ENEE 739C: Advanced Topics in Signal Processing: Coding Theory
Instructor: Alexander Barg

Lecture 5 (**draft;** 9/3/03). Max-likelihood decoding and error exponents: the Binary Symmetric Channel

http://www.enee.umd.edu/~abarg/ENEE739C/course.html

In the previous lecture we proved that arbitrarily reliable transmission over the binary symmetric channel is possible for every rate R below capacity. More precisely we have shown that the error probability P_e of decoding can be made arbitrarily small at the expense of the increasing delay (code length) at the receiving end. Here we will claim a stronger result, that the decrease rate of P_e can be made exponential in the code length n (this is good news because this means that in principle very low error probability of decoding can be achieved by codes of a not very large length).

To establish this result we return to maximum likelihood decoding: for a received vector $\mathbf{y}$ we decode to the unique nearest codevector $\mathbf{c}$; if the nearest vector is not unique, the decoder outputs an arbitrary codeword.

We will prove the following result.

Theorem 1. *Let $\mathcal{C}$ be a random binary linear code with weight profile α_0 used on a binary symmetric channel BSC(p). When the code length $n \to \infty$, the error probability of maximum likelihood decoding of $\mathcal{C}$ behaves as $P_e(\mathcal{C}) \le 2^{-n(E_0(R,p)-o(1))}$, where*

$$E_0(R,p) = \begin{cases} -\delta_{\rm GV}(R)\log_2 2\sqrt{p(1-p)} & 0 \le R \le R_x, \quad \text{(a)} \\ D(\rho_0\|p) + R_{\rm crit} - R & R_x \le R \le R_{\rm crit}, \quad \text{(b)} \\ D(\delta_{\rm GV}(R)\|p) & R_{\rm crit} \le R \le \mathscr{C} = 1-h_2(p), \quad \text{(c)} \end{cases} \tag{1}$$

where

$$R_x = 1 - h_2(\omega_0) \tag{2}$$

$$R_{\rm crit} = 1 - h_2(\rho_0) \tag{3}$$

$$\rho_0 = \frac{\sqrt{p}}{\sqrt{p}+\sqrt{1-p}}, \qquad \omega_0 := 2\rho_0(1-\rho_0) = \frac{2\sqrt{p(1-p)}}{1+2\sqrt{p(1-p)}}.$$

$E_0(R,p)$ is actually a surface in the 3-space (R,p,E). A typified plot of this surface and its section by the plane $p = 0.02$ is given in Fig. 1.

This theorem has a long history in coding theory (see the remarks in the end of this lecture), therefore there is a whole group of terms that come with it. Part (a) of the expression for E_0 is called the *expurgation exponent*, part (b) is called the *random coding bound*, part (c) is called the *sphere packing bound*. The value $R_{\rm crit}$ is called the *critical rate* of the channel. All these terms have a meaning which will become clear later.

This theorem is also of great importance for coding theory, therefore we not only prove it but also explain what we did and what we realized in the course of the proof once we are done.

Proof : With all the preparation we went through the proof is actually fairly easy. Suppose we transmit with a "random linear code", i.e., a code whose weight profile is given in Corollary 2.7. Assume w.l.o.g. that the transmitted vector is the all-zero one. Let $\mathcal{E}_w(\mathbf{c})$ be the event that the received vector $\mathbf{y}$ was decoded incorrectly to a codeword $\mathbf{c}$ of weight w. Again let us use the union bound to write the probability of error as follows:

$$\begin{aligned} P_e &\le \sum_{w=d}^{2d} \sum_{r=\lceil \frac{w-1}{2} \rceil}^{d} A_w P[\mathcal{E}_w(\mathbf{c})|\,\mathrm{wt}(\mathbf{y}) = r] P[\mathrm{wt}(\mathbf{y}) = r] + P[\mathrm{wt}(\mathbf{y}) \ge d] \\ &:= P_1 + P_2, \end{aligned}$$

where $d = \delta_{\rm GV}(R)n$. As before, we obtain

$$P[\mathcal{E}_w(\mathbf{c})|\,\mathrm{wt}(\mathbf{y}) = r] \cong \frac{\binom{w}{w/2}\binom{n-w}{r-w/2}}{\binom{n}{r}}.$$

