly typical with the output**.

- We will show that, for large N, **both probabilities go to zero**
  - as long as there are **fewer than 2^{NC} codewords**
  - the ensemble X **is the optimal input distribution**.

Proof of the noisy-channel coding theorem - Strategy

- We wish to **show that there exists a code** and a **decoder** having **small probability of error**.
- Evaluating the probability of error of **any particular coding and decoding system is not easy**.
- Shannon's innovation was this:
  - instead of constructing a good coding and decoding system and evaluating its error probability,
  - Shannon calculated **the average probability of block error of all codes**, **and proved that this average is small**.
  - **There must then exist individual codes that have small probability of block error**.

Random coding and typical-set decoding

- Consider an **encoding–decoding system**, whose **rate is R'**
  - We fix $P(x)$ and generate the code C that has
    - the set of codewords S with $|S| = 2^{NR'}$ of a block code $(N, K) = (N, NR')$;
    - the $2^{NR'}$ codewords are picked randomly according to the probability distribution

$$P(\mathbf{x}) = \prod_{n=1}^{N} P(x_n)$$

  - The code is know by the sender and the receiver
  - A message s is chosen from $\{1, 2, \dots, 2^{NR'}\}$, and $\mathbf{x}(s)$ is transmitted. The received signal is $\mathbf{y}$

    with
$$P(\mathbf{y} \mid \mathbf{x}^{(s)}) = \prod_{n=1}^{N} P(y_n \mid x_n^{(s)})$$

  - The signal is decoded by **typical-set decoding**
  - A decoding error occurs if $\hat{s} \neq s$

Typical-set decoding

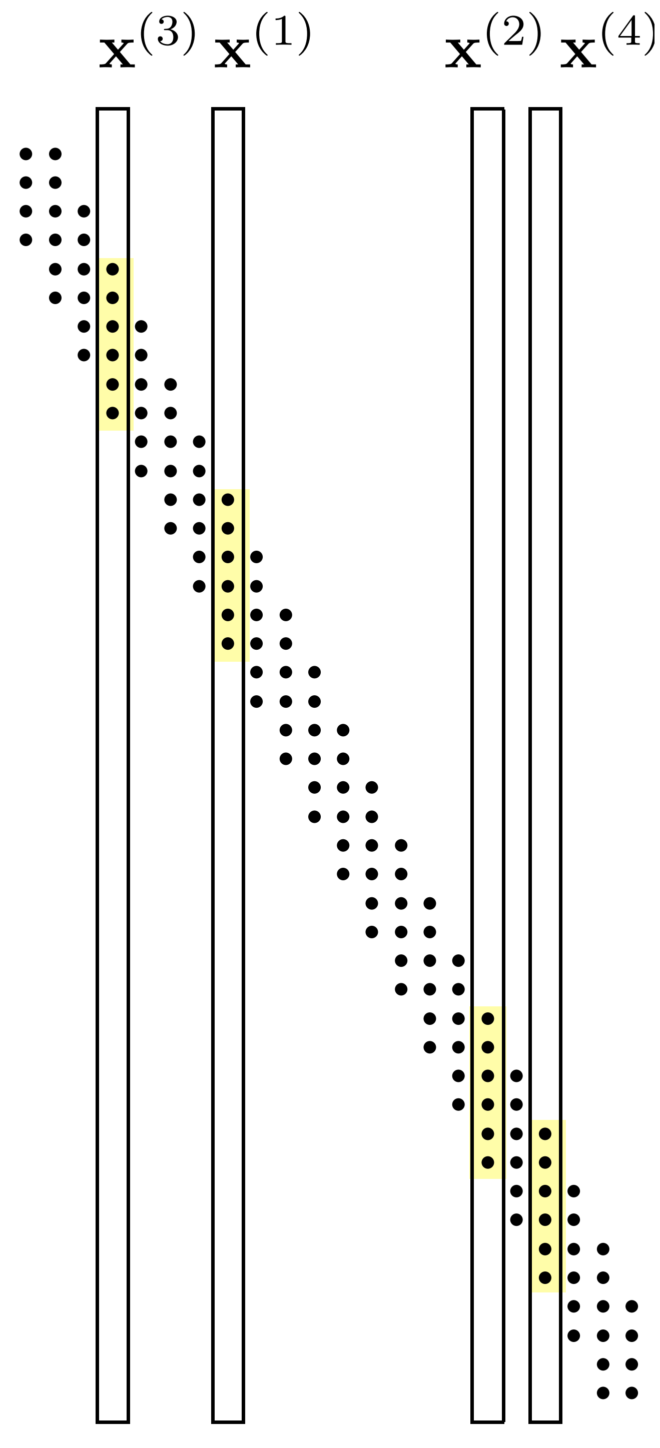
- **Typical-set decoding**
  - Decode $\mathbf{y}$ as $\hat{s}$ if $(\mathbf{x}^{(\hat{s})}, \mathbf{y})$ are jointly typical **and** there is no other s' such that $(\mathbf{x}^{(s')}, \mathbf{y})$ are jointly typical;
  - otherwise declare a failure ($\hat{s} = 0$).
- This is not the optimal decoding algorithm, but it will be good enough, and easier to analyze.
- **Joint typicality**. A pair of sequences $\mathbf{x}$, $\mathbf{y}$ of length N are defined to be jointly typical (to tolerance β) with respect to the distribution $P(x, y)$, if:

$$\mathbf{x} \text{ is typical of } P(\mathbf{x}), \quad \text{i.e.,} \quad \left| \frac{1}{N} \log \frac{1}{P(\mathbf{x})} - H(X) \right| < \beta,$$

