On exponential error bounds for random codes on the BSC

G. David Forney, Jr.

1 Introduction

In this note we will revisit the development of exponential error bounds for random codes on the binary symmetric channel (BSC).

This is a very old problem, whose solution has been well known since the classic work of Shannon [S48], Elias [E55, E56], Fano [F61] and Gallager [G63, G65, G68]. However, some features of this solution that were doubtless known to these early researchers do not appear in standard texts such as [G68], [B87] or [CT91], and appear to be little known today. As the development of "random-like" capacity-approaching codes has reawakened interest in this topic, our aim is to provide a clean, modern development that would not be inappropriate in a first course on information theory and coding.

1.1 Random code ensembles and typical codes

In this section we introduce the appropriate random code ensemble (RCE) for the BSC, and also the random linear code ensemble (LCE). We then discuss the properties of the typical linear code (TRC) from the RCE, and the typical linear code (TLC) from the LCE. In particular, we note that whereas the distance distributions of the RCE or LCE and the TLC are exponentially identical for distances above the Gilbert-Varshamov (GV) distance, they are radically different below the GV distance.

Shannon [S48] introduced the idea of a random code ensemble in which every symbol in every codeword is chosen independently at random according to some input distribution $p(x)$.

On the BSC, the input alphabet is binary, and by symmetry the optimum input distribution is equiprobable. A binary code of length N and rate R is a set of $M = e^{NR}$ binary N-tuples. In a binary equiprobable random code ensemble of length N and rate R, the variables are the NM binary symbols $\{x_k^{(i)}, 1 \le i \le N\}$ of the $M = e^{NR}$ codewords $\{\mathbf{x}^{(i)}, 0 \le i \le M-1\}$, which are independent and identically distributed (iid) random variables chosen according to an equiprobable $\{\frac{1}{2}, \frac{1}{2}\}$ distribution.

In this case the probability that a given random codeword $\mathbf{x}$ of length N will be at Hamming distance $d = N\delta$ from an arbitrary binary N-tuple $\mathbf{b}$ is

$$\Pr\{d_H(\mathbf{x}, \mathbf{b}) = d\} = \binom{N}{d} \left(\frac{1}{2}\right)^d \left(\frac{1}{2}\right)^{N-d} \approx e^{-ND(\delta||\frac{1}{2})},$$

where the exponent $D(\delta||\frac{1}{2})$ is the Kullback-Leibler (KL) divergence

$$D(\delta||\frac{1}{2}) = \delta \log \frac{\delta}{\frac{1}{2}} + (1-\delta) \log \frac{1-\delta}{\frac{1}{2}},$$

logarithms are natural, and the symbol "$\approx$" denotes exponential equality; *i.e.,*

$$\lim_{N\to\infty} -\frac{1}{N} \log \Pr\{d_H(\mathbf{x}, \mathbf{b}) = d\} = D(\delta||\frac{1}{2}).$$

By the general properties of divergences, the exponent $D(\delta||\frac{1}{2})$ is strictly convex and has a minimum of 0 at $\delta = \frac{1}{2}$. It may alternatively be written as

$$D(\delta||\frac{1}{2}) = \log 2 - \mathcal{H}(\delta),$$

where $\mathcal{H}(\delta) = -\delta \log \delta - (1-\delta)\log(1-\delta)$ is the entropy of a binary variable with probabilities $\{\delta, 1-\delta\}$.

Let $\mathbf{x}$ be an arbitrary codeword, and let $\overline{\mathcal{N}(\delta)}$ denote the average number of the remaining $M - 1 \approx e^{NR}$ codewords $\mathbf{x}'$ such that $d_H(\mathbf{x}', \mathbf{x}) = N\delta$. Then

$$\overline{\mathcal{N}(\delta)} = (M-1)\Pr\{d_H(\mathbf{x}', \mathbf{x}) = N\delta\} \approx e^{N(R - D(\delta||\frac{1}{2}))}. \tag{1.1}$$

The exponent $R - D(\delta||\frac{1}{2})$ of $\overline{\mathcal{N}(\delta)}$ is plotted in Figure 1. It is strictly concave and symmetrical about $\delta = \frac{1}{2}$. It is nonnegative for $\delta_{\rm GV}(R) \le \delta \le \frac{1}{2}$, where the (relative) Gilbert-Varshamov distance $\delta_{\rm GV}(R)$ is defined as the $\delta \le \frac{1}{2}$ at which the exponent is 0; *i.e.,* as the solution of

$$D(\delta_{\rm GV}(R)||\frac{1}{2}) = R. \tag{1.2}$$

It is negative for $0 \le \delta < \delta_{\rm GV}(R)$.

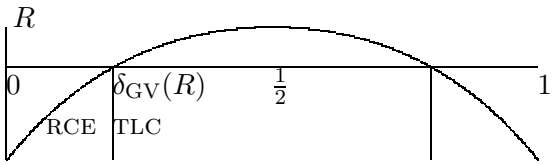

Figure 1. Exponents of average distance distribution $\overline{\mathcal{N}(\delta)}$ for random code ensemble (RCE) and of typical distance distribution $\mathcal{N}_{\rm typ}(\delta)$ for typical linear code (TLC).

If we choose a code at random from the random code ensemble, then for $\delta_{\rm GV}(R) \le \delta \le \frac{1}{2}$ it is highly likely that there will be $\mathcal{N}_{\rm typ}(\delta) \approx \overline{\mathcal{N}(\delta)}$ codewords $\mathbf{x}'$ at distance $N\delta$ from an arbitrary codeword $\mathbf{x}$. On the other hand, for $0 \le \delta < \delta_{\rm GV}(R)$, it is highly likely that there will be no codewords at distance $N\delta$ from an arbitrary codeword $\mathbf{x}$; *i.e.,* $\mathcal{N}_{\rm typ}(\delta) = 0$. Therefore in Figure 1 we indicate that the exponent of $\mathcal{N}_{\rm typ}(\delta)$ goes to $-\infty$ below $\delta_{\rm GV}(R)$. In short, it is highly likely that in a typical random code the minimum distance between an arbitrary codeword and all other codewords will be $N\delta_{\rm GV}(R)$. For $\delta_{\rm GV}(R) \le \delta \le \frac{1}{2}$, there will almost certainly be exponentially many codewords at distance $N\delta$.

Notice that the distribution considered here is not the usual distance distribution of a code, but rather the distance distribution from a given codeword (or from any random binary word) to all other codewords. For our development of RCE bounds, this will actually be the distribution of interest. For random linear codes, we get the same distribution, and moreover the distance distribution from a given codeword is also the distance distribution from every codeword.

For our typical code bounds, we will restrict consideration to linear codes in order to ensure that the typical distance distribution $\mathcal{N}_{\rm typ}(\delta)$ is the same for all codewords. Typical random codes do not have this property. In fact, as Barg has recently shown [B01b], whereas the typical minimum distance from a given codeword in a random code is $N\delta_{\rm GV}(R)$, as shown in Figure 1, the typical minimum distance of the whole random code is much less than $N\delta_{\rm GV}(R)$.

In our subsequent development, we will find the correct error exponents $E_{\rm RCE}(R)$ and $E_{\rm TLC}(R)$ for the random code ensemble and typical linear code, respectively. We will find that the difference in distance distribution exponents illustrated in Figure 1 is reflected in a difference between $E_{\rm RCE}(R)$ and $E_{\rm TLC}(R)$, but only at low rates.

1.2 Summary of results and remarks

For the benefit of the reader who may be wondering what of interest can possibly be said at this date about this old problem, we now summarize our main results and remarks. We do not believe that any of them are new, but on the other hand we believe that most experts in information theory and coding will not have previously appreciated at least some of these points.

1. It is easy to find the correct error exponent $E_{\rm RCE}(R)$ of the RCE via an output-centered analysis that does not explicitly involve the distance distribution $\overline{\mathcal{N}(\delta)}$. The correct exponent is given by

$$E_{\rm RCE}(R) = \begin{cases} R_0 - R, & 0 \le R \le R_{\rm crit}; \\ E_{\rm sp}(R), & R_{\rm crit} \le R \le C, \end{cases} \tag{1.3}$$

where $R_0 = \log 2 - \log(1 + 2\sqrt{p(1-p)})$ is the pairwise error exponent of a BSC with crossover probability p, $R_{\rm crit}$ is the rate at which $\delta_{\rm GV}(R_{\rm crit}) = \tau_{\rm crit}(p)$ (given below in (1.8)), $E_{\rm sp}(R)$ denotes the so-called sphere-packing exponent,

$$E_{\rm sp}(R) = D(\delta_{\rm GV}(R)||p), \tag{1.4}$$

and C (the channel capacity) is the rate at which $\delta_{\rm GV}(C) = p$ and thus $E_{\rm sp}(C) = 0$.

The main points here are:

- The simplicity of the development.
- The fact that the resulting exponent, which is usually characterized as a lower bound on the error exponent of the best possible code (the "reliability function" of the BSC), is in fact the correct error exponent for the RCE. This fact was pointed out by Gallager in [G73], but not in his classic text [G68].
- The breakpoint at $R_{\rm crit}$. For $R_{\rm crit} \le R \le C$, the error probability is dominated by the probability that the number $N\tau$ of channel errors equals or exceeds $N\delta_{\rm GV}(R)$, in which case the number of incorrect codewords at no greater distance from the received codeword becomes exponentially large. For $0 \le R \le R_{\rm crit}$, on the other hand, the error probability is dominated by the probability that there will be $N\tau_{\rm crit}(p)$ channel errors and that a single incorrect codeword will be at the same distance from the received word.

2. The correct error exponent $E_{\rm TLC}(R)$ of the TLC may be found by an input-centered analysis that involves the typical distance distribution $\mathcal{N}_{\rm typ}(\delta)$, combined with a demonstration that $E_{\rm TLC}(R) = E_{\rm RCE}(R)$ at rates $R \ge R_x$. The correct exponent is given by

$$E_{\rm TLC}(R) = \begin{cases} E_x(R), & 0 \le R \le R_x, \\ R_0 - R, & R_x \le R \le R_{\rm crit}; \\ E_{\rm sp}(R), & R_{\rm crit} \le R \le C, \end{cases} \tag{1.5}$$

where $E_x(R)$ is the so-called expurgated exponent given by $E_x(R) = \delta_{\rm GV}(R) D(\frac{1}{2}||p)$, R_x is the rate at which $\delta_{\rm GV}(R_x) = \delta_{\rm crit}(p)$ (given below in (1.11)), and the rest is as before.