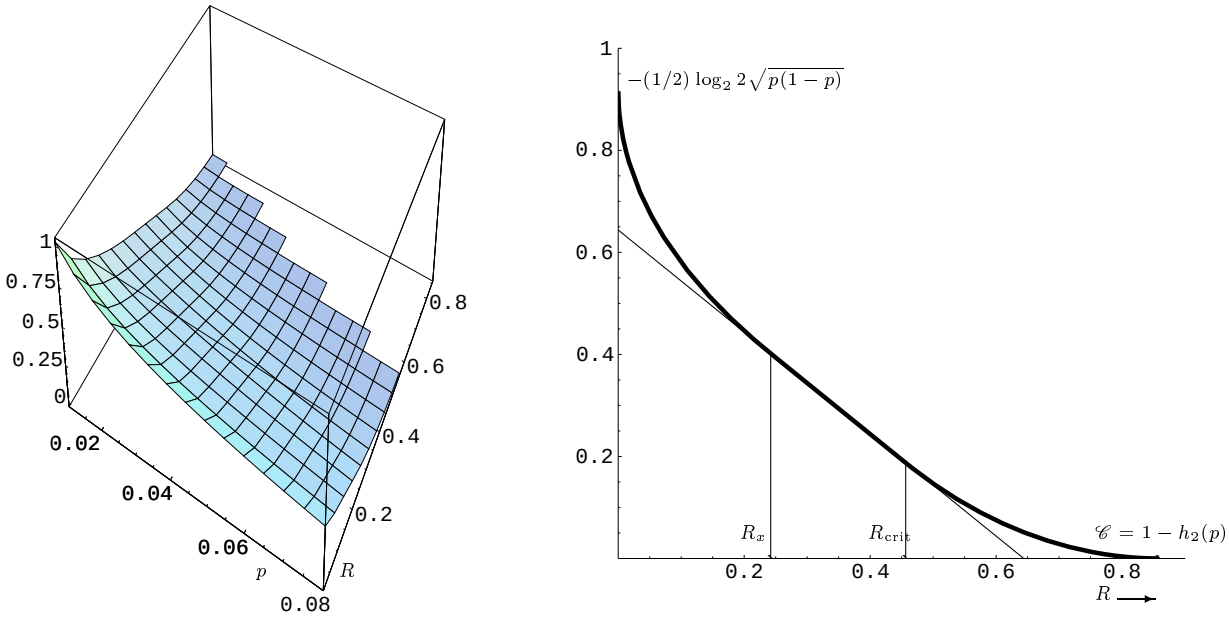


FIGURE 1. a) $E_0(R,p)$; b) section of $E_0(R,p)$ by the plane $p = 0.02$.

Hence

$$P_1 \cong 2^{-n(1-R)} \max_{d/2\le r\le d} \max_{d\le w\le 2r} \binom{n}{w}\binom{w}{w/2}\binom{n-w}{r-w/2} p^r(1-p)^{n-r}. \tag{4}$$

Let $w = \omega n, r = \rho n$. By the Covering Lemma

$$\begin{aligned} P_1 &\cong 2^{-n(1-R)} \max_{d/2\le r\le d} \binom{n}{r}^2 p^r(1-p)^{n-r} \\ &\cong \max_{\delta_{\rm GV}(R)/2\le\rho\le\delta_{\rm GV}(R)} \exp[-n(D(\rho\|p) + (1-R) - h_2(\rho))]. \end{aligned} \tag{5}$$

The unconstrained maximum on ρ on the right-hand side (the minimum of the exponent) is attained for $\rho = \rho_0$, and, by the proof of Covering Lemma, the unconstrained maximum on ω in (4) is attained for $\omega = \omega_0$. The three cases in (1) are realized depending on how ω_0 and ρ_0 are located with respect to the optimization limits.

Part (*b*). The case (1b) corresponds to ω_0, ρ_0 within the limits: $\delta_{\rm GV}(R)/2 \le \rho_0 \le \delta_{\rm GV}(R), \omega_0 \ge \delta_{\rm GV}(R)$. Then the exponent of P_1 is

$$D(\rho_0\|p) + h_2(\delta_{\rm GV}(R)) - h_2(\rho_0), \tag{6}$$

i.e., the random coding exponent of (1b). We need to compare the exponent of P_1 with the exponent $D(\delta_{\rm GV}(R)\|p)$ of P_2. Under the assumption $\rho_0 < \delta_{\rm GV}(R)$ their difference is

$$[D(\rho_0\|p) - h_2(\rho_0)] - [D(\delta_{\rm GV}(R)\|p) - h_2(\delta_{\rm GV}(R))] < 0$$

since $D(x\|p) - h_2(x)$ is an increasing function of x for $x > \rho_0$

$$\Big(\text{indeed, } (D(x\|p) - h_2(x))'_x = \log_2 \frac{x^2(1-p)}{p(1-x)^2} = 0 \text{ for } x = \rho_0 \text{ and } > 0 \text{ for } x > \rho_0.\Big).$$

This proves that the dominating exponent for $R \le R_{\rm crit}$ is given by (6).