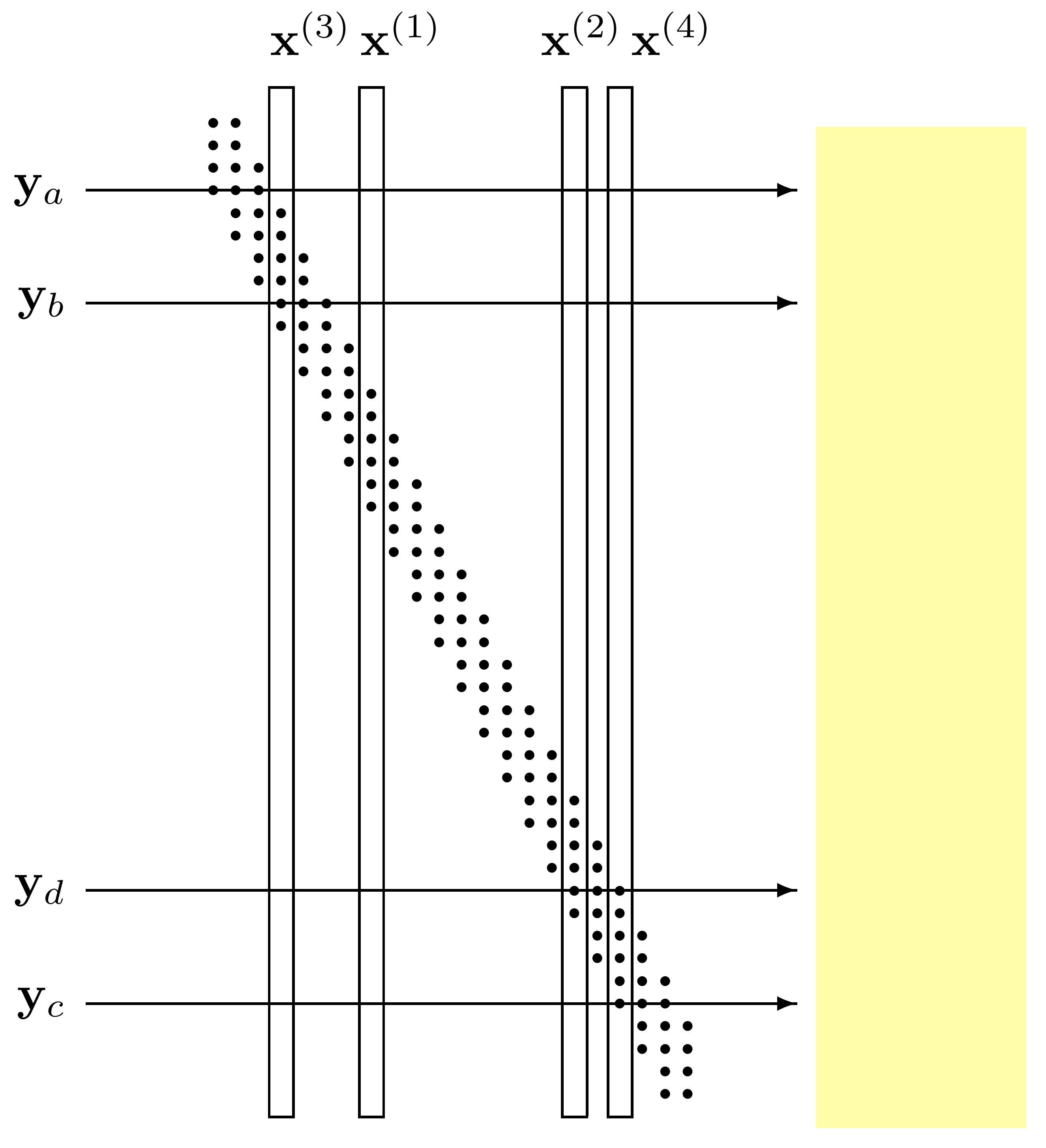
$$\mathbf{y} \text{ is typical of } P(\mathbf{y}), \quad \text{i.e.,} \quad \left| \frac{1}{N} \log \frac{1}{P(\mathbf{y})} - H(Y) \right| < \beta,$$

$$\text{and } \mathbf{x}, \mathbf{y} \text{ is typical of } P(\mathbf{x}, \mathbf{y}), \quad \text{i.e.,} \quad \left| \frac{1}{N} \log \frac{1}{P(\mathbf{x}, \mathbf{y})} - H(X, Y) \right| < \beta.$$