The main points here are:

- The exponent is correct for the TLC and not just a bound;
- The development is best based on the output-centered RCE development when $R > R_x$;
- The only rate interval for which the improved distance distribution of the TLC improves the bound is $0 \le R < R_x$. We will elaborate on this point below.

The exponents $E_{\rm RCE}(R)$ and $E_{\rm TLC}(R)$ are illustrated in Figure 2.

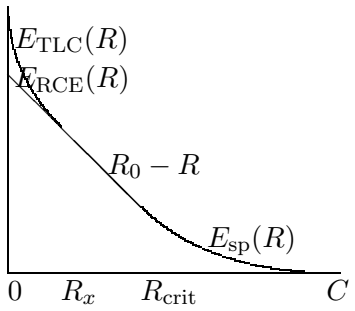

Figure 2. Exponents $E_{\rm RCE}(R)$ and $E_{\rm TLC}(R)$.

3. For both the RCE and the TLC, *given that a decoding error event occurs*, we compute the typical distances between the transmitted (correct) codeword $\mathbf{x}$, the received word $\mathbf{y}$, and the decoded (incorrect) codeword $\mathbf{x}'$. The typical distance $d(\mathbf{x}, \mathbf{x}')$ is denoted by $N\delta_{\rm typ}$, and the typical distance $d(\mathbf{x}, \mathbf{y}) = d(\mathbf{x}', \mathbf{y})$ (or, equivalently, the typical number of channel errors) is denoted by $N\tau_{\rm typ}$. The relative distances $\delta_{\rm typ}$ and $\tau_{\rm typ}$ are illustrated in Figure 3.

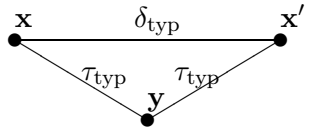

Figure 3. Typical relative distances, given a decoding error, between transmitted (correct) codeword $\mathbf{x}$, received word $\mathbf{y}$, and decoded (incorrect) codeword $\mathbf{x}'$.

For the RCE, we find from the output-centered analysis that

$$\tau_{\rm typ} = \begin{cases} \tau_{\rm crit}(p), & 0 \le R \le R_{\rm crit}; \\ \delta_{\rm GV}(R), & R_{\rm crit} \le R \le C; \end{cases} \tag{1.6}$$

$$\delta_{\rm typ} = 2\tau_{\rm typ}(1 - \tau_{\rm typ}), \tag{1.7}$$

where

$$\tau_{\rm crit}(p) = \frac{\sqrt{p}}{\sqrt{p} + \sqrt{1-p}}, \tag{1.8}$$

and $R_{\rm crit}$ is again the rate at which $\delta_{\rm GV}(R_{\rm crit}) = \tau_{\rm crit}(p)$.

For the TLC, we find from the input-centered analysis that for $0 \le R \le R_{\rm crit}$ we have

$$\delta_{\rm typ} = \begin{cases} \delta_{\rm GV}(R), & 0 \le R \le R_x; \\ \delta_{\rm crit}(p), & R_x \le R \le R_{\rm crit}; \end{cases} \tag{1.9}$$

$$\tau_{\rm typ} = \frac{\delta_{\rm typ}}{2} + p(1 - \delta_{\rm typ}), \tag{1.10}$$

where

$$\delta_{\rm crit}(p) = \frac{2\sqrt{p(1-p)}}{1+2\sqrt{p(1-p)}}, \tag{1.11}$$

and R_x is again the rate at which $\delta_{\rm GV}(R_x) = \delta_{\rm crit}(p)$. For $R_x \le R \le R_{\rm crit}$, these are actually the same as the results for the RCE, although they look quite different. For $R_{\rm crit} \le R \le C$, the results are also the same as the results for the RCE.

The typical relative distances $\tau_{\rm typ}$ and $\delta_{\rm typ}$ are illustrated in Figure 4 as a function of R for both the RCE and the TLC, along with the Gilbert-Varshamov relative distance $\delta_{\rm GV}(R)$.

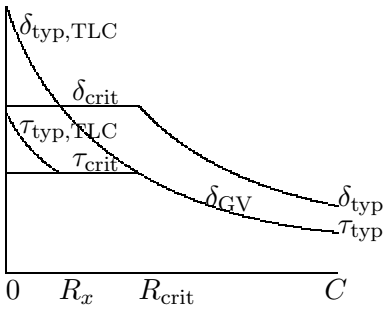

Figure 4. Typical relative distances $\tau_{\rm typ}$ and $\delta_{\rm typ}$ for the RCE and the TLC, with Gilbert-Varshamov relative distance $\delta_{\rm GV}(R)$.

It is interesting to note that the two expressions $\delta_{\rm crit}(p) = 2\tau_{\rm crit}(p)(1-\tau_{\rm crit}(p))$ and $\tau_{\rm crit}(p) = \delta_{\rm crit}(p)/2 + p(1-\delta_{\rm crit}(p))$, which from (1.7) and (1.10) hold simultaneously in the rate interval $R_x \le R \le R_{\rm crit}$, suffice to determine both $\tau_{\rm crit}(p)$ and $\delta_{\rm crit}(p)$ via a simple quadratic equation.

A typical linear code has a minimum distance of $N\delta_{\rm GV}(R)$ between the transmitted codeword and all other codewords. Notice that decoding errors are typically made to minimum-distance codewords only at rates $R \le R_x$. In other words, the improved minimum distance of the TLC over the RCE has an effect only at rates $R \le R_x$; this is why their error exponents differ only in this low-rate region. For rates $R > R_x$, typical decoding errors are to incorrect codewords at relative distances $\delta_{\rm typ} > \delta_{\rm GV}(R)$. The fact that the minimum distance is important only for rates $R \le R_x$, also noted recently by Barg [B01a], does not seem to be widely known; indeed, the literature generally suggests that minimum distance may be important for rates $R \le R_{\rm crit}$, the interval over which the usual union bound gives the correct RCE error exponent.[1]

4. We also find the correct error exponent $E_{{\rm RCE},L}(R)$ of the RCE with list-of-L decoding via a similar analysis. The correct exponent is given by

$$E_{{\rm RCE},L}(R) = \begin{cases} R_{0,L} - LR, & 0 \le R \le R_{{\rm crit},L}; \\ E_{\rm sp}(R), & R_{{\rm crit},L} \le R \le C, \end{cases} \tag{1.12}$$

where $R_{0,L}$ is an exponent to be defined later, $R_{{\rm crit},L}$ is the rate at which

$$\delta_{\rm GV}(R_{{\rm crit},L}) = \tau_{{\rm crit},L}(p) = \frac{p^{1/(1+L)}}{p^{1/(1+L)} + (1-p)^{1/(1+L)}},$$

[1] For instance, in his 1987 information theory text [B87, pp. 186-187], Blahut says:

> Speaking very loosely, and only from the nature of these bounds [on the reliability function], it appears that the performance of good codes of large blocklength has a different cause above $R_{\rm crit}$ than below. Below $R_{\rm crit}$, the average probability of error of a code is dominated by the fact that a typical codeword has a few neighboring codewords with which it is often confused. Then it seems important to design the code so that all codewords are far apart in the sense of a distance known as *Bhattacharyya distance* [which is equivalent to Hamming distance for the BSC].

and $E_{\rm sp}(R)$ is again the sphere-packing exponent of (1.4).

The main points here are:

- Again, the resulting exponent is in fact the correct list-of-L exponent for the RCE.
- The breakpoint $R_{{\rm crit},L}$ decreases monotonically with increasing L, and approaches 0 for large L. Thus with list-of-L decoding we can follow the sphere-packing exponent down to as low a rate as we like. (In practice, the list size L does not need to become very large.)
- The fact that list-of-2 decoding improves the exponent for rates $0 \le R < R_{\rm crit}$ proves that in this rate interval, given an ordinary decoding error event, it is highly likely that there is only one incorrect codeword closer to the received word than the correct codeword.
- On the other hand, no matter how large L is, the list-of-L error exponent is no better than that of ordinary decoding for rates $R_{\rm crit} \le R \le C$. This confirms that in this rate interval, given an ordinary decoding error event, it becomes highly likely as $N \to \infty$ that the number of incorrect codewords closer to the received word than to the correct codeword will be larger than any finite L.

5. Finally, we connect these bounds with the well-known parametric bounds of Gallager [G65], [G68]. In particular, we characterize Gallager's parameter ρ as a Lagrange multiplier, and show that the sphere-packing exponent of (1.4) may be written as a dual convex maximization problem, as follows:

$$E_{\rm sp}(R) = \max_{\rho>0} E_0(\rho) - \rho R, \tag{1.13}$$

where

$$E_0(\rho) = \rho \log 2 - \log \left(p^{1/(1+\rho)} + (1-p)^{1/(1+\rho)} \right)^{1+\rho} \tag{1.14}$$

is Gallager's function for a BSC with crossover probability p. Similarly,

$$E_{\rm RCE}(R) = \max_{0<\rho\le 1} E_0(\rho) - \rho R; \tag{1.15}$$

$$E_{{\rm RCE},L}(R) = \max_{0<\rho\le L} E_0(\rho) - \rho R. \tag{1.16}$$

The main point here is that the parameter ρ appears naturally as a Lagrange multiplier, as it often does in convex optimization problems, whereas Gallager introduces ρ without motivation.

2 Analysis of random codes on the BSC

A binary code $\mathcal{C}$ of length N and rate R nats per symbol is a set of $M = e^{NR}$ codewords $\{\mathbf{x}^{(i)} \in (\mathbb{F}_2)^N, 0 \le i \le M-1\}$.