Part (*c*). Suppose now that $\omega_0 \ge \delta_{\rm GV}(R)$ and $\rho_0 \ge \delta_{\rm GV}(R)$, i.e., $R \ge R_{\rm crit}$. In this case the exponent of P_1 is dominated by the term with $\rho = \delta_{\rm GV}(R)$. Then we obtain that the exponents of P_1 and P_2 are both equal to the sphere-packing exponent of (1c).

Part (a). If $\omega_0 \le \delta_{\rm GV}$, i.e., $R \le R_x$, the maximum on ω is attained for $\omega = \delta_{\rm GV}$, and we get

$$\sum_{r=\rho n \ge d/2} \binom{n\delta_{\rm GV}}{n\delta_{\rm GV}/2} \binom{n(1-\delta_{\rm GV})}{n(\rho-\delta_{\rm GV}/2)} p^{\rho n}(1-p)^{n(1-\rho)}.$$

This is maximized when $\rho - \delta_{\rm GV}/2 = (1-\delta_{\rm GV})p$, i.e., for

$$\rho = (1-\delta_{\rm GV})p + \delta_{\rm GV}/2.$$

Substituting, we obtain the expurgation exponent of (1a). To finish off this case, we need to show that the exponent $D(\delta_{\rm GV}\|p)$ of the term $P[{\rm wt}(\mathbf{y}) \ge d]$ is greater for $\omega_0 \le \delta_{\rm GV} \le 1/2$ than $-\delta_{\rm GV}(R)\log_2 2\sqrt{p(1-p)}$. This was done in Lemma 3 of lecture 3. ∎

The reader is advised to spend some time on the above proof. The intuition gathered from it serves a basis of many insights into the design of communication systems as well as numerous research problems in coding theory. First it is possible to draw conclusions about the nature of the error events for ML decoding of random codes. The capacity region of the BSC is given on the (R,p)-plane by

$$0 \le R,\ 0 \le p \le 1/2;\ R + h_2(p) \le 1.$$

According to the three cases in the theorem, this region can be partitioned naturally into the regions of low noise A, moderate noise B, and high noise C, where

$$\begin{aligned} A &= \{(R,p) : R \le 1 - h_2(\omega_0)\},\\ B &= \{(R,p) : 1 - h_2(\omega_0) \le R \le 1 - h_2(\rho_0)\},\\ C &= \{(R,p) : 1 - h_2(\rho_0) \le R \le 1 - h_2(p)\}, \end{aligned}$$

see Fig. 2.

As n increases, within each region the error events are dominated by a particular (relative) weight $\omega_{\rm typ}$ of incorrectly decoded codewords. Moreover, the relative weight $\rho_{\rm typ}$ of error vectors that form the main contribution to the error rate also converges to a particular value. We have, for the regions A, B, and C, respectively,

$$\begin{array}{lll} \omega_0 < \delta_{\rm GV}, & \rho_{\rm typ} = (1-\delta_{\rm GV})p + \frac{1}{2}\delta_{\rm GV}, & \omega_{\rm typ} = \delta_{\rm GV},\\ \omega_0 \ge \delta_{\rm GV},\ \rho_0 < \delta_{\rm GV}, & \rho_{\rm typ} = \rho_0, & \omega_{\rm typ} = 2\rho_0(1-\rho_0),\\ \omega_0 \ge \delta_{\rm GV},\ \rho_0 \ge \delta_{\rm GV}, & \rho_{\rm typ} = \delta_{\rm GV}, & \omega_{\rm typ} = 2\delta_{\rm GV}(1-\delta_{\rm GV}). \end{array}$$

When the code is used in the low-noise region, the typical relative weight of incorrectly decoded codewords is $\delta_{\rm GV}(R)n$, i.e., it does not depend on the noise level in the channel. In the moderate-noise region, the typical weight of incorrect codewords is ρ_0 and in the high-noise region it is $\delta_{\rm GV}(R)$. We observe therefore that for $R > R_x$ the error probability does not depend on the minimum distance of the code. The quantity $2\delta_{\rm GV}(R)(1-\delta_{\rm GV}(R))$ is sometimes called the *Elias radius*.