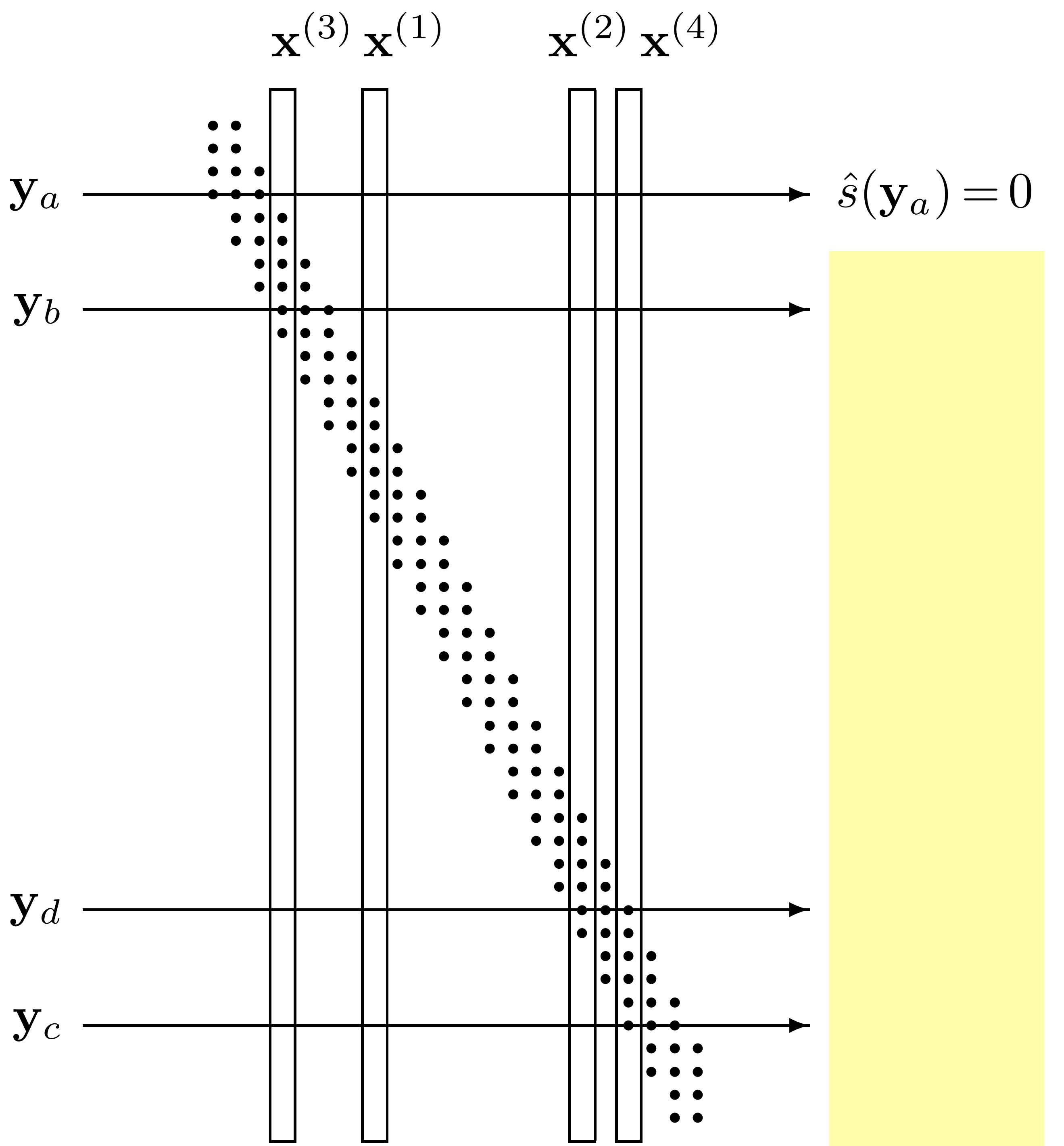
Random coding and typical-set decoding



Random coding and typical-set decoding

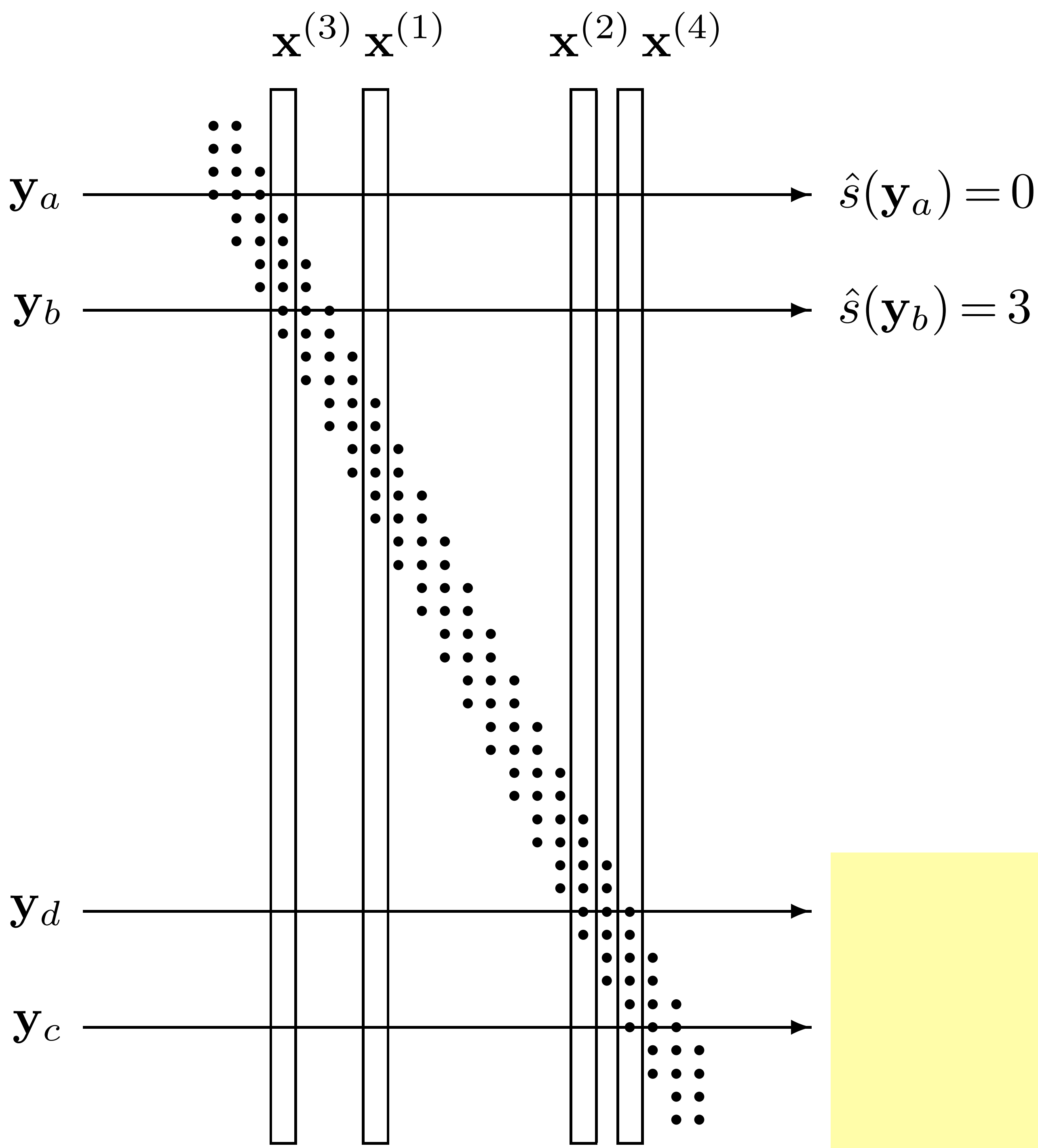


Random coding and typical-set decoding



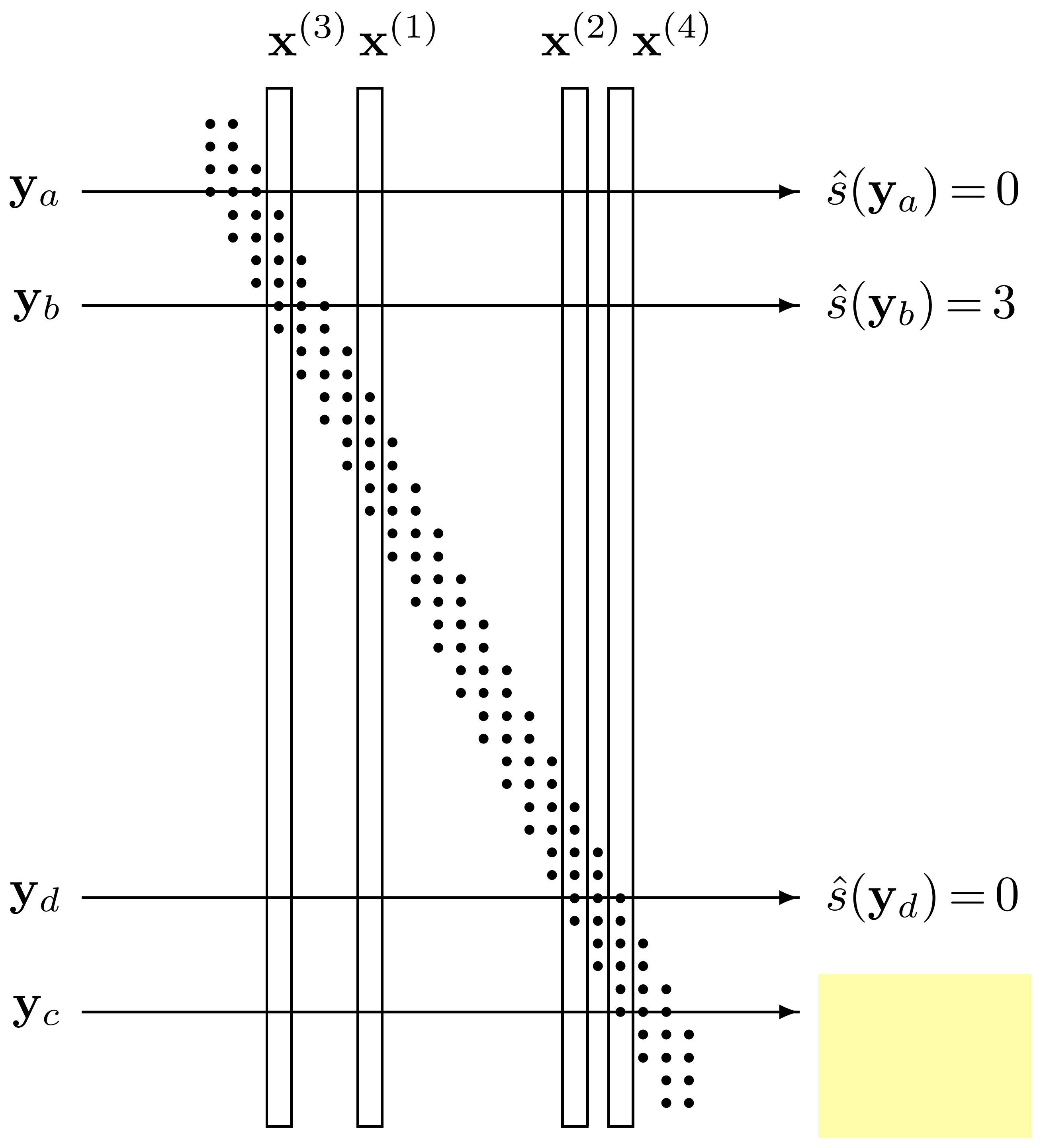
A sequence that is not jointly typical with any codeword

