

# Teaching freshers information theory: Shannon's theorem for linear codes on the BEC

**Conference Paper**

**Author(s):**

Sayir, Jossy

**Publication date:**

2026-02-25

**Permanent link:**

<https://doi.org/https://doi.org/10.3929/ethz-c-000796746>

**Rights / license:**

[In Copyright - Non-Commercial Use Permitted](#)

# Teaching freshers information theory: Shannon’s theorem for linear codes on the BEC

Josy Sayir

University of Cambridge

Probability, Systems, Information & Inference Lab ( $\Psi^2$ )

Department of Engineering

email: js851@cam.ac.uk

**Abstract**—In this paper on information theory education, an outline of a new course for first year students at the University of Cambridge is described. The coding theorem is derived for the special case of the Binary Erasure Channel using random binary linear block coding.

## I. INTRODUCTION

I am designing a new undergraduate course on information theory for first year engineers. As part of this course, I am proposing to derive the channel coding theorem for the Binary Erasure Channel (BEC) using random linear codes, for which asymptotically vanishing error probability can be established using a fairly simple counting argument. This paper will discuss the motivation for this course and the approach taken to teach this part of the course.

My aim in presenting this paper on information theory education at the International Zurich Seminar (IZS) is to seek advice and feedback on this teaching project. While reading this paper or attending my presentation, please reflect on how you and your students learned these concepts. What do you think is a hard concept to digest? Are there shortcuts, illustrations, metaphors, or examples that help clarify these concepts? If you have any tips, I would love to hear from you.

## II. COURSE BACKGROUND

As part of our undergraduate reform of the Engineering Tripos at the University of Cambridge, we are introducing an 8 lecture block on Information Theory to be taught in the third trimester of the first year. This lecture block will be accompanied by two Examples Papers taught by supervisors across 29 Cambridge colleges. The reformed tripos is scheduled to be deployed starting in October 2027, and we are currently in the process of developing lecture materials, with the aim of finalising them over the coming months so supervisors can be trained to help teach the material.

Like most information theory courses, this course will cover source coding and point-to-point channel coding. However, the principal aim is to awaken students’ interest in the area rather than to teach them information theory comprehensively. An auxiliary aim is to motivate them to learn probability theory and linear algebra by showing them applications of these mathematical tools early on. As such, it is essential that the material remain accessible mainly with pre-university level mathematics, and that students have a positive first encounter

with probability and matrices. We currently have information theory courses in the 3rd and 4th year of the tripos and, while some modifications of the 3rd year course are expected, the plan is to maintain these courses, allowing students to gain a more comprehensive understanding later on.

The Cambridge Engineering tripos is a general-entry course, so that some of the students taking this information theory course will end up specialising in mechanical engineering, structural engineering, electrical engineering, while some (currently about half) will specialise in information engineering and hence remain in related areas. The idea of teaching information theory so early to such a broad range of future engineers prompted some debate, but ultimately the field provides a good example of how engineering intuition and rigorous analysis lead to groundbreaking engineering innovations, and knowing this would benefit all engineers.

The source coding part of the course follows a classical design and will not be discussed in detail. We reserve the concepts of typicality and the asymptotic equipartition property for our advanced course, though many consider this the cornerstone of the field. These concepts are hard to grasp without probability background. Variable length coding and prefix-free codes are easier to understand with no prior background.

This paper focuses on the channel coding theorem. Standard teaching of this theorem relies on typical sequences or on advanced bounding techniques for random block codes. We chose to teach the special case of linear coding for the Binary Erasure Channel (BEC), because this simple example suffices to develop full intuition as to why random coding works and why the error probability decays with increasing block length.

I developed the background for this course while working at the “Forschungszentrum Telekommunikation Wien” (Vienna Telecommunications Research Centre, ftw.) in the early 2000s. This paper will be presented at IZS within a session entitled “From Theory to Wireless (FTW)” that brings together alumni of this former research centre. At the time, my colleague Ralf Müller (now a professor at the Friedrich-Alexander-Universität Erlangen-Nürnberg) was working on the application of random matrix theory to Multiple-Input Multiple-Output (MIMO) wireless communications, and I was teaching a university course on Turbo Coding and Low-Density Parity-Check (LDPC) codes together with Gottfried Lechner (now a professor at the University of South Australia / Adelaide

University). Developments in the understanding of LDPC codes relied heavily on concepts developed for the BEC, and Ralf and I got talking about how one could prove the coding theorem for the BEC using random matrices. I remember laughing about Ralf’s surprise when we discovered that, unlike for complex random matrices, a large random square matrix over  $\text{GF}(2)$  is regular with probability 0.2887881, not 0 or 1. More will be said about this in the next sections.