The geometry of decoding for $R < R_{\rm crit}$ and for $R > R_{\rm crit}$ is of very different nature. Consider an error event that corresponds to the moderate-noise region. Its probability is dominated by errors y of relative weight ρ_0. From the proof of the theorem and the Covering Lemma it can be seen that the number of points of the sphere $\mathcal{S}_{\rho_0 n}$ that are decoded incorrectly behaves exponentially as

$$\frac{\binom{n}{\rho_0 n}}{\binom{n}{\delta_{\rm GV} n}} \binom{n}{\rho_0 n};$$

hence their fraction has the same exponent as $\binom{n}{\rho_0 n}/\binom{n}{\delta_{\rm GV} n}$. We see that for $\rho_0 < \delta_{\rm GV}$ an exponentially small fraction of error vectors $\mathbf{y}$ of weight $\rho_0 n$ leads to a decoding error. In case of such an error the weight of the incorrect codeword $\mathbf{c}'$ output by the decoder with probability $\to 1$ is close to $\omega_0 n$. More precisely, if $\mathcal{E}$ is a decoding error and $\mathbf{c}'$ the output of the decoder then the probability

$$\Pr\left[\left|\frac{{\rm wt}(\mathbf{c}')}{n} - \omega_0\right| \ge \alpha \mid \mathcal{E}\right] \le 2^{-nc(\alpha)} \qquad (\alpha > 0).$$

Similarly if $\mathbf{y}$ is the channel output, then typically ${\rm d}(\mathbf{y},\mathbf{c}')/n \to \rho_0$.

So far we have looked at the case of $R, R_{\rm crit}$ or $\rho_0 < \delta_{\rm GV}(R)$. Once R exceeds $R_{\rm crit}$ or $\rho_0 \ge \delta_{\rm GV}$, almost every code point on the sphere $\mathcal{S}_{\rho_0 n}$ leads to a decoding error (again by the covering lemma). Reliable transmission is still possible due to the fact that the total probability of the sphere $\mathcal{S}_{\rho_0 n}$ is exponentially small,

so points on this sphere are received from the channel in an exponentially small fraction of transmissions. For every such points $\mathbf{y}$ there are exponentially many nonzero code vectors which are at least as close to $\mathbf{y}$ as is 0 (see Fig. 2). This jump from one incorrect candidate to exponentially many probably warranted the term "critical rate".

With some additional argument [11] it is possible to prove that for a random linear code and $R < R_{\rm crit}$ the bounding technique used (the union bound) is in fact exponentially tight. With probability one there is at most one nonzero codeword $\mathbf{c}'$ which is as close or closer to $\mathbf{y}$ than 0. Error vectors $\mathbf{y}$ that are incorrectly decoded to $\mathbf{c}'$ occupy the same fraction

$$\cong \frac{\binom{n}{\rho_0 n}}{\binom{n}{\delta_{\rm GV} n}} \frac{\binom{n}{\delta_{\rm GV} n}}{\binom{n}{w}} = \frac{\binom{n}{\rho_0 n}}{\binom{n}{w}}$$

of the sphere $\mathcal{S}_{\rho_0 n}$ for almost every $\mathbf{c}' \in \mathcal{S}_{\omega_0 n}$.

The analysis performed above for a fixed channel and changing code (rate) can be reversed. Namely we can fix a code and change the noise; then the critical probability

$$p_{\rm crit} = \frac{\delta_{\rm GV}^2}{\delta_{\rm GV}^2 + (1-\delta_{\rm GV})^2}$$

and so on.

It is important to realize what we did in this lecture. First we used the result that there exists a code with weight profile α_0 essentially the same as the average weight profile over the ensemble of linear codes. Then, *for that code* we estimated from above the error probability of ML decoding. What if instead of this two-stage procedure we compute right away the average (say over linear codes) error probability of ML decoding? In brief, the result will be the same (see [8] for the first and [10] for the second approach). The same dichotomy applies to the probability of undetected error (see the end of Lecture 3). While the proof we gave is very simple, averaging P_e over code ensembles is substantially more complicated [13].