Random coding and typical-set decoding



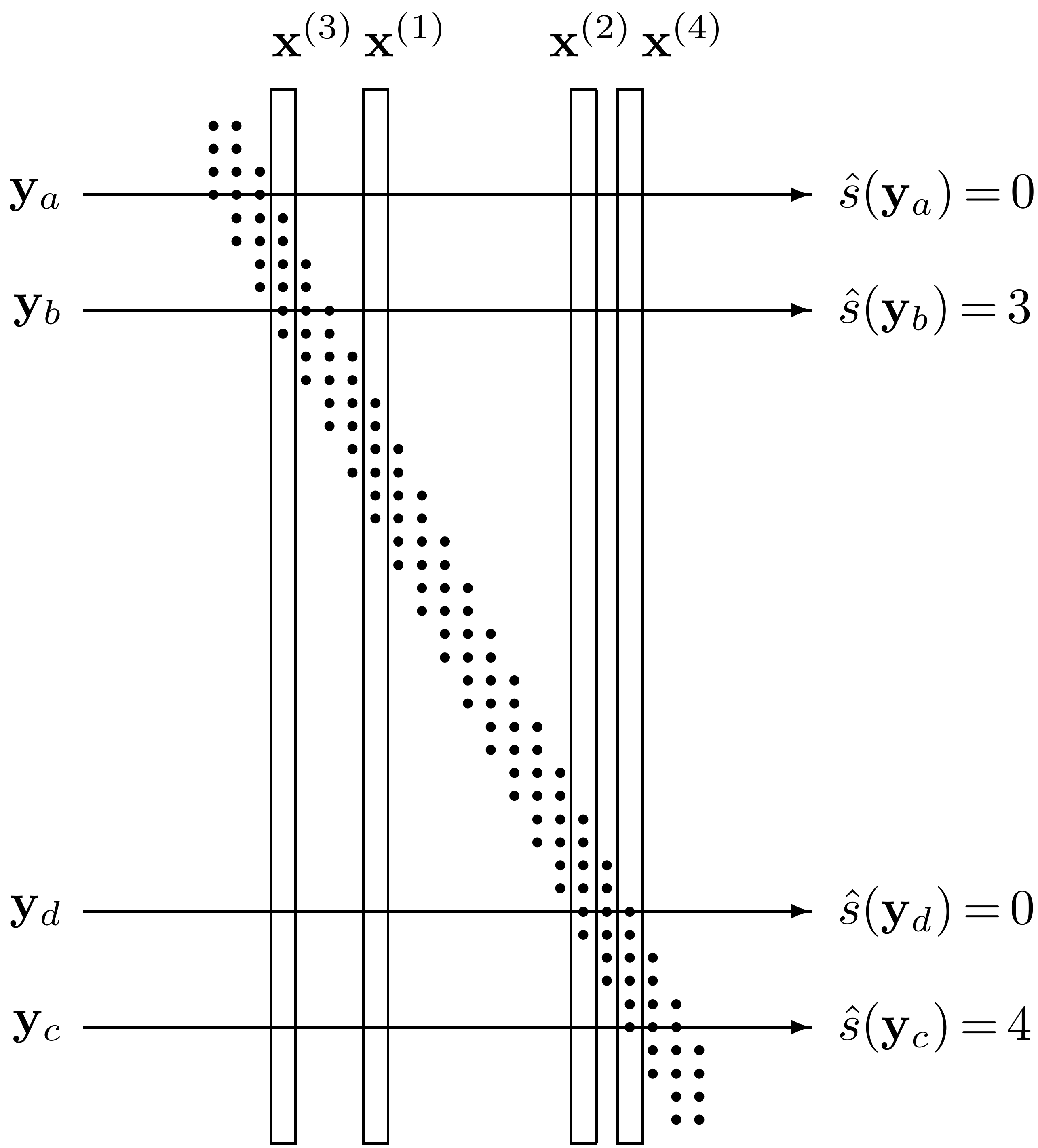
A sequence that is jointly typical with only one codeword

Random coding and typical-set decoding



A sequence that is jointly typical with more that one codeword

Random coding and typical-set decoding



A sequence that is jointly typical with only one codeword

Three Probabilities of error

- Probability of **block error** **for a particular code** C

$$p_{\mathrm{B}}(\mathcal{C}) \equiv P(\hat{s} \neq s \,|\, \mathcal{C})$$

  - This is a difficult quantity to evaluate for any given code.

- The **average** of Probability of block error over all codes of this block:

$$\langle p_{\mathrm{B}} \rangle \equiv \sum_{\mathcal{C}} P(\hat{s} \neq s \,|\, \mathcal{C}) P(\mathcal{C})$$

  - This quantity is much easier to evaluate than the first quantity $p_B(C)$.

- The **maximal** block error probability of a code C

$$p_{\mathrm{BM}}(\mathcal{C}) \equiv \max_{s} P(\hat{s} \neq s \,|\, s, \mathcal{C})$$

  - Is the quantity we are most interested in: we wish **to show that there exists a code** C **with the required rate whose maximal block error probability is small**

Three Probabilities of error - **Strategy**

- **How to show that maximal block error probability is small**?
  - By **first** finding the **average block error probability**, ⟨pB⟩.
  - Once we have **shown that this can be made smaller than a desired small number**, we immediately **deduce that there must exist at least one code *C* whose block error probability is also less than this small number.**
  - This code, whose block error probability is satisfactorily small but whose **maximal block error probability is unknown** (and could conceivably be enormous), can be **modified to make a code of slightly smaller rate whose maximal block error probability is also guaranteed to be small**.
  - We modify the code by throwing away the worst 50% of its codewords.

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Probability of error of typical-set decoder

- There are two sources of error when we use typical-set decoding.
  - The output $\mathbf{y}$ is **not jointly typical with the transmitted codeword $\mathbf{x}(s)$**
  - There is **some other codeword** in C that is jointly typical with $\mathbf{y}$.
- By the symmetry of the code construction, the average probability of error averaged over all codewords **does not depend on the selected value of s**;
  - **We can assume without loss of generality that $s = 1$.**

Joint typicality theorem (review)

- **The jointly-typical set** $J_{N\beta}$ is the set of all jointly-typical sequence pairs of length N.

- **Joint typicality theorem.** Let $\mathbf{x}$, $\mathbf{y}$ be drawn from the ensemble $(XY)^N$ defined by

$$P(\mathbf{x}, \mathbf{y}) = \prod_{n=1}^{N} P(x_n, y_n).$$

then,

  - The probability that $\mathbf{x}$, $\mathbf{y}$ are jointly typical (to tolerance β) tends to 1 as $N \to \infty$

  - The number of jointly-typical sequences $|J_{N\beta}|$ is close to $2^{NH(X,Y)}$

$$\left| J_{N\beta} \right| \leq 2^{N(H(X,Y)+\beta)}$$

  - If $\mathbf{x}' \sim X^N$ and $\mathbf{y}' \sim Y^N$, i.e., $\mathbf{x}'$ and $\mathbf{y}'$ are independent samples with the same marginal distribution as $P(x, y)$, then the probability that $(\mathbf{x}', \mathbf{y}')$ lands in the jointly-typical set is about $2^{-NI(X;Y)}$.

$$P\left[(\mathbf{x}', \mathbf{y}') \in J_{N\beta} \right] \leq 2^{-N(I(X;Y)-3\beta)}$$

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Probability of error of typical-set decoder

- There are two sources of error when we use typical-set decoding.
  - The output $\mathbf{y}$ is **not jointly typical with the transmitted codeword $\mathbf{x}(s)$**
  - There is **some other codeword** in C that is jointly typical with $\mathbf{y}$.
- By the symmetry of the code construction, the average probability of error averaged over all codewords **does not depend on the selected value of s**;
  - **We can assume without loss of generality that $s = 1$.**
- The probability that the input $\mathbf{x}^{(1)}$ and the output $\mathbf{y}$ are not jointly typical vanishes, by the joint typicality theorem's first part.
  - Let δ, be the upper bound on this probability, satisfying $\delta \to 0$ as $N \to \infty$;
  - For any desired δ, we can find a block length $N(\delta)$ such that the $P((\mathbf{x}^{(1)}, \mathbf{y}) \notin J_{N\beta}) \leq \delta$.
- The probability that $\mathbf{x}^{(s')}$ and $\mathbf{y}$ are jointly typical, for a given $s' \neq 1$ is $\leq 2^{-N(I(X;Y)-3\beta)}$, by the joint typicality theorem's first part 3. And there are $(2^{NR'} - 1)$ rival values of s' to worry about.