### III. CHANNEL, CAPACITY AND THE CODING THEOREM

Peter Elias introduced the Binary Erasure Channel (BEC) in [1] and [2]. Information theorists all know this channel as a sort of thought experiment akin to those used to illustrate Einstein’s relativity, rather than a model for a real-world channel. Careful justification will be required to maintain first year students’ attention when presented with a channel with no apparent applications. In particular, I count on discussing channels more generally. When proposing to restrict our attention to BECs, I will note that all of our analysis also applies to non-binary erasure channels for which there are real world applications.

To discuss channel capacity, several options are possible. Having introduced entropy in the first half of the course, we could show students the expression  $C = \max_{p_x}[H(Y) - H(Y|X)]$ . However, this expression may prove difficult for two reasons: first, it relies on conditional entropies and although we will need conditional probabilities later in this paper to discuss probability of error, this will be high level and intuitive without fully introducing properties of conditional and joint probability. Second, maximising an expression over a range of probability distributions is quite an advanced concept that requires students to be comfortable with constrained multivariate function optimisation. More likely, we will defer this expression to the advanced course and use a different argument to show the capacity of the BEC specifically, namely that if the transmitter knows the position of the erasures (essentially a BEC with noiseless feedback but we probably wouldn’t use that terminology yet) then they can achieve a rate of  $R = 1 - p$  by transmitting uncoded data in the non-erased positions, where  $p$  is the probability of erasure. This approach also has the benefit of not requiring an explicit converse, since any data rate in excess of  $1 - p$  would have to use the erased positions, but erasures carry no information and hence that portion of the transmitted data would be “lost” or, equivalently, recoverable with an error probability of  $1/2$ .

At this stage, block coding can be introduced as a sort of “magic trick” promise: with this technique, we will be able to transmit data at a rate almost equal to  $1 - p$  even if the position of the erasures are not known to the transmitter. Analogies with language will be used to explain the concept and usefulness of redundancy in coding, where block encoding is akin to translating a perfectly efficient language onto a language that can help recover information for all likely patterns of erasures.

### IV. EQUATIONS, MATRICES, LINEAR SPACES AND RANK

In late 2015, I tested a version of this BEC-based course on a class of 3rd year students in Cambridge. At the time, my late

colleague David MacKay had been diagnosed with a terminal illness and was unable to teach the courses he had intended to teach. In order to keep involved in teaching, he proposed to serve as my “teaching buddy”, attending my lectures and providing feedback to improve my courses. His feedback proved invaluable. In particular my 4th year course on Reed Solomon coding has been drastically improved, resulting in a spectacular improvement in the proficiency of students as verified by exam performance. However, his reaction to the BEC-based course shocked me. Instead of praising it as I had hoped, he reported not understanding much, explaining he had never fully understood the concept of *matrix rank* and hence considered this to be an “advanced” linear algebra concept that could not be taken for granted. Reflecting on David’s famous information theory course, which I had audited a few years earlier, I realised that I had been in a similar position in that I found his “easy” analysis of the Binary Symmetric Channel (BSC) hard to follow because it relied on combinatorial counting arguments that he considered trivial, but that never felt comfortable to me. The reason why I feel comfortable with matrix rank as a concept is that I learned linear algebra in the first half of my first year of university, taught by the legendary Professor Konrad Voss at ETH Zurich, whereas counting and combinatorics is something I misunderstood in secondary school and never quite caught up with during my university education. In contrast, in Cambridge, linear algebra is currently taught as part of a second year module and students consider it an “advanced” part of mathematics (as did David). This goes to show that the difference between “easy” and “advanced” often depends on the stage of one’s education when the material is taught.