How good is the error bound derived above? Are there better code families which have smaller error rate on the same channel? In other words, letting

$$P_e(n,R,p) = \min_{\substack{\mathcal{C}\subset\mathscr{H}_2^n \\ |C|=2^{nR}}} P_e(\mathcal{C})$$

$$E_e(R,p) = \lim_{n\to\infty} \frac{1}{n} \log_2 \frac{1}{P_e(n,R,p)}$$

we would like to know if $E_e(R,p) \stackrel{?}{=} E_0(R,p)$. At this moment the answer to this question is not known although many conjecture that this is the equality is indeed true. It is not so difficult to prove ([2, 10, 15]), by providing a matching lower bound on the error probability, that for $R = 0; R \geq R_{\rm crit}$ the bound is actually tight, so what remains is to deal with the case $0 < R < R_{\rm crit}$. Caution: this problem is very difficult.

What to remember:

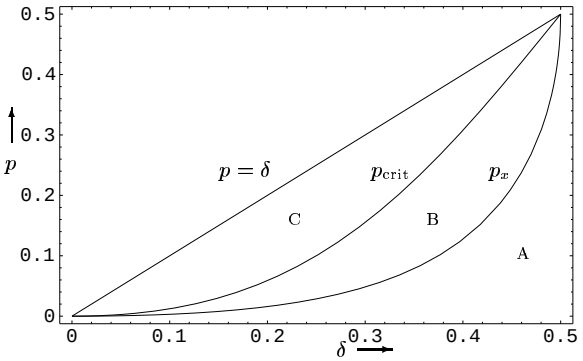
Maximum likelihood decoding of a typical binary linear code on a binary symmetric channel has error probability that falls exponentially with the code length. The exponent of this probability consists of three pieces:
• for low rates (below R_x) the error rate is determined by errors to codewords at a minimum distance from the transmitted word.
• For medium rates (below $R_{\rm crit}$) the error rate is determined by errors made to codewords of some (relative) weight between the GV distance $\delta_{\rm GV}(R)$ and the Elias radius $\delta_E(R)$.
• For rates above $R_{\rm crit}$ the errors are typically made to codewords of relative weight $\delta_E(R)$.

Historical remarks. The body of results discussed in this and the next lectures has a convoluted history. Even though Shannon's theorems prove that it is possible to transmit at any rate below capacity with arbitrarily small error probability (at the expense of growing block length and hence the delay at the receiving end), in his first information theory papers Shannon did not look at the decrease rate of the error.

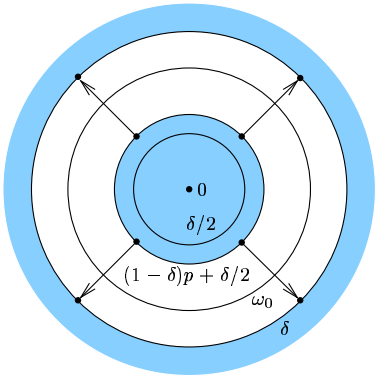
A. Fenstein [7] was the first to derive an exponential estimate of the error probability. P. Elias [5] derived the bound for the binary symmetric channel (except the low-rate part). C. Shannon [14] found a lower bound on the error exponent of the power-constrained Gaussian channel. R. Fano [6] derived error exponents for an arbitrary discrete memoryless channel (DMC). R. Gallager[9] (see also [10]) found a simple proof, and a different algebraic form of the random coding exponent for an arbitrary DMC (and also for finite-state channels). R. Gallager [8] was also the first to study error probability for a particular (linear) code with a given weight distribution. V. D. Goppa [12] suggested nonprobabilistic max-mutual information decoding which was used by I. Csiszár and coauthors [4] (see [3, 1]) to derive an equivalent form of the error exponents essentially in a similar way we did it for the BSC. We will cover this material in the next lecture.