In Shannon's random code ensemble (RCE), each symbol of each codeword $\mathbf{x}^{(i)} \in \mathcal{C}$ is chosen independently at random according to an equiprobable $\{\frac{1}{2}, \frac{1}{2}\}$ distribution. It is helpful to think of first choosing the codeword $\mathbf{x}^{(0)}$ which is to be transmitted, which we will call the correct codeword, and then choosing the other codewords $\mathbf{x}^{(1)}, \mathbf{x}^{(2)}, \ldots, \mathbf{x}^{(M-1)}$, which we will call the incorrect codewords. The receiver will be informed of the code $\mathcal{C}$, but not of which word is transmitted.

The correct codeword $\mathbf{x}^{(0)}$ is transmitted over a binary symmetric channel (BSC) with error probability p. If the number of channel errors is $t = N\tau$, then the received word $\mathbf{y}$ will be at Hamming distance $N\tau$ from the correct codeword $\mathbf{x}^{(0)}$.

The receiver decodes the received word $\mathbf{y}$ to the unique closest codeword $\mathbf{x}^{(i)} \in \mathcal{C}$, if there is one. A minimum-distance decoding error $\mathcal{E}$ occurs if the correct codeword $\mathbf{x}^{(0)}$ is not the unique closest codeword to $\mathbf{y}$; *i.e.,* if the minimum distance $d = N\delta$ between $\mathbf{y}$ and any incorrect codeword $\mathbf{x}^{(i)}$ is less than or equal to $N\tau$. In other words, the decoding error event is $\mathcal{E} = \{\delta \leq \tau\}$.

2.1 Finding the correct exponent

We will start with an elementary analysis of the probability of error $\Pr(\mathcal{E})$ for this simple RCE/BSC model. We will be interested in the case in which N is large, we will ignore integer constraints, and we will be interested only in determining the correct exponent of quantities that increase or decrease exponentially with N. The *exponent* of a quantity $Q(N)$ that decreases exponentially with N is defined as

$$E_Q = \lim_{N\to\infty} -\frac{\log Q(N)}{N},$$

where "log" denotes a natural logarithm. We express such an asymptotic equality by the notation

$$Q(N) \approx e^{-NE_Q}.$$

For the RCE/BSC model, using an output-centered analysis, we need only two well-known lemmas to find the correct exponent:

Lemma 2.1 (Chernoff exponent) *The correct exponent for the probability that the sum of N iid binary $\{0,1\}$-valued $\{1-p,p\}$-distributed random variables x_i equals or exceeds a threshold $t = N\tau \geq Np$ is the Kullback-Leibler (KL) divergence*

$$D(\tau||p) = \tau \log \frac{\tau}{p} + (1-\tau)\log\frac{1-\tau}{1-p}; \tag{2.17}$$

i.e., if $w(\mathbf{x}) = \sum_i x_i$ is the Hamming weight of $\mathbf{x} = (x_1, \ldots, x_N)$, then

$$\Pr\{w(\mathbf{x}) \geq N\tau\} \approx e^{-ND(\tau||p)}.$$

Similarly, if $\tau \leq p$, then

$$\Pr\{w(\mathbf{x}) \leq N\tau\} \approx e^{-ND(\tau||p)}.$$

As a function of τ in the range $0 < \tau < 1$, $D(\tau||p)$ is analytic, strictly convex, and attains its minimum of 0 at $\tau = p$.

Lemma 2.2 (Union bound exponent) *The correct exponent for the probability that any of $M \approx e^{NR}$ independent events occurs, each event having probability $p \approx e^{-NE}$ with $E > 0$, is*

$$E_{\mathrm{UB}}(R,E) = \max\{E-R, 0\}. \tag{2.18}$$

In other words,

$$1-(1-p)^M \approx e^{-NE_{\mathrm{UB}}(R,E)} = \begin{cases} e^{-N(E-R)}, & \text{if } E \geq R; \\ 1, & \text{if } E \leq R. \end{cases}$$

2.2 Output-centered analysis

The RCE/BSC model is a system consisting of one correct codeword $\mathbf{x}^{(0)}$, a received word $\mathbf{y}$, and a set of $M-1 \approx e^{NR}$ incorrect codewords $\mathbf{x}^{(i)}$. The probability distribution of the whole system factors as follows:

$$p(\mathbf{x}^{(0)}, \mathbf{y}, \{\mathbf{x}^{(i)}\}) = p(\mathbf{x}^{(0)}, \mathbf{y}) \prod_i p(\mathbf{x}^{(i)});$$

i.e., the correct codeword $\mathbf{x}^{(0)}$ and the received word $\mathbf{y}$ are dependent, but the incorrect codewords are independent of $\mathbf{x}^{(0)}$ and $\mathbf{y}$.

Because we are concerned only with the distances of the codewords from $\mathbf{y}$, it makes sense to translate all codewords by $\mathbf{y}$. We will call this an "output-centered analysis."

The translated correct codeword is the channel noise word $\mathbf{n} = \mathbf{x}^{(0)} \oplus \mathbf{y}$. The noise word $\mathbf{n}$ is independent of $\mathbf{y}$, and

$$p(\mathbf{n}) = p^{w(\mathbf{n})}(1-p)^{N-w(\mathbf{n})},$$

where $w(\mathbf{n})$ is the Hamming weight of $\mathbf{n}$ (the number $t = N\tau$ of channel errors). All received words $\mathbf{y} \in (\mathbb{F}_2)^N$ are equiprobable: $p(\mathbf{y}) = 2^{-N}$.

The translated incorrect codewords $\mathbf{z}^{(i)} = \mathbf{x}^{(i)} \oplus \mathbf{y}$ are still independent of $\mathbf{y}$ and equiprobable; *i.e.,* the statistics of the output-translated incorrect codeword set are identical to those of the incorrect codeword set.

It therefore makes sense to think of the whole system as consisting of two independent subsystems, one comprising the noise word $\mathbf{n}$, and the other the output-translated incorrect codeword set $\{\mathbf{z}^{(i)}\}$.

In summary, we now have one subsystem consisting of the noise word $\mathbf{n}$, and a second independent subsystem consisting of $M-1 \approx e^{NR}$ translated incorrect codewords $\mathbf{z}^{(i)}$. A decoding error event $\mathcal{E} = \{\delta \le \tau\}$ occurs if the weight $N\tau$ of $\mathbf{n}$ is greater than or equal to the minimum weight $N\delta$ of the $\mathbf{z}^{(i)}$. We can thus compute $\Pr\{\mathcal{E}\}$ by computing the distributions of τ and δ independently, and then comparing them.

2.3 The subsystem exponents

In the correct subsystem, the Hamming weight $w(\mathbf{n})$ is the sum of N independent, identically distributed (iid) random variables X with alphabet $\{0,1\}$ and probabilities $\{1-p, p\}$. Thus by the Chernoff exponent lemma, for $\gamma > p$ we have

$$\Pr\{w(\mathbf{n}) = N\tau \ge N\gamma\} \approx e^{-ND(\gamma||p)}. \tag{2.19}$$

where the exponent $D(\gamma||p)$ is a strictly convex continuous function of γ whose value and derivative at $\gamma = p$ are both equal to zero. Therefore for $\gamma > p$, $D(\gamma||p)$ is strictly positive and has strictly positive first and second derivatives.

The incorrect subsystem has $M-1 \approx e^{NR}$ independent translated incorrect codewords, each of which has 2^N equiprobable configurations $\mathbf{z}^{(i)} \in (\mathbb{F}_2)^N$.

By the Chernoff exponent lemma, for $\gamma < \frac{1}{2}$, the probability that the Hamming weight $w(\mathbf{z}^{(i)})$ of a given translated incorrect codeword $\mathbf{z}^{(i)}$ is less than or equal to $N\gamma$ is

$$\Pr\{w(\mathbf{z}^{(i)}) \le N\gamma\} \approx e^{-ND(\gamma||\frac{1}{2})}.$$

where the exponent $D(\gamma||\frac{1}{2})$ is a strictly convex function of γ whose value and derivative at $\gamma = \frac{1}{2}$ are both equal to zero. Therefore for $\gamma < \frac{1}{2}$, $D(\gamma||\frac{1}{2})$ is strictly positive and has a strictly negative first derivative and a strictly positive second derivative.

The probability that the minimum translated incorrect codeword weight $d = N\delta$ is less than $N\gamma$ is the probability that any of the $M - 1 \approx e^{NR}$ independent translated incorrect codewords has weight less than $N\gamma$. Therefore, by the union bound exponent lemma, the exponent of this probability is

$$E_{\rm UB}(R,\gamma) = \begin{cases} D(\gamma||\frac{1}{2}) - R, & D(\gamma||\frac{1}{2}) \geq R, \\ 0, & D(\gamma||\frac{1}{2}) \leq R. \end{cases} \tag{2.20}$$

The breakpoint occurs at the γ for which $D(\gamma||\frac{1}{2}) = R$, which is called the (relative) Gilbert-Varshamov distance $\delta_{\rm GV}(R)$; *i.e.,* $D(\delta_{\rm GV}(R)||\frac{1}{2}) = R$. Thus $\delta_{\rm GV}(0) = \frac{1}{2}$, and $\delta_{\rm GV}(R)$ is strictly decreasing and strictly convex as a function of R.

In summary,

$$\Pr\{\min_i w(\mathbf{z}^{(i)}) \leq N\gamma\} \approx \begin{cases} e^{-N(D(\gamma||\frac{1}{2})-R)}, & \gamma \leq \delta_{\rm GV}(R), \\ 1, & \gamma \geq \delta_{\rm GV}(R). \end{cases} \tag{2.21}$$

2.3.1 Weight distribution of translated incorrect codewords

It is instructive to consider also the average number $\overline{\mathcal{N}(\gamma)}$ of translated incorrect codewords with relative weight $\gamma < \frac{1}{2}$. Since there are $M - 1 \approx e^{NR}$ incorrect codewords,

$$\overline{\mathcal{N}(\gamma)} \approx e^{NR} e^{-ND(\gamma||\frac{1}{2})} = e^{-N(D(\gamma||\frac{1}{2})-R)}.$$

Thus $\overline{\mathcal{N}(\gamma)}$ is exponentially large for $R > D(\gamma||\frac{1}{2})$, or equivalently $\gamma > \delta_{\rm GV}(R)$, and exponentially small when $\gamma > \delta_{\rm GV}(R)$. (In fact, this is the same distribution as shown in Figure 1 for the distance between the correct codeword and the incorrect codewords with the RCE.) Therefore the minimum weight $N\gamma$ of the translated incorrect codewords is almost certain to be $N\delta_{\rm GV}(R)$.