In our following discussion with David, he proposed that I introduce the concepts of matrix manipulation, rank, and inversion based on simple examples using small binary matrices, relating them to systems of equations that all students have learned to solve during their secondary education and hence typically feel comfortable with. My aim in this part of the course is to put David’s proposal into practice and develop the following concepts using examples on small matrices:

- matrix representation of systems of equation
- sets of solutions, linear spaces, subspaces and dimension
- column and row rank, rank of a matrix
- solving linear equations using Gauss elimination steps (at this point I may introduce an example in  $\text{GF}(3)$  instead of  $\text{GF}(2)$ , to include the “division by pivot” step that doesn’t exist in binary.)

Having familiarised students with matrix manipulation, I will introduce the concept of a linear code, encoder, and decoding over the BEC by solving linear equations: if a codeword  $\mathbf{c} = \mathbf{u}\mathbf{G}$  is transmitted over a BEC, where  $\mathbf{u}$  is a length  $K$  information word and  $\mathbf{G}$  is a binary  $K \times N$  encoding/generator matrix, and the received word  $\mathbf{r}$  has erasures at positions  $j_1, \dots, j_\ell$ , then the encoded information can be recovered if the system of equations

$$\tilde{\mathbf{r}} = \mathbf{u}\tilde{\mathbf{G}} \quad (1)$$

has a unique solution  $\mathbf{u}$ , where  $\tilde{\mathbf{r}}$  is the received word with the erasures removed and  $\tilde{\mathbf{G}}$  is the encoding matrix with columns  $j_1, \dots, j_\ell$  removed.

V. RANDOM CODING AND DECODING

The random coding experiment to bound the probability of decoding of failure is as follows:

- 1) pick a  $K \times N$  random binary matrix  $\mathbf{G}$  with entries chosen independently and uniformly at random, and  $K = \lfloor RN \rfloor$  where  $R$  is the code rate (where the effect of integer rounding can be neglected in our analysis for large  $N$ )
- 2) the channel picks  $L$  positions to erase  $j_1, \dots, j_L$  where  $L$  is a binomially distributed random variable with mean  $pN$  and variance  $p(1-p)N$
- 3) if the resulting  $K \times (N-L)$  matrix  $\tilde{\mathbf{G}}$  has rank  $K$ , then there is a unique solution to (1) and the information word can be recovered

The choice of an initial random matrix  $\mathbf{G}$  followed by the channel choice of  $L$  erasure positions is equivalent to directly choosing an  $K \times (N-L)$  binary matrix  $\tilde{\mathbf{G}}$  with entries picked independently and uniformly at random, so the probability of success is the probability that a  $K \times (N-L)$  matrix chosen in this manner has rank  $K$ .

The probabilistic analysis of this experiment is in two parts: the “nice” part is to determine the probability of success for a given  $L$ , while the “boring” part is taking the average over  $P_L(\ell)$ . I call the second part boring because I have not found a pedagogically effective way of teaching this to first years. Several options are possible:

- 1) explain why, when  $R = 1 - p - \varepsilon$  for a small  $\varepsilon$ , the probability that the aspect ratio deviation  $D = (N-L) - K$  exceed a threshold  $\delta N$  for  $\delta < \varepsilon$ , i.e., the probability that  $L < N(p + \varepsilon - \delta)$ , goes to 1 as  $N$  goes to infinity. This can be illustrated graphically by showing that the binomial distribution becomes thinner with respect to a proportional deviation  $N(\varepsilon - \delta)$  from the mean.
- 2) equivalently, one can show that the deviation  $N(\varepsilon - \delta)$  can be made to be any multiple of the standard deviation  $\sqrt{p(1-p)N}$ .
- 3) a more formal standard derivation uses Chebyshev’s inequality (to obtain a bound on the probability of failure proportional to  $1/N$ ) or a Chernoff-style bound (to obtain an exponential bound). The difficulty here is that the concept of general tail bounds on distributions is a little hard to digest before one has had time to learn probability theory more formally.
- 4) using specific bounds on the tail of a binomial would be less abstract and hence easier to understand, but these bounds tend to be quite technical and hard to work with.

I will probably use the first and second approaches during lectures, and possibly give an optional exercise guiding students through tail bounding methods for those who want to explore this topic in more depth.