Probability of error of typical-set decoder

- The probability that the input $\mathbf{x}^{(1)}$ and the output $\mathbf{y}$ are not jointly typical vanishes, by the joint typicality theorem's first part.
  - Let δ, be the upper bound on this probability, satisfying $\delta \to 0$ as $N \to \infty$;
  - For any desired δ, we can find a block length $N(\delta)$ such that the $P((\mathbf{x}^{(1)}, \mathbf{y}) \notin J_{N\beta}) \leq \delta$.
- The probability that $\mathbf{x}^{(s')}$ and $\mathbf{y}$ are jointly typical, for a given $s' \neq 1$ is $\leq 2^{-N(I(X;Y)-3\beta)}$, by the joint typicality theorem's part 3. And there are $(2^{NR'} - 1)$ rival values of s' to worry about.
- Thus the average probability of error $\langle p_B \rangle$ satisfies

$$\begin{aligned} \langle p_{\mathrm{B}} \rangle \quad &\leq \quad \delta + \sum_{s'=2}^{2^{NR'}} 2^{-N(I(X;Y)-3\beta)} \\ &\leq \quad \delta + 2^{-N(I(X;Y)-R'-3\beta)}. \end{aligned}$$

- $<p_B>$ can be made $< 2\delta$ by increasing N if $R' < I(X; Y) - 3\beta$.

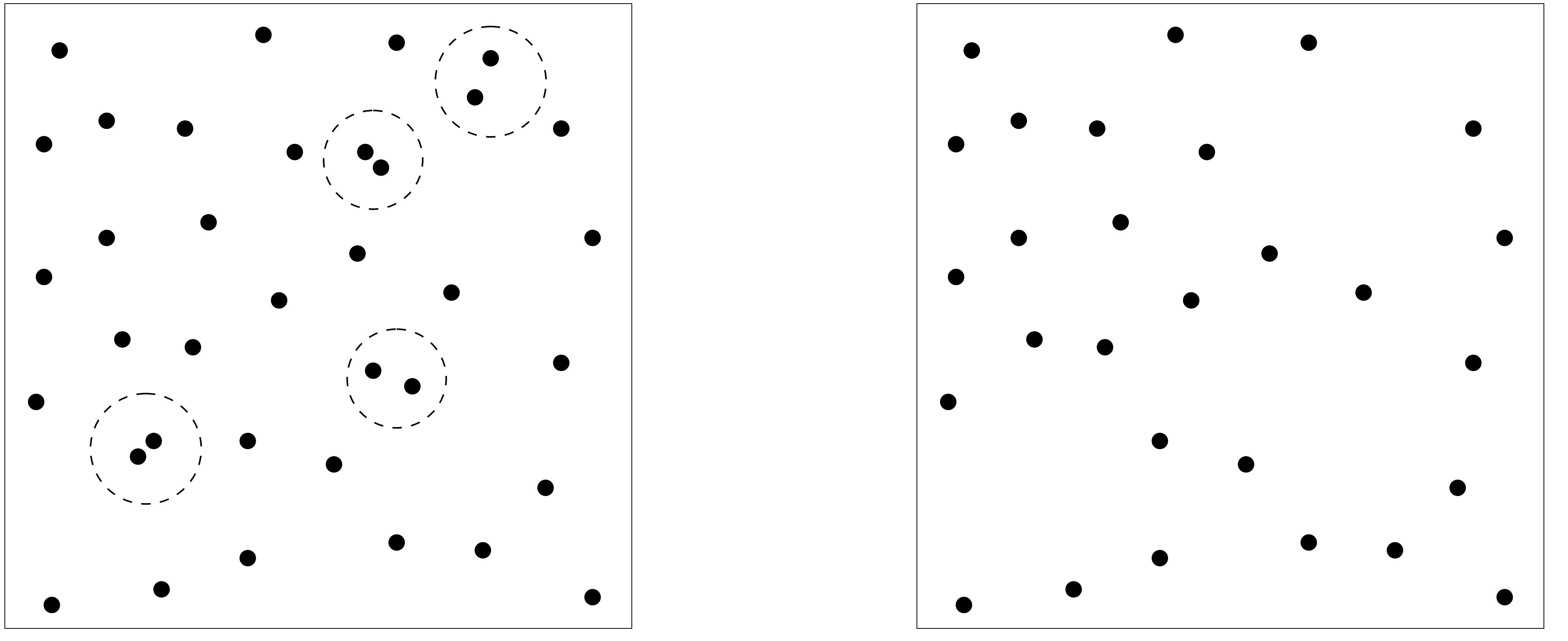
Probability of error of typical-set decoder

- Thus the average probability of error $\langle p_B \rangle$ satisfies

$$\begin{aligned}\langle p_{\mathrm{B}} \rangle &\leq \delta + \sum_{s'=2}^{2^{NR'}} 2^{-N(I(X;Y)-3\beta)} \\ &\leq \delta + 2^{-N(I(X;Y)-R'-3\beta)}.\end{aligned}$$

- $<p_B>$ can be made $< 2\delta$ by increasing N if $R' < I(X; Y) - 3\beta$.
- We choose $P(x)$ in the proof to be the optimal input distribution of the channel.
  - Then the condition $R' < I(X; Y) - 3\beta$ becomes $R' < C - 3\beta$
- Since the **average probability of error over all codes is $< 2\delta$**, there **must exist a code with mean probability of block error $p_B(C) < 2\delta$**.
- To show that **not only the average** but **also the maximal probability of error**, p_{BM}, can be made small we modify this code by **throwing away the worst half of the codewords** – the ones most likely to produce errors.