References

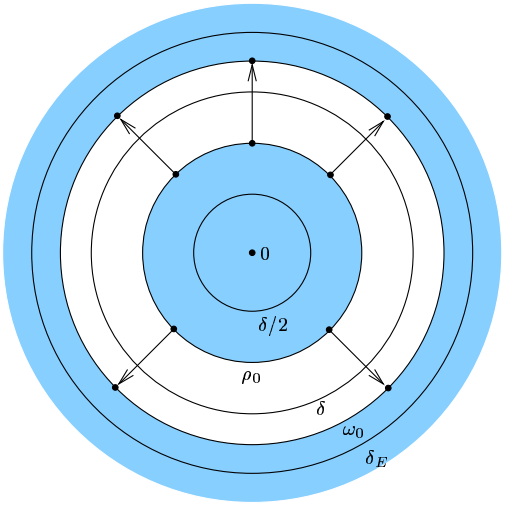
1. I. Csiszár, *The method of types*, IEEE Trans. Inform. Theory **44** (1998), no. 6, 2505–2523, Information theory: 1948–1998. MR **99j:**94016
2. I. Csiszár and J. Körner, *Graph decomposition: A new key to coding theorems*, IEEE Trans. Inform. Theory **27** (1981), no. 1, 5–12.
3. ———, *Information theory. Coding theorems for discrete memoryless channels*, Akadémiai Kiadó, Budapest, 1981.
4. I. Csiszár, J Körner, and Marton K., *A new look at the error exponent of a discrete memoryless channel*, unpublished. Presented at the IEEE International Symposium on Information Theory, October 1977, Cornell Univ., Ithaca, NY.
5. P. Elias, *Coding for two noisy channels*, Information Theory. Third London Symposium (C. Cherry, ed.), Academic Press, New York, 1956, pp. 61–74.
6. R. M. Fano, *Transmission of information*, Wiley, New York, 1961.
7. A. Feinstein, *A new basic theorem of information theory*, IRE Trans. Inform. Theory **PGIT-4** (1954), 2–22.
8. R. G. Gallager, *Low-density parity-check codes*, MIT Press, Cambridge, MA, 1963.
9. ———, *A simple derivation of the coding theorem and some applications*, IEEE Trans. Inform. Theory **11** (1965), 3–18.
10. ———, *Information theory and reliable communication*, John Wiley & Sons, New York e.a., 1968.
11. ———, *The random coding bound is tight for the average code*, IEEE Trans. Inform. Theory (1973), no. 2, 244–246.
12. V. D. Goppa, *Nonprobabilistic mutual information without memory*, Problems of Control and Information Theory **4** (1975), no. 2, 97–102. MR 56 #11530
13. V. I. Levenshtein, *Bounds on the probability of undetected error*, Problems of Information Transmission **13** (1977), no. 1, 3–18.
14. C. E. Shannon, *Probability of error for optimal codes in a Gaussian channel*, Bell Syst. Techn. Journ. **38** (1959), no. 3, 611–656.
15. A. J. Viterbi and J. K. Omura, *Principles of digital communication and coding*, McGraw-Hill, 1979.



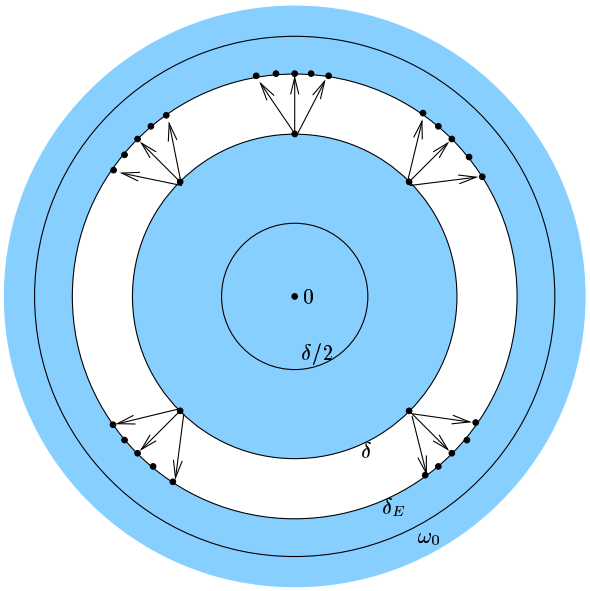
Regions of low (A), moderate (B), and high (C) noise for random codes



Low noise: $0 < p \le p_x$
$\rho_{\rm typ} = ((1-\delta)p + \delta/2)$; $\omega_{\rm typ} = \delta$



Moderate noise: $p_x < p \le p_{\rm crit}$
$\rho_{\rm typ} = \rho_0$; $\omega_{\rm typ} = \omega_0$



High noise: $p_{\rm crit} < p \le \delta$
$\rho_{\rm typ} = \delta$; $\omega_{\rm typ} = \delta_E$

FIGURE 2. Decoding geometry of random codes; in the case of decoding error the most likely weight of the error pattern in the channel is $w(\mathbf{e})$ and the weight of the decoder output is $w(\mathbf{c})$. $\delta = \delta_{\rm GV}(R)$ and $\delta_E = 2\delta_{\rm GV}(1-\delta_{\rm GV})$ is the Elias radius.