The “nice” part of the probability analysis, for a given  $L$ , goes as follows. Let  $M = N - L$  be the number of columns

in the matrix  $\tilde{\mathbf{G}}$  corresponding to non-erased symbols in the received word  $\mathbf{r}$ . The question is now, given that the matrix was constructed by selecting its entries independently and uniformly at random, to determine the probability that this matrix has full row rank  $K$ . Clearly, if  $M < K$ , this probability is 0. Assuming  $M \geq K$ , we can now visualise the matrix as follows

$$\tilde{\mathbf{G}} = \begin{matrix} & \overbrace{\begin{matrix} \square & \square & \square & \dots & \square \\ \square & \square & \square & \dots & \square \\ \square & \square & \square & \dots & \square \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \square & \square & \square & \dots & \square \end{matrix}}^M & \begin{matrix} 1 \times M \\ 2 \times M \\ 3 \times M \\ \vdots \\ K \times M \end{matrix} \end{matrix}$$

and note that

- there are  $2^M - 1$  possible choices for the first row (all binary vectors of length  $M$  except the all-zero vector) that would yield a  $1 \times M$  matrix of rank 1
- there are  $2^M - 2$  possible choices for the second row (all binary vectors of length  $M$  except the all-zero vector and the vector picked as the first row) that would yield a  $2 \times M$  matrix of rank 2
- there are  $2^M - 2^{k-1}$  possible choices for the  $k$ -th row (all binary vectors of length  $M$  except any binary linear combinations of the first  $k-1$  rows) that would yield a  $k \times M$  matrix of rank  $k$ .

Therefore, the total number of invertible  $K \times M$  matrices is

$$\prod_{k=1}^K (2^M - 2^{k-1}).$$

Dividing by the total number  $2^{KM}$  of  $K \times M$  matrices, we obtain a probability that a random  $K \times M$  matrix is invertible

$$\begin{aligned} P_{\text{invertible}}(K \times M) &= \frac{\prod_{k=1}^K (2^M - 2^{k-1})}{2^{KM}} \\ &= \prod_{k=1}^K (1 - 2^{k-1-M}) \\ &= \prod_{m=M-K+1}^M (1 - 2^{-m}) \end{aligned} \quad (2)$$

For  $M = K$ , i.e., a square  $K \times K$  matrix, this gives the product

$$\frac{1}{2} \times \frac{3}{4} \times \frac{7}{8} \times \frac{15}{16} \times \dots \times \frac{2^K - 1}{2^K}. \quad (3)$$

The limit of the product above as  $K$  goes to infinity is the “magic number”  $\alpha = 0.288788095086602$ . Attempting to fully analyse this limit as  $M$  goes to infinity for all  $K \leq M$  would lead us down a mathematical rabbit hole involving concepts such as Euler’s function,  $q$ -Pochhammer symbols, and the pentagonal number theorem. This is unnecessary and beyond the scope of our course. A few observations will be made to help students understand the significance of this result:

- if the aspect ratio deviation is  $D = M - K = 1$ , then the limit in (3) becomes  $2\alpha$
- if  $D = 2$  then the limit becomes  $2 \cdot \frac{4}{3} \cdot \alpha$

- as the aspect ratio deviation  $D$  increases, the probability that the  $K \times M$  matrix is invertible approaches 1
- remember that  $M = N - L$  so  $D = M - K = N - L - K = N(1 - R) - L$ . In our analysis of the “boring” averaging over  $L$ , we determined that  $P(L < N(p + \epsilon - \delta))$  goes to 1 as  $N$  goes to infinity, hence with a probability approaching 1,

$$D > \delta N$$

which grows with  $N$ , hence the probability that the  $K \times M$  matrix is invertible, which is also the probability of decoding success given  $L$ , goes to 1 as  $N$  goes to infinity

The argument above suffices to establish a vanishing probability of decoding failure conceptually and this approach will be taken during lectures. It gives students a full intuition of why increasing block length results in an improved probability of successful decoding for a given rate below capacity, thereby giving them a working understanding of the coding theorem for the BEC.