2.4 Channel capacity

We have now shown that the probability that the number $w(\mathbf{n}) = N\tau$ of channel errors is greater than $N\gamma$ is exponentially small for $\gamma > p$, whereas it is almost 1 for $\gamma < p$. Similarly, the probability that the minimum translated incorrect codeword weight $\min_i w(\mathbf{z}^{(i)}) = N\delta$ is less than $N\gamma$ is exponentially small for $\gamma < \delta_{\rm GV}(R)$, whereas it is almost 1 for $\gamma > \delta_{\rm GV}(R)$.

Consequently, if $p < \delta_{\rm GV}(R)$, the probability of decoding error $\Pr(\mathcal{E})$ is exponentially small. Indeed, let us take some γ such that $p < \gamma < \delta_{\rm GV}(R)$, and use the following suboptimum "typical-set" decoding rule: if there is a single codeword within distance $N\gamma$ of $\mathbf{y}$, then decode to it; otherwise decoding fails. Since the probability that $w(\mathbf{n}) \geq N\gamma$ is exponentially small and the probability that $\min_i w(\mathbf{z}^{(i)}) \leq N\gamma$ is also exponentially small, it is clear that the probability of decoding error even with this suboptimal rule decreases exponentially with N.

On the other hand, if $p > \delta_{\rm GV}(R)$, then for any γ such that $p > \gamma > \delta_{\rm GV}(R)$, we have $\Pr\{w(\mathbf{n}) \geq N\gamma\} \approx 1$ and $\Pr\{\min_i w(\mathbf{z}^{(i)}) \leq N\gamma\} \approx 1$, so it is clear that $\Pr(\mathcal{E}) \approx 1$.

The rate C at which $p = \delta_{\rm GV}(C)$, namely $C = \log 2 - \mathcal{H}(p)$, is called the channel capacity of the BSC. We have thus shown that $\Pr(\mathcal{E})$ is exponentially small for $R < C$ and $\Pr(\mathcal{E}) \approx 1$ for $R > C$. Our next goal is the find the correct exponent when $R < C$, or equivalently when $\delta_{\rm GV}(R) > p$.

2.5 The correct exponent for the RCE/BSC model

Assuming that $p < \delta_{\rm GV}(R)$, we wish to find the correct error exponent for $\Pr(\mathcal{E}) = \Pr\{\tau \geq \delta\}$. We shall say that an error event $\mathcal{E}_\gamma$ of type γ occurs if $\tau \geq \gamma$ and $\delta \leq \gamma$. Since these two events are independent with exponents given in (2.19) and (2.21), respectively, the joint error probability for $\mathcal{E}_\gamma$ has the joint exponent

$$E_\gamma = D(\gamma||p) + E_i(\gamma) = \begin{cases} D(\gamma||p) + D(\gamma||\frac{1}{2}) - R, & p < \gamma \leq \delta_{\rm GV}(R); \\ D(\gamma||p), & \gamma \geq \delta_{\rm GV}(R). \end{cases}$$

The dominating exponent will therefore be

$$E_{\rm RCE}(R) = \min_{p<\gamma\leq\delta_{\rm GV}(R)} D(\gamma||p) + D(\gamma||\frac{1}{2}) - R.$$

and we will then have

$$\Pr(\mathcal{E}) \approx e^{-NE_{\rm RCE}(R)}.$$

In the range $p < \gamma \leq \delta_{\rm GV}(R) \leq \frac{1}{2}$, we know that $D(\gamma||p)$ is strictly convex, positive, and strictly increasing as a function of γ, whereas $D(\gamma||\frac{1}{2})$ is strictly convex, positive, and strictly decreasing as a function of γ. The minimum of E_γ therefore occurs either at the γ for which $D(\gamma||p) + D(\gamma||\frac{1}{2})$ is minimum, which we will call $\tau_{\rm crit}$, or at the breakpoint $\gamma = \delta_{\rm GV}(R)$, if $\tau_{\rm crit} \geq \delta_{\rm GV}(R)$. We can evaluate $\tau_{\rm crit}$ by solving $D'(\tau_{\rm crit}||p) = -D'(\tau_{\rm crit}||\frac{1}{2})$, which yields (1.8):

$$\tau_{\rm crit} = \frac{\sqrt{p}}{\sqrt{p} + \sqrt{1-p}}.$$

Finally, we define $R_{\rm crit}$ as the rate that satisfies $\tau_{\rm crit} = \delta_{\rm GV}(R_{\rm crit})$.

To illustrate this minimization, we plot in Figure 1 the two exponents

$$\begin{aligned} E_{\rm I}(\gamma) &= D(\gamma||p); \\ E_{\rm II}(\gamma) &= D(\gamma||p) + D(\gamma||\frac{1}{2}) - R, \end{aligned}$$

versus γ for $R > R_{\rm crit}$, $R = R_{\rm crit}$, and $R < R_{\rm crit}$. These two exponents are equal at $\gamma = \delta_{\rm GV}(R)$, and E_γ is equal to their maximum.

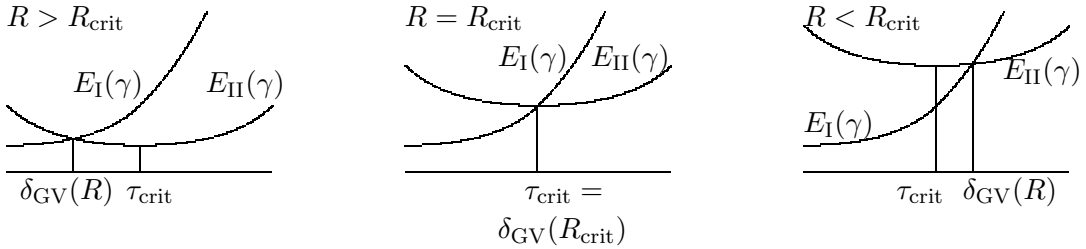

Figure 1. $E_{\rm I}(\gamma)$ and $E_{\rm II}(\gamma)$ as functions of γ for $R > R_{\rm crit}$, $R = R_{\rm crit}$ and $R < R_{\rm crit}$.

For $R > R_{\rm crit}$, the minimum of E_γ occurs at $\gamma = \delta_{\rm GV}(R)$, and equals $E_{\rm RCE}(R) = D(\delta_{\rm GV}(R)||p)$. Alternatively, since $\delta_{\rm GV}(R)$ satisfies $R = D(\delta_{\rm GV}(R)||\frac{1}{2})$, the exponent-rate curve $E_{\rm RCE}(R)$ may be expressed parametrically in this range by

$$\begin{aligned} E_{\rm RCE}(\gamma) &= D(\gamma||p); \\ R(\gamma) &= D(\gamma||\frac{1}{2}), \end{aligned} \tag{2.22}$$

for $p < \gamma < \tau_{\rm crit}$.

For $R < R_{\rm crit}$, the minimum of E_γ occurs at $\gamma = \tau_{\rm crit}$, and equals $E_{\rm RCE}(R) = R_0 - R$, where

$$R_0 = D(\tau_{\rm crit}||p) + D(\tau_{\rm crit}||\frac{1}{2}) = \log 2 - \log(1 + 2\sqrt{p(1-p)}). \tag{2.23}$$

At $R = R_{\rm crit}$, these two curves coincide, since $\tau_{\rm crit} = \delta_{\rm GV}(R_{\rm crit})$ and $D(\delta_{\rm GV}(R_{\rm crit})||\frac{1}{2}) = R_{\rm crit}$.

2.6 Discussion

We conclude that the dominant error mechanisms differ in the low-rate regime $0 \le R < R_{\rm crit}$ and the high-rate regime $R_{\rm crit} < R < C$.

In the high-rate regime, the error probability is dominated by the probability that the number $N\tau$ of channel errors reaches the Gilbert-Varshamov distance, $\tau \ge \delta_{\rm GV}(R)$. We call this a Type I error event. Conditioned on this event, the error probability is ≈ 1, and with high probability there will be an exponentially large number of translated incorrect codewords with weights $\le \tau$. The probability of a Type I error event is thus simply

$$\Pr(\mathcal{E}_{\rm I}) \approx e^{-ND(\delta_{\rm GV}(R)||p)}.$$

In the low-rate regime, the error probability is dominated by the probability that the relative weight of the translated correct codeword is $\tau \approx \tau_{\rm crit}$, and that the minimum relative weight of the $M \approx e^{NR}$ translated incorrect codewords is also $\delta \approx \tau_{\rm crit}$. We call this a Type II error event.

The probability of a Type II error event is

$$\Pr(\mathcal{E}_{\rm II}) \approx Me^{-NR_0},$$

where R_0 is defined by (2.23). This is just M times the pairwise error probability e^{-NR_0} that the relative weight of the translated correct codeword is $\tau \approx \tau_{\rm crit}$, and that the relative weight of a single output-translated incorrect codeword is $\approx \tau_{\rm crit}$. Therefore in this regime the error exponent is correctly given by a pairwise error analysis with the usual union bound.

Finally, we note that the breakpoint $R_{\rm crit}$ is the rate at which the union bound on the minimum incorrect weight

$$\Pr\{\min_i w(\mathbf{z}^{(i)}) \le N\gamma\} \approx e^{-N(D(\gamma||\frac{1}{2})-R)}$$

blows up for $\tau_{\rm crit}$, namely the dominating γ for the joint exponent $D(\gamma||p) + D(\gamma||\frac{1}{2})$. This is why the usual union bound analysis gives the incorrect exponent for $R > R_{\rm crit}$.