Probability of error of typical-set decoder

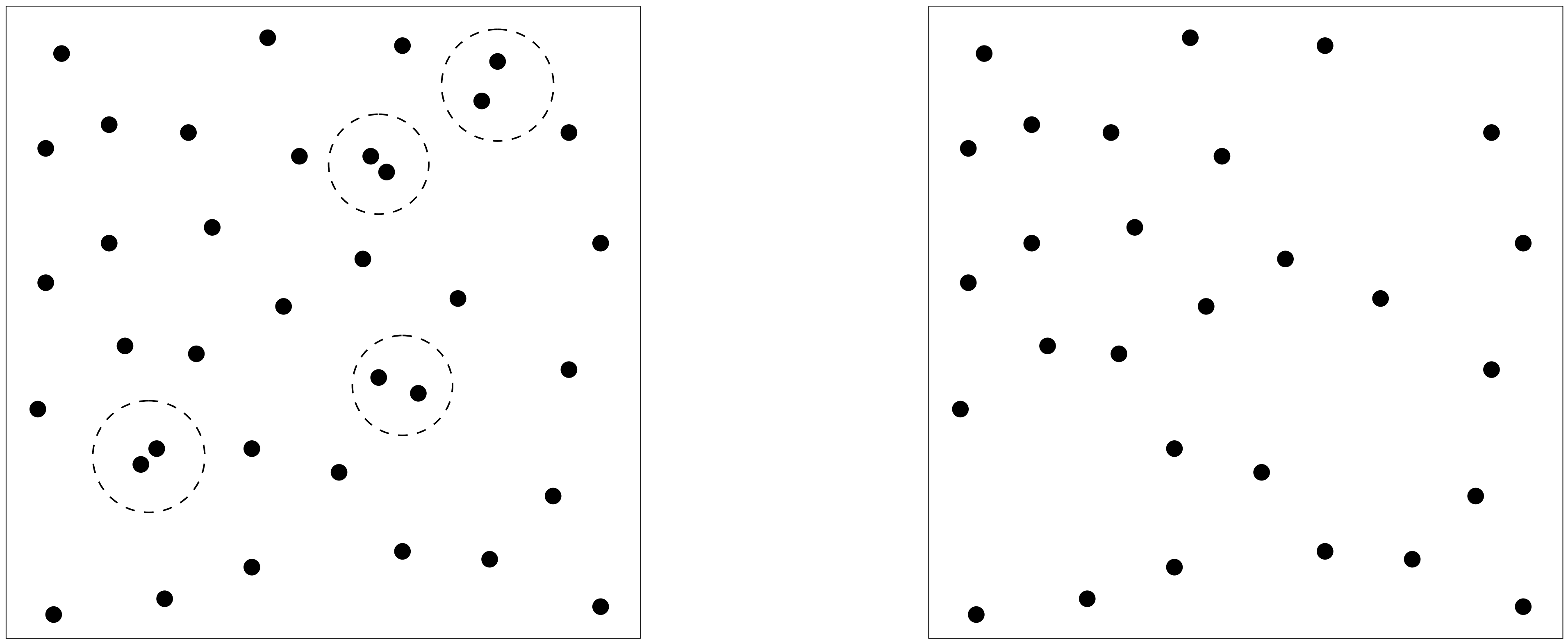


⇒

(a) A random code . . . (b) expurgated

- Those that remain must all have *conditional* probability of error less than 4δ.
- These **remaining codewords to define a new code**. This new code has $2^{NR'-1}$ codewords.
- The rate is reduced from R’ to R’ - $1/N$ and achieved $p_{BM} < 4\delta$.

Probability of error of typical-set decoder



(a) A random code . . . $\Rightarrow$ (b) expurgated

- Those that remain must all have *conditional* probability of error less than 4δ.
- These **remaining codewords to define a new code**. This new code has $2^{NR'-1}$ codewords.
- The rate is reduced from R' to R' - $1/N$ and achieved $p_{BM} < 4\delta$.

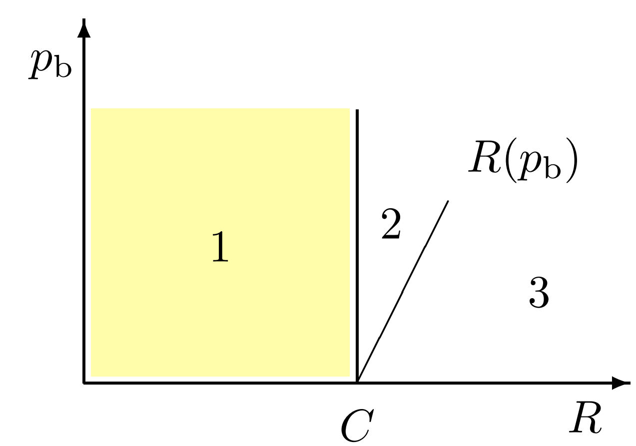
The theorem

Part 1 - Positive

- For every discrete memoryless channel, the channel capacity

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

has the following property. For any $\varepsilon > 0$ and $R < C$, for large enough N, there exists a code of length N and rate $\geq R$ and a decoding algorithm, such that the maximal probability of block error is $< \varepsilon$.



- We can '*construct*' a code of rate $R' - 1/N$,

where $R' < C - 3\beta$, with maximal probability of error $< 4\delta$.

We obtain the theorem as stated by setting $R' = (R + C)/2$,

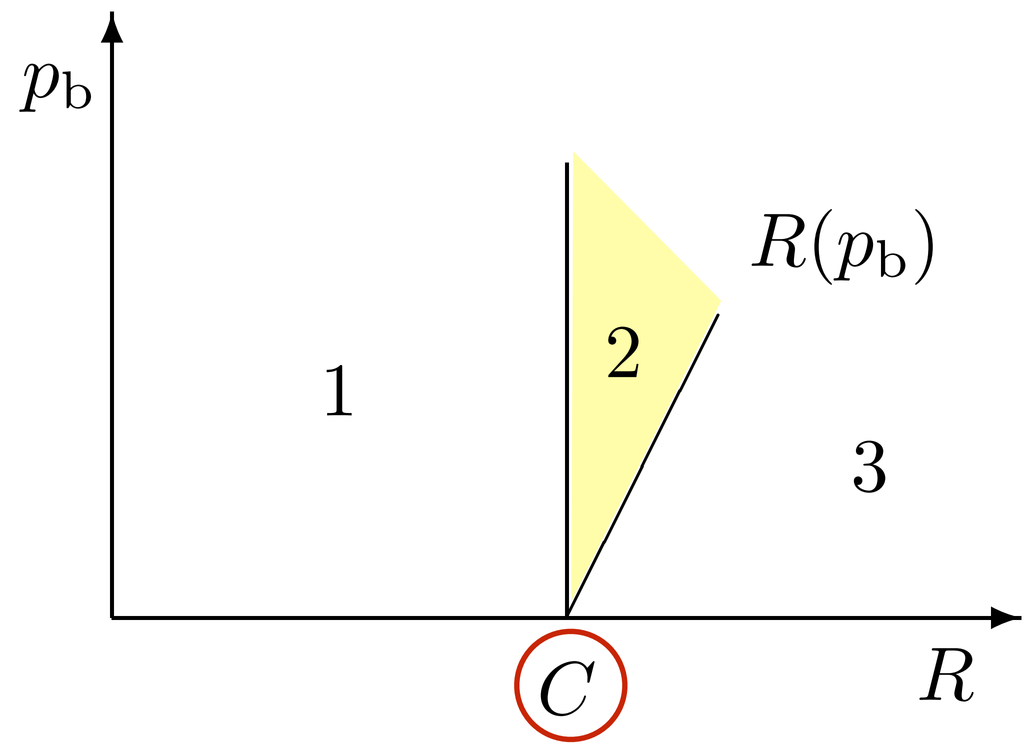
$\delta = \varepsilon/4$, $\beta < (C - R')/3$, and N sufficiently large for the remaining conditions to hold.