A slightly more formal analysis of the infinite product (3) for  $K \rightarrow \infty$  will be given as an exercise, inviting students to note the following

- turn the infinite product (3) into a sum by taking logs
- use the power series expansion of  $\log(1-x)$  to give upper bounds of the form

$$\log(1-x) \leq -x - \frac{x^2}{2} - \frac{x^3}{3} \leq -x - \frac{x^2}{2} \leq -x$$

for non-negative  $x$ , where the linear upper bound  $\log(1-x) \leq x$  gives a bound of  $1/e$  on (3) and the bound can be tightened by either including more terms of the power series expansion of  $\log(1-x)$  or by substituting exact values for a few initial terms

- lower bounds are harder to obtain. We use a tilting trick: consider that for every  $\beta > 1$ ,

$$\log(1-x) \geq -x - \beta \frac{x^2}{2}.$$

for a range of  $x$ . This can be visualised by considering the quadratic upper bound and noting that the factor  $\beta$  tilts the upper bound so it dips below  $\log(1-x)$  and becomes a lower bound for a range until it crosses the curve again. In (3) we only need a lower bound for  $0 \leq x \leq \frac{1}{2}$ , so that  $\beta = 2$  for example does the trick (see Fig. 1).

- showing that (3) is decreasing in  $K$  together with the lower bound established using the tilting trick shows that the infinite product converges to a constant
- the tilting bound and the linear upper bound can be used to derive an upper bound on the probability of decoding failure that is exponential in  $\delta N$ .

## VI. PRACTICE AND EXAMINING

The intended learning outcomes of this part of the course are to gain an understanding of linear coding, of decoding for the binary erasure channel, and of the asymptotics of random coding that demonstrate a vanishing probability of decoding failure as  $N$  grows to infinity as long as  $R < 1 - p$ .

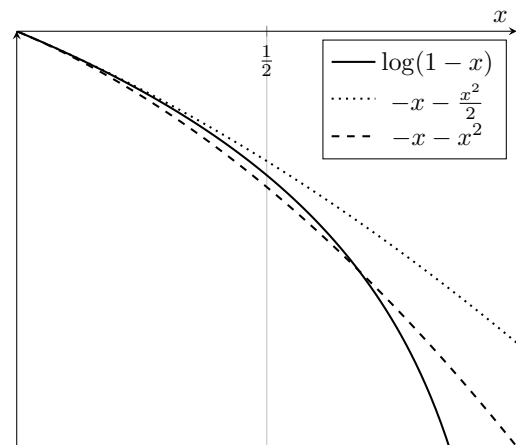


Fig. 1. Quadratic upper bound and tilted lower bound on  $\log(1-x)$  for  $\beta = 2$

The outcomes can be tested through exercises, e.g.,

- check the understanding of rate, dimension, number of codewords for a code described by its encoder matrix  $\mathbf{G}$
- decode a received sequence  $\mathbf{r}$  through Gauss elimination for a small code, in cases where the erasures result in a rank  $K$  residual matrix  $\tilde{\mathbf{G}}$  or otherwise
- determine the probability that a random  $K \times M$  matrix is invertible for reasonably small numbers  $K$  and  $M$ , or equivalently, ask this in the coding context by specifying the dimensions  $K \times N$  of a random encoder matrix  $\mathbf{G}$  and a number  $L$  of erasures
- test understanding of the coding theorem by asking to confirm what happens to the probability of failure as  $N$  goes to infinity for  $R < 1 - p$ , what happens for  $R = 1 - p$  (the probability of success tends to  $\alpha$ ) or what happens for  $R > 1 - p$  (the probability of success tends to 0 because you have more variables than equations in (1))

I have described two “looking further” exercises in the paper but these will not be part of the core learning outcomes and it is not expected that all students will complete them.

The topics of this course, both source and the channel coding parts, lend themselves beautifully to being translated into software practicals. Students could be asked to implement, or complete, programs that implement elementary data compression techniques and decoding of linear codes for the BEC via Gauss elimination. The undergraduate reform of the Engineering Tripos at the University of Cambridge foresees more opportunities for project-based learning but these plans are still being finalised so I am unable to say yet whether practical implementations of the techniques in this course will form part of the undergraduate experience in Cambridge.

## REFERENCES

- [1] P. Elias, *Coding for Two Noisy Channels*, Third London Symposium on Information Theory, The Royal Institution, London, September 12-17, 1955.
- [2] P. Elias, *The Noisy Channel Coding Theorem for Erasure Channels*, The American Mathematical Monthly, Vol. 81, No. 8 (Oct., 1974), pp. 853-862.