2.7 Minimum distance

We note that the arguments above do not directly involve the minimum distance $d_{\min}(\mathcal{C})$ of the random code $\mathcal{C}$; rather, they compare the distance $N\tau$ between the correct codeword $\mathbf{x}^{(0)}$ and the received word $\mathbf{y}$ to the minimum distance $N\delta$ between the incorrect codewords $\{\mathbf{x}^{(i)}\}$ and $\mathbf{y}$.

To consider distances between the correct and incorrect codewords, it is preferable to use an input-centered model in which all words are translated by the correct codeword $\mathbf{x}$. The correct codeword translates to $\mathbf{0}$, the received word $\mathbf{y}$ translates to the noise word $\mathbf{n}$, and the incorrect codewords translate to the input-translated incorrect codewords $\mathbf{w}^{(i)} = \mathbf{x}^{(i)} \oplus \mathbf{x}$, whose weights are the distances $w(\mathbf{w}^{(i)}) = d_H(\mathbf{x}^{(i)}, \mathbf{x})$.

Since the statistics of the input-translated incorrect codewords $\{\mathbf{w}^{(i)}\}$ are identical to those of the output-translated incorrect codewords $\{\mathbf{z}^{(i)}\}$, we can conclude by the same arguments as in our analysis of the incorrect subsystem that $\min_i\{w(\mathbf{w}^{(i)})\}$ is almost surely $\approx N\delta_{\rm GV}(R)$ as $N \to \infty$.

The output-centered analysis shows that for rates $R \geq R_{\rm crit}$, Type I errors predominate; *i.e.,* the probability of error is approximately equal to the probability that the weight $w(\mathbf{n})$ of the noise word exceeds $N\delta_{\rm GV}(R)$. Therefore we can think of a sphere of radius $N\delta_{\rm GV}(R)$ around the correct codeword $\mathbf{x}^{(0)}$ in which errors rarely occur, even though the previous paragraph shows that there are in fact exponentially many incorrect codewords at or beyond the surface of this sphere. Each of these incorrect codewords $\mathbf{x}^{(i)}$ cuts a spherical cap (the intersection of the sphere with the half-space of words closer to $\mathbf{x}^{(i)}$ than to $\mathbf{x}^{(0)}$) out of the sphere to form the actual hard decoding region of $\mathbf{x}^{(0)}$. However, our analysis shows that the probability of $\mathbf{y}$ falling in this hard decoding region is exponentially equal to that of falling in the whole sphere.

If a decoding error occurs to an output-translated incorrect word $\mathbf{z}^{(i)}$, how far is $\mathbf{z}^{(i)}$ likely to be from $\mathbf{n}$, or equivalently $\mathbf{x}^{(i)}$ from $\mathbf{x}^{(0)}$, or equivalently $\mathbf{w}^{(i)}$ from $\mathbf{0}$?

For $R \geq R_{\rm crit}$, the most likely error events occur when $w(\mathbf{n}) \approx N\delta_{\rm GV}(R)$ and $w(\mathbf{z}^{(i)}) \approx N\delta_{\rm GV}(R)$. By the triangle inequality, we have $w(\mathbf{n} \oplus \mathbf{z}^{(i)} = \mathbf{x}^{(0)} \oplus \mathbf{x}^{(i)} = \mathbf{w}^{(i)}) \leq 2N\delta_{\rm GV}(R)$. Moreover, by symmetry between coordinate positions, the typical distance in this case will be $w(\mathbf{w}^{(i)}) \approx 2N\delta_{\rm GV}(R)(1-\delta_{\rm GV}(R))$, the expected distance between two random binary N-tuples chosen independently according to the distribution $\{p(0) = 1-\delta_{\rm GV}(R), p(1) = \delta_{\rm GV}(R)\}$. This typical distance $2N\delta_{\rm GV}(R)(1-\delta_{\rm GV}(R))$ is sometimes called the Elias distance.

If on the other hand $R \leq R_{\rm crit}$, then Type II errors predominate; *i.e.,* the dominant errors occur when $w(\mathbf{n}) \approx \tau_{\rm crit}$ and there is a single output-translated incorrect word $\mathbf{z}^{(i)}$ such that $w(\mathbf{z}^{(i)}) \approx \tau_{\rm crit}$, where for $R < R_{\rm crit}$ we have $\tau_{\rm crit} < \delta_{\rm GV}(R)$. So in this case there is typically a single incorrect word at distance $N\tau_{\rm crit}$ from $\mathbf{y}$, inside the sphere of radius $N\delta_{\rm GV}(R)$ around $\mathbf{y}$, and the correct word is at the same distance.

The typical distance in this case will be $w(\mathbf{w}^{(i)}) \approx 2N\tau_{\rm crit}(1-\tau_{\rm crit}) = \delta_{\rm crit}$, the expected distance between two random binary N-tuples chosen independently according to the distribution $\{p(0) = 1-\tau_{\rm crit}, p(1) = \tau_{\rm crit}\}$. Explicitly, we obtain (1.11):

$$\delta_{\rm crit} = 2\tau_{\rm crit}(1-\tau_{\rm crit}) = \frac{2\sqrt{p(1-p)}}{1+2\sqrt{p(1-p)}}.$$

Note that the critical distance $\delta_{\rm crit}$ does not depend on R, but only on p; *i.e.,* the critical distance remains constant for rates $R \leq R_{\rm crit}$. Note further that this critical distance may still be larger than $\delta_{\rm GV}(R)$. Indeed, since $\tau_{\rm crit} = \delta_{\rm GV}(R_{\rm crit})$, we have $\delta_{\rm crit} = 2\delta_{\rm GV}(R_{\rm crit})(1-\delta_{\rm GV}(R_{\rm crit}))$, the Elias distance at rate $R_{\rm crit}$, which surely exceeds $\delta_{\rm GV}(R_{\rm crit})$. Thus for a range of rates R less than $R_{\rm crit}$, decoding errors *do not* mainly occur to minimum-distance neighbors.

2.8 The typical linear code

Let us define R_x as the rate such that $\delta_{\rm crit} = \delta_{\rm GV}(R_x)$, or equivalently $R_x = \log 2 - \mathcal{H}(\delta_{\rm crit})$. Then for $R < R_x$, errors in the RCE/BSC channel model are typically made to incorrect codewords at distance $\delta_{\rm crit} < \delta_{\rm GV}(R)$, the typical minimum distance in the random linear code ensemble.

Therefore in this range it makes sense to replace the random code ensemble by a code with an improved distance distribution. We will say that a typical code of rate R is a code that

has a typical number $\mathcal{N}_{\rm typ}(\delta)$ of incorrect codewords at distance $N\delta$ from the correct codeword, namely:

$$\mathcal{N}_{\rm typ}(\delta) \approx \begin{cases} e^{-N(D(\delta||\frac{1}{2})-R)}, & \delta \geq \delta_{\rm GV}(R); \\ 0, & \delta < \delta_{\rm GV}(R). \end{cases}$$

The exponent of $\mathcal{N}_{\rm typ}(\delta)$ approaches 0 as $R \downarrow R_{\rm crit}$, and equals $-\infty$ for $R < R_{\rm crit}$ (see Figure 1).

The nicest way of showing that a code with such a distance distribution exists is to construct a random linear code, by randomly choosing $K = \log_2 M$ binary N-tuples as generators. Any input-translated incorrect word $\mathbf{w}^{(i)}$ is then equally likely to be any binary N-tuple, so the expected number of incorrect words at distance $N\delta$ is again $\overline{\mathcal{N}(\delta)} \approx e^{N(R-D(\delta||\frac{1}{2}))}$. Thus as $N \to \infty$ we will almost surely obtain a code with minimum distance $N\delta_{\rm GV}(R)$, and with a typical distance distribution for $\delta \geq \delta_{\rm GV}(R)$. Moreover, this distance distribution will be common to all codewords, not just the transmitted codeword. We call such a code a typical linear code (TLC).

Alternatively, we may construct a code with a typical distance distribution from a random code using the method of "expurgation."

We now analyze the probability of decoding error for both the TLC and the RCE models in the low-rate region $R \leq R_{\rm crit}$, using an input-centered analysis and the usual union bound. We say that an error $\mathcal{E}_\delta$ of type δ occurs if an error is made to an incorrect codeword at distance $N\delta$. The number of such codewords is given by $\mathcal{N}(\delta)$ or $\overline{\mathcal{N}(\delta)}$, and the probability of decoding error to a given such word is the probability of $N\delta/2$ or more channel errors in $N\delta$ given positions, which is $\approx e^{-N\delta D(\frac{1}{2}||p)}$. Thus the exponent for the TLC will be

$$E_{\rm TLC}(R) = \min_{\delta \geq \delta_{\rm GV}(R)} D(\delta||\frac{1}{2}) - R + \delta D(\frac{1}{2}||p),$$

while for the RCE it is simply

$$E_{\rm RCE}(R) = \min_{\delta} D(\delta||\frac{1}{2}) - R + \delta D(\frac{1}{2}||p).$$

Differentiating with respect to δ, we find that the unconstrained minimum is achieved for δ equal to

$$\delta_{\rm crit} = \frac{2\sqrt{p(1-p)}}{1+2\sqrt{p(1-p)}},$$

which agrees with our previous result for the typical distance of RCE errors in the low-rate regime. Moreover, we may verify that the two expressions for the pairwise exponent R_0 are equal:

$$D(\delta_{\rm crit}||\frac{1}{2}) + \delta_{\rm crit} D(\frac{1}{2}||p) = D(\tau_{\rm crit}||p) + D(\tau_{\rm crit}||\frac{1}{2}) = R_0.$$

Finally, from an input-centered perspective the typical noise word has $N\delta/2$ errors in the given $N\delta$ places and typically $pN(1-\delta)$ errors in the remaining places, so we find

$$\tau_{\rm crit} = \frac{1}{2}\delta_{\rm crit} + p(1-\delta_{\rm crit}) = \frac{\sqrt{p}}{\sqrt{p}+\sqrt{1-p}},$$

consistent with our previous result.