Communication (with errors) above capacity

Communication (with errors) above capacity

- We have shown that **we can turn any noisy channel into an essentially noiseless binary** channel with **rate up to *C*** bits per cycle.
- We **now extend the right-hand boundary of the region of achievability** at **non-zero error probabilities**
- If a probability of bit error p_b is acceptable, rates up to $R(p_b)$ are achievable, where

$$R(p_b) = \frac{1}{1 - H_2(p_b)}$$



Communication (with errors) above capacity

- We know we can make the noisy channel into a perfect channel with a smaller rate !
- It is sufficient to consider communication with errors over a **noiseless channel**.
- How fast can we communicate over a noiseless channel, if we are allowed to make errors ?
- Consider a **noiseless binary channel**
  - Assume that we force communication at a rate greater than its capacity of 1 bit.
  - For example, if we require the sender to attempt to communicate at ***R* = 2 bits per cycle** then he must effectively **throw away half of the information**.
  - One simple strategy is to communicate a fraction $1/R$ of the source bits, and ignore the rest. The receiver guesses the missing fraction $1 - 1/R$ at random,

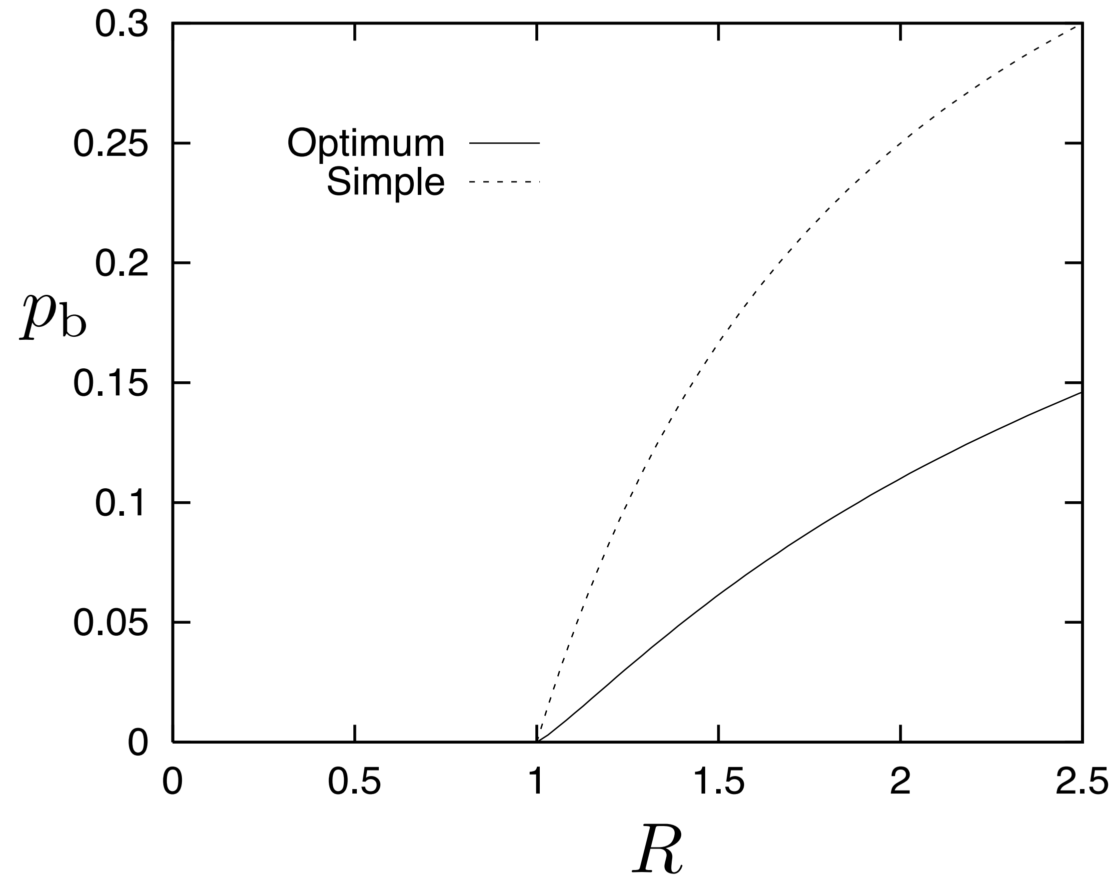
$$p_{\mathrm{b}} = \frac{1}{2}(1 - 1/R)$$

FACULDADE DE CIÊNCIAS E TECNOLOGIA UNIVERSIDADE NOVA DE LISBOA

Communication (with errors) above capacity

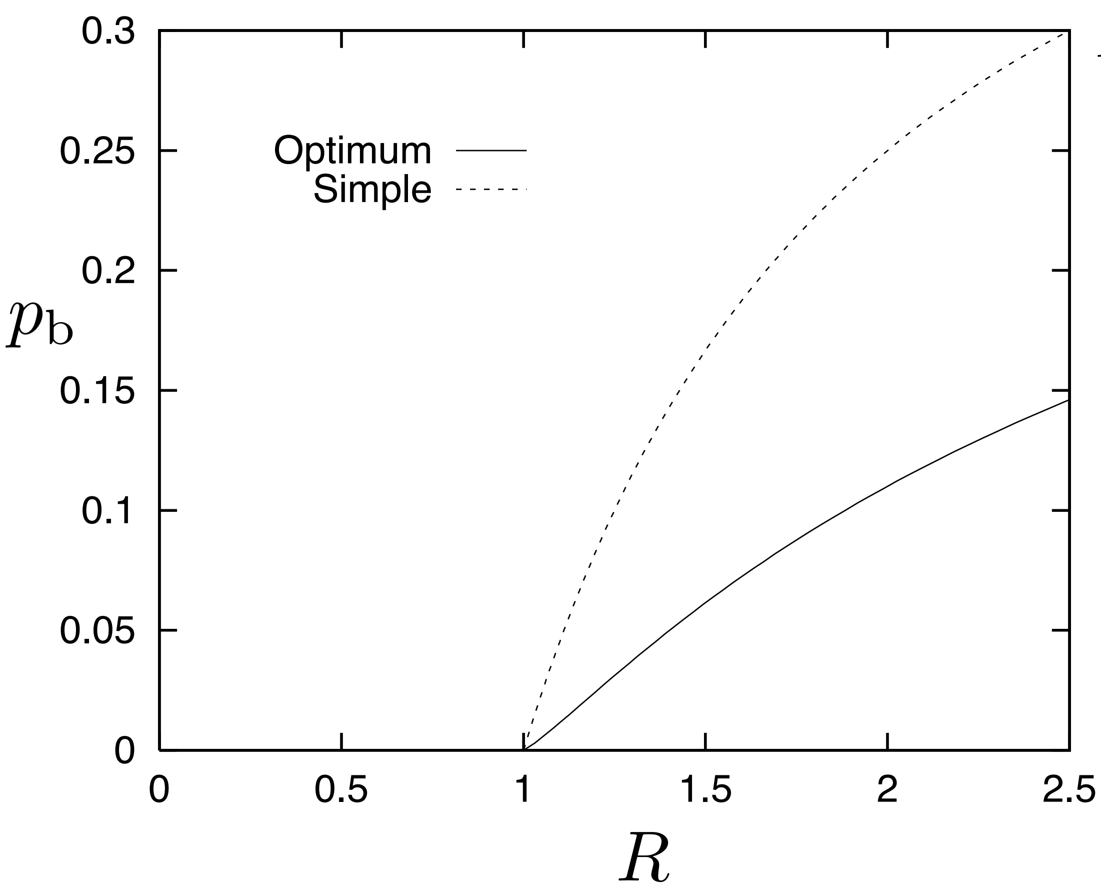
- Consider a **noiseless binary channel**
  - One simple strategy is to communicate a fraction $1/R$ of the source bits, and ignore the rest. The receiver guesses the missing fraction $1 - 1/R$ at random,

$$p_{\mathrm{b}} = \frac{1}{2}(1 - 1/R)$$



Communication (with errors) above capacity

- Consider a **noiseless binary channel**
  - One simple strategy is to communicate a fraction $1/R$ of the source bits, and ignore the rest. The receiver guesses the missing fraction $1 - 1/R$ at random,



$$p_{\mathrm{b}} = \frac{1}{2}(1 - 1/R)$$

$$R(p_b) = \frac{1}{1 - H_2(p_b)}$$

Can be improved up to

Communication (with errors) above capacity

- For any Channel, we **can extend the right-hand boundary of the region of achievability** at **non-zero error probabilities**
  - If a probability of bit error p_b is acceptable, rates up to $R(p_b)$ are achievable, where