Interestingly, the output-centered and input-centered relations

$$\begin{aligned} \delta_{\rm crit} &= 2\tau_{\rm crit}(1-\tau_{\rm crit}); \\ \tau_{\rm crit} &= \frac{1}{2}\delta_{\rm crit} + p(1-\delta_{\rm crit}), \end{aligned}$$

actually suffice to specify both $\delta_{\rm crit}$ and $\tau_{\rm crit}$ via a simple quadratic equation, namely

$$\tau_{\rm crit} = \tau_{\rm crit}(1-\tau_{\rm crit}) + p(\tau_{\rm crit}^2 + (1-\tau_{\rm crit})^2).$$

In summary, this input-centered union bound analysis gives the same result as our previous output-centered analysis for the RCE, which we know to be exact, so the union bound does not blow up for either the RCE or the TLC distribution.

When $\delta_{\rm crit} \geq \delta_{\rm GV}(R)$, both the RCE and the TLC models have the same typical distance $\overline{\mathcal{N}(\delta_{\rm crit})} = \mathcal{N}(\delta_{\rm crit})$, so the exponent is the same. However, when $\delta_{\rm crit} \leq \delta_{\rm GV}(R)$, or equivalently $R \leq R_x$, the probability of decoding error with a typical linear code will dominated by the probability of making an error to a minimum-distance codeword. The exponent of the number of codewords at the minimum distance $\delta_{\rm GV}(R)$ is 0, and the exponent of the event of $\geq \delta_{\rm GV}(R)/2$ errors in a given $\delta_{\rm GV}(R)$ places is

$$E_{\rm TLC}(R) = \delta_{\rm GV}(R) D(\frac{1}{2}||p) = -\delta_{\rm GV}(R) \log 2\sqrt{p(1-p)}, \quad R \leq R_x. \tag{2.24}$$

Alternatively, since $\delta_{\rm GV}(R)$ satisfies $R = D(\delta_{\rm GV}(R)||\frac{1}{2})$, the exponent-rate curve $E_{\rm TLC}(R)$ may be expressed parametrically in this range by

$$\begin{aligned} E_{\rm TLC}(\gamma) &= \gamma D(\frac{1}{2}||p); \\ R(\gamma) &= D(\gamma||\frac{1}{2}), \end{aligned} \tag{2.25}$$

for $\gamma \geq \delta_{\rm crit}$.

For $R \leq R_x$, $E_{\rm TLC}(R)$ is also called the expurgated exponent $E_x(R)$. For $R = R_x$, the expurgated exponent $E_x(R_x)$ is equal to the RCE exponent $E_{\rm RCE}(R_x) = R_0 - R_x$, while for $R < R_x$, $E_x(R)$ exceeds $E_{\rm RCE}(R) = R_0 - R$.

2.9 Sphere-packing bound

The sphere-packing bound is a lower bound on decoding error probability that shows that the exponent developed above is the best possible for $R \geq R_{\rm crit}$.

Any decoding rule partitions the 2^N possible received words into e^{NR} disjoint subsets, namely the decision regions corresponding to each codeword. To minimize the average error probability, it is easy to see that the best case would be if all decoding regions were of equal size $e^{N(\log 2 - R)}$, and as spherical as possible. In this case, since the size of a Hamming sphere of radius $N\gamma$ is $\approx e^{N\mathcal{H}(\gamma)}$, where $\mathcal{H}(\gamma) = -\gamma \log \gamma - (1-\gamma)\log(1-\gamma)$, the radius of the sphere would be approximately $N\gamma(R)$, where

$$\mathcal{H}(\gamma(R)) = \log 2 - R.$$

Since $D(\gamma||\frac{1}{2}) = \log 2 - \mathcal{H}(\gamma)$, this expression is equivalent to $D(\gamma(R)||\frac{1}{2}) = R$. It follows that $\gamma(R) = \delta_{\rm GV}(R)$. Thus in this best possible case, the decoding error probability would be

$$\Pr(\mathcal{E}) \approx e^{-ND(\delta_{\rm GV}(R)||p)}.$$

Thus the sphere-packing exponent is

$$E_{\rm sp}(R) = D(\delta_{\rm GV}(R)||p) \tag{2.26}$$

for all rates $0 \le R < C$, which is equal to $E_{\rm RCE}(R)$ for $R_{\rm crit} \le R < C$. Since $E_{\rm sp}(R)$ is an upper bound on the error exponent for any code $\mathcal{C}$, we conclude that no code can be exponentially better than a random code for $R_{\rm crit} \le R < C$.

Alternatively, since $D(\delta_{\rm GV}(R)||\frac{1}{2}) = R$, the sphere-packing exponent may be expressed in the parametric form

$$\begin{aligned} E_{\rm sp}(R) &= D(\gamma||p); \\ R &= D(\gamma||\frac{1}{2}), \end{aligned} \tag{2.27}$$

for $p < \gamma \le 1/2$.

2.10 Gallager's bounds

In this section we show that Gallager's well-known bounding technique may be viewed as a Lagrangian solution of the constrained maximization problems discussed above.

The sphere-packing exponent may be written as

$$E_{\rm sp}(R) = \min_{\gamma > p: D(\gamma||\frac{1}{2}) = R} D(\gamma||p).$$

Rather than simply solving for γ, we may introduce a Lagrange multiplier $\rho > 0$ for the constraint $D(\gamma||\frac{1}{2}) = R$, which gives the Lagrangian

$$E(\gamma, \rho, R) = D(\gamma||p) + \rho(D(\gamma||\frac{1}{2}) - R).$$

$E(\gamma, \rho, R)$ is strictly convex as a function of γ, and linear as a function of ρ.

We can then express $E_{\rm sp}(R)$ as the solution to an unconstrained convex maximization problem

$$E_{\rm sp}(R) = \max_{\rho > 0} \min_{\gamma > p} E(\gamma, \rho, R).$$

Because of the convexity of $E(\gamma, \rho, R)$, this optimization problem has a unique solution at a unique parameter pair (ρ, γ).

Minimizing $E(\gamma, \rho, R)$ over γ for a given ρ by taking derivatives, we find that

$$\gamma(\rho) = \frac{p^{1/(1+\rho)}}{p^{1/(1+\rho)} + (1-p)^{1/(1+\rho)}}.$$

Substituting this value into the exponent, we obtain

$$E(\gamma(\rho), \rho, R) = E_0(\rho) - \rho R,$$

where

$$E_0(\rho) = \rho \log 2 - \log \left(p^{1/(1+\rho)} + (1-p)^{1/(1+\rho)} \right)^{1+\rho}$$

is Gallager's function for a BSC with parameter p. Then $E_{\rm sp}(R)$ is the solution to the dual convex optimization problem

$$E_{\rm sp}(R) = \max_{\rho > 0} E_0(\rho) - \rho R.$$

This expresses the sphere-packing exponent as the upper envelope of a set of straight lines $E_0(\rho) - \rho R$ of slope $-\rho$. From this we get the parametric equations

$$\begin{aligned} E_{\rm sp}(\rho) &= E_0(\rho) - \rho E_0'(\rho); \\ R(\rho) &= E_0'(\rho). \end{aligned}$$

Similarly, the RCE exponent is

$$E_{\rm RCE}(R) = \min_{\gamma > p} D(\gamma||p) + \max\{D(\gamma||\frac{1}{2}) - R, 0\}.$$

We have already seen that the minimum occurs either at $\gamma = \tau_{\rm crit}$ or $\gamma = \delta_{\rm GV}(R)$, implying that $D(\gamma||\frac{1}{2}) = R$. But rather than simply solving for the minimizing γ, we may again introduce a range-limited Lagrange multiplier ρ with the range $0 \le \rho \le 1$. Since

$$\max\{D(\gamma||\frac{1}{2}) - R, 0\} = \max_{0 \le \rho \le 1} \rho(D(\gamma||\frac{1}{2}) - R),$$

the RCE exponent may then be expressed as

$$E_{\rm RCE}(R) = \min_{\gamma > p} \max_{0 \le \rho \le 1} E(\gamma, \rho, R).$$

Again, because of the convexity of $E(\gamma, \rho, R)$, we may optimize over γ at a given ρ to obtain

$$E_{\rm RCE}(R) = \max_{0 \le \rho \le 1} E_0(\rho) - \rho R.$$

This shows that $E_{\rm RCE}(R) = E_{\rm sp}(R)$ whenever the optimizing ρ is less than or equal to 1, namely when

$$R \ge E_0'(1).$$

It is straightforward to show that $E_0'(1) = R_{\rm crit}$. For $R < R_{\rm crit}$, the optimizing ρ for $E_{\rm sp}(R)$ is greater than 1, so the optimizing ρ for $E_{\rm RCE}(R)$ is 1, which yields

$$E_{\rm RCE}(R) = E_0(1) - R, \quad R \le R_{\rm crit}.$$

It is straightforward to show that $E_0(1) = R_0$.

2.11 List-of-L decoding

Analysis of list-of-L decoding is a sensitive tool for showing that, given a decoding error event, the number of incorrect codewords that are closer to the received word $\mathbf{y}$ than the correct codeword is almost surely 1 for $R < R_{\rm crit}$, but almost surely exponentially large for $R > R_{\rm crit}$.

In list-of-L decoding, the decoder puts out a list of the L closest codewords to the received word rather than the single closest codeword. A list decoding error event $\mathcal{E}_L$ occurs if the correct codeword is not on the list.

Again using the RCE model and an output-centered perspective, let $N\tau = w(\mathbf{n})$, and let $N\delta_L$ be the Lth largest output-translated incorrect codeword weight $w(\mathbf{z}^{(i)})$. Then $\mathcal{E}_L = \{\tau \ge \delta_L\}$.

Since there are $\binom{M}{L}$ possible ways to select L incorrect codewords out of M and the probability that all L have relative weight $\delta \le \gamma$ is $\approx (e^{-ND(\gamma||\frac{1}{2})})^L$, the union bound gives

$$\Pr(\delta_L \le \gamma) \le \binom{M}{L} e^{-NLD(\gamma||\frac{1}{2})} \approx \frac{1}{L!} e^{-NL(D(\gamma||\frac{1}{2}) - R)}.$$

For small (non-exponential) values of L, we may ignore the constant $L!$.

Again, it is straightforward to show that the union bound gives the correct exponent provided that the exponent is nonnegative. Therefore

$$\Pr(\delta_L \le \gamma) \approx e^{-NLE_{\rm UB}(\gamma)};$$

i.e., the list-of-L exponent is L times that of (2.20) for list-of-1 decoding, namely

$$E_{\rm UB}(\gamma) = \max\{0, D(\gamma||\frac{1}{2}) - R\}.$$

We then proceed as before. We say that a list-of-L error event $\mathcal{E}_{\gamma,L}$ occurs if $\tau \ge \gamma$ and $\delta_L \le \gamma$. Since these two events are independent, the joint error probability for $\mathcal{E}_{\gamma,L}$ is given by

$$\Pr(\mathcal{E}_{\gamma,L}) \approx e^{-ND(\gamma||p)} e^{-NLE_{\rm UB}(\gamma)},$$

where we assume that $\delta_{\rm GV}(R) > p$. The dominating type of error event then corresponds to the γ that minimizes the joint exponent

$$E_{\gamma,L} = D(\gamma||p) + LE_{\rm UB}(\gamma) = \begin{cases} D(\gamma||p) + LD(\gamma||\frac{1}{2}) - LR, & p < \gamma \le \delta_{\rm GV}(R); \\ D(\gamma||p), & \gamma \ge \delta_{\rm GV}(R). \end{cases}$$

We thus have

$$\Pr(\mathcal{E}_L) \approx e^{-NE_L(R)},$$

where

$$E_L(R) = \min_{p<\gamma\le\frac{1}{2}} E_{\gamma,L}.$$

The minimum of $E_{\gamma,L}$ now occurs either at the γ for which $D(\gamma||p) + LD(\gamma||\frac{1}{2})$ is minimum, which we will call $\tau_{{\rm crit},L}$, or at the breakpoint, $\gamma = \delta_{\rm GV}(R)$. We can evaluate $\tau_{{\rm crit},L}$ by solving $D'(\gamma||p) = -LD'(\gamma||\frac{1}{2})$, which yields

$$\tau_{{\rm crit},L} = \frac{p^{1/(1+L)}}{p^{1/(1+L)} + (1-p)^{1/(1+L)}}.$$

Thus $\tau_{{\rm crit},L}$ increases with L and approaches $\frac{1}{2}$ as L becomes large. We further define $R_{{\rm crit},L}$ as the value of R for which $\tau_{{\rm crit},L} = \delta_{\rm GV}(R_{{\rm crit},L})$; thus $R_{{\rm crit},L}$ decreases with L and approaches 0 as L becomes large.

The two exponents

$$\begin{aligned} E_{{\rm I},L}(\gamma) &= D(\gamma||p); \\ E_{{\rm II},L}(\gamma) &= D(\gamma||p) + LD(\gamma||\frac{1}{2}) - LR, \end{aligned}$$

are equal at $\gamma = \delta_{\rm GV}(R)$, and $E_{\gamma,L}$ is equal to their maximum.

For $R > R_{{\rm crit},L}$, the minimum of $E_{\gamma,L}$ occurs at $\gamma = \delta_{\rm GV}(R)$, and equals $E_L(R) = D(\delta_{\rm GV}(R)||p)$. Thus $E_L(R) = E_{\rm sp}(R)$ for $R_{{\rm crit},L} \le R < C$, which is a greater interval than for list-of-1 decoding.

For $R < R_{{\rm crit},L}$, the minimum of $E_{\gamma,L}$ occurs at $\gamma = \tau_{{\rm crit},L}$, and equals $E(R) = R_{0,L} - LR$, where

$$R_{0,L} = D(\tau_{{\rm crit},L}||p) + LD(\tau_{{\rm crit},L}||\frac{1}{2}).$$

The exponent $R_{0,L}$ increases with L.

At $R = R_{\mathrm{crit},L}$, these exponents coincide, since $\tau_{\mathrm{crit},L} = \delta_{\mathrm{GV}}(R_{\mathrm{crit},L})$ and $D(\delta_{\mathrm{GV}}(R_{\mathrm{crit},L})||\frac{1}{2}) = R_{\mathrm{crit},L}$.

We conclude that with a small (non-exponential) list size L, the list-of-L decoding exponent $E_L(R)$ can follow the sphere-packing exponent $E_{\mathrm{sp}}(R)$ down to arbitrarily low rates $R > 0$. This improves on $E(R)$ for $R < R_{\mathrm{crit}}$, where Type II error events predominate, but not for $R \geq R_{\mathrm{crit}}$, where Type I error events predominate. This conclusion agrees with our observation that given an ordinary decoding error, for $R \geq R_{\mathrm{crit}}$ there are typically an exponentially large number of incorrect codewords closer to $\mathbf{y}$ than the correct codeword, but for $R < R_{\mathrm{crit}}$ it is unlikely that there will be more than one incorrect codeword closer than the correct codeword. In the latter case, list decoding can help greatly.

Acknowledgment

I am grateful to A. Barg for preprints of [B01a], which contains a number of helpful observations on error bounds for the TLC, and [B01b], which shows that the typical random code has much worse minimum distance that the typical random linear code, and for additional useful comments.

Appendix A: The Chernoff bound and convex optimization

In the correct subsystem, the Hamming weight $w(\mathbf{n})$ is the sum of N iid binary random variables $w(n_k)$ with a Bernoulli distribution $\{1-p, p\}$. Similarly, in the incorrect subsystem, the Hamming weight $w(\mathbf{z}^{(i)})$ of any translated incorrect codeword is the sum of N iid equiprobable binary random variables. We are interested in the probability of large fluctuations of such sums.

The Chernoff bound of large deviation theory gives both an upper bound and the correct exponent for such fluctuations, as we shall show by giving also a lower bound. The derivation has much in common with the derivation of exponential error bounds in the main text. This is not surprising, as both are instances of convex optimization.

A.1 The Chernoff exponent

Let S be the sum of N iid random variables X_k, each with the same alphabet $\mathcal{X}$ and probability distribution $\{p(x), x \in \mathcal{X}\}$. Since the indicator function $\Phi(S \geq N\tau)$ of the event $\{S \geq N\tau\}$ is bounded by

$$\Phi(S \geq N\tau) \leq e^{s(S-N\tau)}$$

for any $s \geq 0$, we have

$$\Pr\{S \geq N\tau\} = \overline{\Phi(S \geq N\tau)} \leq \overline{e^{s(S-N\tau)}} = e^{-N(s\tau-\mu(s))}, \quad s \geq 0,$$

where the overbar denotes expectation, and $\mu(s)$ denotes the semi-invariant moment-generating function of X:

$$\mu(s) = \log \overline{e^{sx}} = \log \sum_x p(x) e^{sx}, \quad s \geq 0.$$

The semi-invariant moment-generating function may be written as $\mu(s) = \log Z(s)$, where

$$Z(s) = \overline{e^{sx}} = \sum_x p(x) e^{sx}, \quad s \geq 0.$$

The notation $Z(s)$ is used because in the terminology of statistical physics, $Z(s)$ is the partition function for the tilted probability distribution $q(x, s) \propto p(x)e^{sx}$; *i.e.,*

$$q(x, s) = \frac{p(x)e^{sx}}{Z(s)}.$$

In statistics, $Z(s)$ is called the moment-generating function of the random variable X.

We may optimize the exponent over $s \geq 0$ to obtain the Chernoff exponent

$$E_c(\tau) = \max_{s \geq 0} s\tau - \mu(s);$$

then we have the Chernoff bound

$$\Pr\{S \geq N\tau\} \leq e^{-NE_c(\tau)}.$$

We can show that this bound is tight by the following lower bound. Let $\mathbf{q} \in \mathcal{P}$ be an arbitrary probability distribution over $\mathcal{X}$, where $\mathcal{P} = \{q(x), x \in \mathcal{X}\}$ is the set of all probability distributions over $\mathcal{X}$. If $\mathcal{X}$ is discrete, we shall say that an N-tuple $\mathbf{x} = (x_1, x_2, \ldots, x_N)$ of elements of $\mathcal{X}$ is of type $\mathbf{q}$ if the number $n(x)$ of appearances of x is equal to $Nq(x)$ for all $x \in \mathcal{X}$.

For N large, the total number of sequences of type $\mathbf{q}$ is a multinomial coefficient $C_{\mathbf{q}}$ which is exponentially equal to

$$C_{\mathbf{q}} \approx e^{N\mathcal{H}(\mathbf{q})},$$

where $\mathcal{H}(\mathbf{q})$ is the entropy of the distribution $\mathbf{q}$. The total probability $P(\mathbf{q}|\mathbf{p})$ under $\mathbf{p}$ of all sequences of type $\mathbf{q}$ is therefore equal to

$$P(\mathbf{q}|\mathbf{p}) = C_{\mathbf{q}} \prod_x p(x)^{Nq(x)} \approx e^{-ND(\mathbf{q}||\mathbf{p})},$$

where the exponent $D(\mathbf{q}||\mathbf{p})$ is the KL divergence

$$D(\mathbf{q}||\mathbf{p}) = \sum_x q(x) \log \frac{q(x)}{p(x)} = \mathsf{E}_{\mathbf{q}}[-\log p(x)] - \mathcal{H}(\mathbf{q}).$$

Now let $\mathbf{q}$ be any probability distribution over $\mathcal{X}$ such that

$$\mathsf{E}_{\mathbf{q}}[x] = \sum_x xq(x) \geq \tau.$$

Then the probability of the event $\{S \geq N\tau\}$ is lowerbounded by the probability of an N-tuple of type $\mathbf{q}$:

$$\Pr\{S \geq N\tau\} \geq \min_{\mathbf{q}\in\mathcal{P}:\mathsf{E}_{\mathbf{q}}[x]\geq\tau} P(\mathbf{q}|\mathbf{p}).$$

Thus we obtain a lower bound involving the minimal type exponent

$$E_t(\tau) = \min_{\mathbf{q}\in\mathcal{P}:\mathsf{E}_{\mathbf{q}}[x]\geq\tau} D(\mathbf{q}||\mathbf{p}).$$

We now show that $E_c(\tau) = E_t(\tau)$ by showing that both are equal to

$$E(\tau) = \max_{s\geq 0} \min_{\mathbf{q}\in\mathcal{P}} D(\mathbf{q}||\mathbf{p}) - s(\mathsf{E}_{\mathbf{q}}[x] - \tau).$$

Since $D(\mathbf{q}||\mathbf{p})$ is a strictly convex function of $\mathbf{q}$ (since $\mathcal{H}(\mathbf{q})$ is strictly concave), and $\mathsf{E}_{\mathbf{q}}[x]$ is linear in $\mathbf{q}$ and $s(\mathsf{E}_{\mathbf{q}}[x] - \tau)$ is linear in s, this is a well-behaved convex optimization problem over a convex region that has a solution at a unique $(s, \mathbf{q})$.