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

Computing capacity

Computing capacity

- How can we compute the capacity of a given discrete memoryless channel?
  - We need to find its optimal input distribution.
  - In general we can find the optimal input distribution by a computer search, making use of the **derivative of the mutual information with respect to the input probabilities**.
- Since $I(X;Y)$ is concave $\frown$ in the input distribution p, any probability distribution p at which $I(X;Y)$ is stationary must be a global maximum of $I(X;Y)$.
- So it is **tempting** to put the derivative of $I(X;Y)$ into a routine that finds a local maximum of $I(X;Y)$, that is, an input distribution $P(x)$ such that

$$\frac{\partial I(X;Y)}{\partial p_i} = \lambda \ \text{ for all } i,$$

where λ is a Lagrange multiplier associated with the constraint $\sum_i p_i = 1$

Computing capacity

- Since $I(X;Y)$ is concave $\frown$ in the input distribution p, any probability distribution p at which $I(X;Y)$ is stationary must be a global maximum of $I(X;Y)$.
- So it is **tempting** to put the derivative of $I(X;Y)$ into a routine that finds a local maximum of $I(X;Y)$, that is, an input distribution $P(x)$ such that

$$\frac{\partial I(X;Y)}{\partial p_i} = \lambda \quad \text{for all } i,$$

where λ is a Lagrange multiplier associated with the constraint $\sum_i p_i = 1$

- However, this approach may fail to find the right answer, because $I(X;Y)$ might be maximized by a distribution that has $p_i = 0$ for some inputs.
- The optimization routine must therefore take account of the possibility that, as we go up hill on $I(X;Y)$, we may run into the inequality constraints $p_i \geq 0$.

Computing capacity - Results that may help

- All outputs must be used
- $I(X\,;\,Y)$ is a convex $\smile$ function of the channel parameters.
- There may be several optimal input distributions, but they all look the same at the output
- **A discrete memoryless channel is a symmetric channel** if the set of outputs can be partitioned into subsets in such a way that for each subset the matrix of transition probabilities has the property that each row (if more than 1) is a permutation of each other row and each column is a permutation of each other column.

Computing capacity - symmetric channel

- An example of a Symmetric Channel

$$
\begin{array}{rclrcl}
P(y=0 \mid x=0) & = & 0.7\,; & P(y=0 \mid x=1) & = & 0.1\,; \\
P(y=\texttt{?} \mid x=0) & = & 0.2\,; & P(y=\texttt{?} \mid x=1) & = & 0.2\,; \\
P(y=1 \mid x=0) & = & 0.1\,; & P(y=1 \mid x=1) & = & 0.7.
\end{array}
\tag{10.23}
$$

is a symmetric channel because its outputs can be partitioned into $(0,1)$ and `?`, so that the matrix can be rewritten:

$$
\begin{array}{rclrcl}
\hline
P(y=0 \mid x=0) & = & 0.7\,; & P(y=0 \mid x=1) & = & 0.1\,; \\
P(y=1 \mid x=0) & = & 0.1\,; & P(y=1 \mid x=1) & = & 0.7\,; \\
\hline
P(y=\texttt{?} \mid x=0) & = & 0.2\,; & P(y=\texttt{?} \mid x=1) & = & 0.2. \\
\hline
\end{array}
\tag{10.24}
$$

Other coding theorems

Other coding theorems

- The noisy-channel coding theorem is quite general, applying to any discrete memoryless channel; but it is not very specific.
- The theorem **only says** that reliable communication with error probability ε and rate R can be achieved by using codes with sufficiently large block length N.
- The theorem **does not say how large N needs to be** to achieve given values of R and ε.
- Presumably, the **smaller ε is and the closer R is to C**, the **larger N has to be**

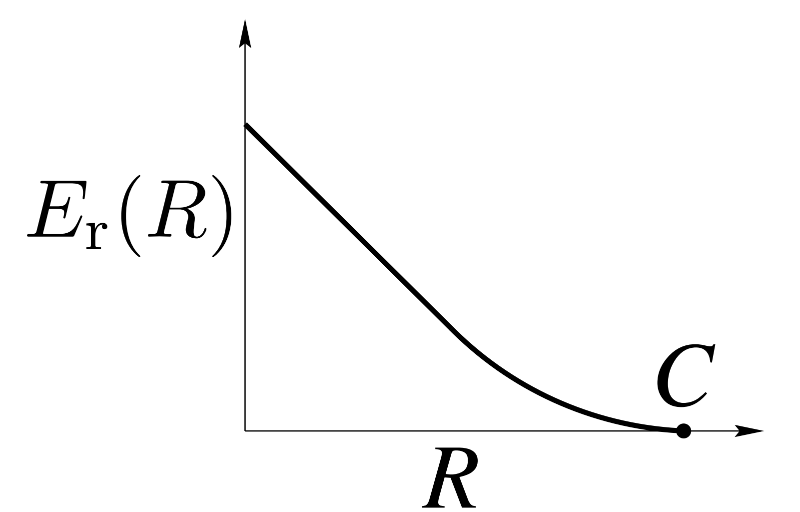
Noisy-channel coding theorem – explicit N-dependence

- For a discrete memoryless channel, a block length N and a rate R, there exist block codes of length N **whose average probability of error** satisfies:

$$p_{\mathrm{B}} \leq \exp\left[-N E_{\mathrm{r}}(R)\right]$$

  where $E_r(R)$ is the *random-coding exponent* of the channel, a convex $\smile$, decreasing, positive function of R for $0 \leq R < C$.

- The random-coding exponent is also known as the **reliability function**
- $E_r(R)$ approaches zero as $R \to C$;
- The computation of the random-coding exponent for interesting channels is a challenging task



Lower bounds on the error probability as a function of N

- For any code with block length N on a discrete memoryless channel, the probability of error assuming **all source messages are used with equal probability** satisfies

$$p_{\mathrm{B}} \gtrsim \exp[-N E_{\mathrm{sp}}(R)],$$

- where the function $E_{sp}(R)$, the *sphere-packing* exponent of the channel, is a convex $\smile$, decreasing, positive function of R for $0 \leq R < C$.

Further Reading and Summary

Q&A

Further Reading

- **Recommend Readings**
  - Information Theory, Inference, and Learning Algorithms from David MacKay, 2015, pages 161 - 173.
- **Supplemental readings:**

What you should know

- The three parts of noisy-channel coding theorem
- The concept of Jointly-typical sequences
- What is Random coding and typical-set decoding
- What is the general ideia of noisy-channel coding theorem’s demonstration

Further Reading and Summary

Q&A