For a given s, we can optimize $\mathbf{q}$ by differentiating with respect to each $q(x)$ and imposing the condition that $\mathbf{q}$ be a probability distribution. The result is that

$$q(x, s) = \frac{p(x)e^{sx}}{Z(s)},$$

where $Z(s) = \sum_x p(x)e^{sx}$. That is, the optimum $\mathbf{q}(s)$ is a tilted probability distribution with "tilt" e^{sx}. For such a tilted distribution $\mathbf{q}(s)$, we have

$$\mathcal{H}(\mathbf{q}(s)) = \mathsf{E}_{\mathbf{q}(s)}[-\log q(s, x)] = \mu(s) + \mathsf{E}_{\mathbf{q}(s)}[-\log p(x)] - s\mathsf{E}_{\mathbf{q}(s)}[x],$$

where we use $\mu(s) = \log Z(s)$, and thus

$$D(\mathbf{q}(s)||\mathbf{p}) - s(\mathsf{E}_{\mathbf{q}(s)}[x] - \tau) = s\tau - \mu(s).$$

It follows that

$$E(\tau) = \max_{s\geq 0} s\tau - \mu(s) = E_c(\tau).$$

On the other hand, we can show that $E(\tau) = E_t(\tau)$ as follows. Since $D(\mathbf{q}||\mathbf{p}) - s(\mathsf{E}_\mathbf{q}[x] - \tau)$ decreases with s if $\mathsf{E}_\mathbf{q}[x] > \tau$ and increases with s if $\mathsf{E}_\mathbf{q}[x] < \tau$, the minimum can only occur when $\mathsf{E}_\mathbf{q}[x] = \tau$. Thus

$$E(\tau) = \min_{\mathbf{q}\in\mathcal{P}:\mathsf{E}_\mathbf{q}[x]=\tau} D(\mathbf{q}||\mathbf{p}) = E_t(\tau).$$

Thus the Chernoff bound exponent is the correct exponent for the probability that $S \geq N\tau$; *i.e.,*

$$\lim_{N\to\infty} \frac{-\log \Pr\{S \geq N\tau\}}{N} = E_c(\tau).$$

We write this as

$$\Pr\{S \geq N\tau\} \approx e^{-NE_c(\tau)}.$$

The language of convex optimization theory sheds some light on this development. The primal problem is to minimize $D(\mathbf{q}||\mathbf{p})$ subject to the constraint that $\mathbf{q}$ be a probability distribution such that $\mathsf{E}_\mathbf{q}[x] \geq \tau$. To solve this problem, we introduce a Lagrange multiplier s and a Lagrangian $D(\mathbf{q}||\mathbf{p}) - s(\mathsf{E}_\mathbf{q}[x] - \tau)$. This yields the dual problem of maximizing $s\tau - \mu(s)$ over the dual variable s. Since the primal problem was strictly convex, the solutions to the primal and dual problems are the same.

A.2 Properties of the Chernoff exponent

Let $X(s)$ be the random variable with the same alphabet as X but with the tilted probability distribution $q(x,s) = p(x)e^{sx}/Z(s)$; then $X(0) = X$. It is easy to see that $Z(0) = 1$, $Z'(s)/Z(s) = \overline{X(s)}$, the mean of $X(s)$, and $Z''(s)/Z(s) = \overline{X^2(s)}$, the second moment of $X(s)$. Consequently

$$\begin{aligned}
\mu(0) &= 0; \\
\mu'(s) &= \frac{Z'(s)}{Z(s)} = \overline{X(s)}; \\
\mu''(s) &= \frac{Z''(s)}{Z(s)} - \left(\frac{Z'(s)}{Z(s)}\right)^2 = \overline{X^2(s)} - \overline{X(s)}^2.
\end{aligned}$$

Thus the second derivative $\mu''(s)$ is the variance of $X(s)$, which is strictly positive unless X is deterministic. We conclude that if X is a nondeterministic random variable with mean $\overline{X}$, then $\mu(s)$ is a strictly convex function of s that equals 0 at $s = 0$ and whose derivative at $s = 0$ is $\overline{X}$.

It follows that the function $s\tau - \mu(s)$ is a strictly concave function of s that equals 0 at $s = 0$ and whose derivative at $s = 0$ is $\tau - \overline{X}$. Thus if $\tau > \overline{X}$, the function $s\tau - \mu(s)$ has a unique maximum which is strictly positive, which occurs at the $s(\tau)$ for which

$$\tau = \mu'(s(\tau)) = \overline{X(s(\tau))},$$

and which equals

$$E_c(\tau) = s(\tau)\tau - \mu(s(\tau)).$$

In other words, we tilt the probability distribution until the tilted mean $\overline{X(s)}$ is equal to τ; this determines $s(\tau)$ and thus $E_c(\tau)$.

In convex optimization theory, $E_c(\tau)$ and $\mu(s)$ are called conjugate functions. It is easy to show from the properties of $\mu(s)$ that $E_c(\tau)$ is a strictly convex function of τ that equals 0 at $\tau = \overline{X}$ and whose derivative at $\tau = \overline{X}$ is 0.

Appendix B: The union bound exponent

In this appendix we prove the lemma given in the text:

Lemma 2.1 (Union bound exponent) *The correct exponent for the probability that any of $M \approx e^{NR}$ independent events occurs, each event having probability $p \approx e^{-NE}$ with $E > 0$, is*

$$E_{\rm UB}(R, E) = \max\{E - R, 0\}.$$

In other words,

$$1 - (1-p)^M \approx e^{-NE_{\rm UB}(R,E)} = \begin{cases} e^{-N(E-R)}, & \text{if } E \geq R; \\ 1, & \text{if } E \leq R. \end{cases}$$

Given an event $\mathcal{E}$ which is a union of subevents $\mathcal{E}_i$, the usual union bound is

$$\Pr(\mathcal{E}) \leq \sum_i \Pr(\mathcal{E}_i).$$

If there are M subevents $\mathcal{E}_i$ and each has the same probability p, then $\Pr(\mathcal{E}) \leq Mp$.

If all subevents are independent, then we have the exact expression

$$\Pr(\mathcal{E}) = 1 - (1-p)^M = Mp - \binom{M}{2}p^2 + \binom{M}{3}p^3 - \cdots,$$

since $\overline{\mathcal{E}}$ is the event that each of the subevents fails to occur. If $Mp < 1$, then this expression may be crudely lowerbounded by

$$\Pr(\mathcal{E}) > Mp - (Mp)^2 - (Mp)^3 - \cdots = Mp\left(\frac{1-2Mp}{1-Mp}\right).$$

Thus if $M \approx e^{NR}$ and $p \approx e^{-NE}$ with $E > R$, then we have

$$Mp \approx e^{-N(E-R)} \geq \Pr(\mathcal{E}) > Mp\left(\frac{1-2Mp}{1-Mp}\right) \approx e^{-N(E-R)},$$

so the correct exponent for $\Pr(\mathcal{E})$ is $E - R$ when $E > R$.

On the other hand, since $1 - p \leq e^{-p}$ (which follows from the basic inequality $\ln x \leq x - 1$), we have

$$\Pr(\mathcal{E}) = 1 - (1-p)^M \geq 1 - e^{-Mp}.$$

Thus if $M \approx e^{NR}$ and $p \approx e^{-NE}$ with $E < R$, then we have $\Pr(\mathcal{E}) \to 1$ as $N \to \infty$, so the correct exponent for $\Pr(\mathcal{E})$ is 0 when $E < R$.

The exponent of 0 for $E = R$ follows by taking the limit from either side.

References

[B01a] A. Barg, "Error bounds for the Hamming space." Working notes, June 2001.

[B01b] A. Barg, "On the size of a typical random code." Working notes, Aug. 2001.

[B87] R. E. Blahut, *Principles and Practice of Information Theory.* Reading, MA: Addison-Wesley, 1987.

[CT91] T. Cover and J. Thomas, *Elements of Information Theory.* New York, Wiley, 1991.

[E55] P. Elias, "Coding for noisy channels," *IRE Convention Record*, Pt. 4, pp. 37-46, 1955.

[E56] P. Elias, "Coding for two noisy channels," in *Information Theory* (C. Cherry, ed.), pp. 61-76. London: Butterworth, 1956.

[F61] R. E. Fano, *Transmission of Information.* Cambridge, MA: MIT Press, 1961.

[G63] R. G. Gallager, *Low-Density Parity-Check Codes.* Cambridge, MA: MIT Press, 1963.

[G65] R. G. Gallager, "A simple derivation of the coding theorem and some applications," *IRE Trans. Inform. Theory*, vol. IT-11, pp. 3-18, Jan. 1965.

[G68] R. G. Gallager, *Information Theory and Reliable Communication.* New York: Wiley, 1968.

[G73] R. G. Gallager, "The random coding bound is tight for the average code," *IEEE Trans. Inform. Theory*, vol. IT-19, pp. 244-246, Mar. 1973.

[S48] C. E. Shannon, "A mathematical theory of communication," *Bell Syst. Tech. J.*, vol. 27, pp. 379-423, 623-656, 1948.