Coding for New Applications in Storage Media

Artyom Sharov

Coding for New Applications in Storage Media

Doctoral Thesis Research Proposal

Artyom Sharov

Under the supervision of Prof. Ron M. Roth

Computer Science Department
Technion – Israel Institute of Technology
Haifa 32000, Israel
e-mail: sharov@cs.technion.ac.il

May 2011

Contents

List of Figures

List of Tables

Notation and Abbreviations

$\mathbb{Z}$	—	the set of integers
$\mathbb{Z}^+$	—	the set of positive integers
$\mathbb{R}$	—	the set of real numbers
$[s]$	—	the set $\{0, 1, \ldots, s-1\}$
Σ	—	a finite alphabet
Σ^n	—	the set of words of length n over alphabet Σ
$\lambda(A)$	—	the Perron eigenvalue of a square matrix A
$\mathcal{G} = (V, E)$	—	a finite labeled graph with set of states V and set of edges E
$A_{\mathcal{G}}$	—	the adjacency matrix of graph $\mathcal{G}$
$M_q(n, t)$	—	the largest size of any t-grain-correcting code over q-ary alphabet
$\mathsf{M}_q(n, t)$	—	the largest size of any t-grain-detecting code over q-ary alphabet
$R_q(\tau)$	—	the asymptotic rate of $\lceil \tau n \rceil$-grain-correcting codes over q-ary alphabet
$\mathsf{R}_q(\tau)$	—	the asymptotic rate of $\lceil \tau n \rceil$-grain-detecting codes over q-ary alphabet
WOM	—	write-once memory
WEM	—	write-efficient memory
i.i.d.	—	independent and identically distributed
cws	—	confusable in a wider sense

Chapter 1

Introduction

Conventional magnetic recording media are composed of basic magnetizable units called *grains* which might be random in size and shape. Information is stored on the medium through a write mechanism that sets the magnetic polarities of the grains. Each grain can be magnetized to take on one of the two possible types of magnetic polarity. Thus, each grain represents at most one bit of information. If the boundaries of the grains had been known to the write and the readback apparatuses, then, theoretically, it would have been possible to attain the storage capacity of one bit per grain.

However, even if the write and the readback apparatuses were aware of the shapes and locations of specific grains in the medium, it would still be impossible to attain the capacity of one bit per grain since the existing technologies are still incapable of setting magnetic polarities of regions as small as a single grain. In current technologies, the writing in a magnetic medium is carried out by dividing the medium into periodically partitioned *cells* and writing one bit of data into these cells. Since a cell is typically larger in size than a grain, the writing of a bit into a cell boils down to uniformly magnetizing all the grains inside the cell (whereas the grains straddling the boundary between cells can be neglected). For a general background on the subject of magnetic recording we refer the reader to [35].

Recently, Wood *et al.* [36] proposed an approach whereby magnetizing areas as small as the size of grains are achievable, thereby effectively creating a different type of medium where the grain polarity is determined by the last bit written into the grain. To simplify the original two-dimensional problem, we will model the new medium as a one-dimensional array of bit cells over which the granular structure is imposed by means of grouping adjacent cells into grains of arbitrary lengths. Reading information from the chain of cells is considered reliable whereas writing within a cell overwrites the contents of all the cells pertaining to the same grain, thereby introducing *substitution errors*. This combinatorial error model was considered by Mazumdar *et al.* [26]; we elaborate on the model and the results obtained in [26] in Section 1.1.

A different way of modeling the same medium is a probabilistic one. Instead of introducing the granular structure, we can assume that errors can occur with a prescribed probability distribution only in the cells whose value differs from the value of the previously-written cell. This characterization introduces a one-dimensional *channel*, that is, a "black box" with inputs and outputs such that an output can differ from the corresponding input due to the adversary distortions of the medium. We elaborate on the probabilistic model, suggested by Iyengar *et al.* [13], in Section 1.2; the survey of the results of [13] is presented therein as well.

Another example of a (relatively) recently emerged storage media are flash memories. Interestingly, the research on the subject starts with the paper of Rivest and Shamir [28] in 1982 who (in the context of digital optical disks) considered the *write-once memory* (in short, WOM) whose basic storage elements could irreversibly transit from state 0 to state 1. An even earlier example of a WOM is the punch card. A *flash memory* is built of blocks of memory cells where each cell can take one of the possible q values from 0 to $q-1$ (determined by the appropriate threshold voltage). Unlike in WOMs, the memory cells can all be erased (i.e. set to value 0), yet the erasure operation is several orders of magnitude slower than the writing, hence, to save time, it makes sense to erase cells in blocks, rather than individually. Any transition other than erasure cannot decrease the values of memory cells. An additional characterization of the flash memory is that a memory cell can endure only a limited number of erasures (typically, $\sim 10^5$), posing the problem of the optimal leveling of the wear among the memory cells in the flash memory. We review the notable results related to this medium in Section 1.3.

Our own contribution and insights into the posed problem are unveiled in Section 2; therein we somewhat extend the combinatorial model of [26]. The main inspiration for our results is the theoretical platform established in [24] (which, in turn, relies on the idea of [19]) central points of which are recapped in Section 2.1.

Chapter 3 lists the objectives that we set for the future research.

1.1 Combinatorial error model

In what follows, we will define a somewhat generalized version of the combinatorial error model of the one-dimensional granular medium considered by Mazumdar *et al.* [26].

Let $[s]$ denote the set $\{0, 1, \ldots, s-1\}$ for any positive integer s. Let q be a positive integer, and let $\Sigma = [q]$ be an alphabet. A *grain (of length* 2*)* ending at location e in a word $\mathbf{x} = (x_i)_{i \in [n]}$ of length n over Σ causes the value of x_e to equal that of x_{e-1}. Given n consecutive positions on the medium (where words of length n over Σ are to be written), define a *grain pattern* as a set $\mathcal{S} \subseteq [n] \setminus \{0\}$ containing all the grain locations in these n positions. We will commonly refer to the elements of $\mathcal{S}$ (which indicate grain locations) simply as grains. Thus, a grain pattern $\mathcal{S}$ inflicts errors to a word $\mathbf{x} = (x_i)_{i \in [n]}$ over Σ by means of the smearing operator $\sigma = \sigma_{\mathcal{S}}$ that yields an output word $\mathbf{y} = (y_i)_{i \in [n]} = \sigma(\mathbf{x})$ over Σ in the following way: for any index $e \in [n] \setminus \{0\}$, $y_e = x_{e-1}$ if $e \in \mathcal{S}$ and $y_e = x_e$ otherwise. We will say that a grain pattern has *overlaps* if there exist two grains $e, e' \in \mathcal{S}$ such that $e' = e+1$; otherwise the grain pattern will be called *nonoverlapping*.

Example 1.1. Let $\Sigma = [3]$ $(q = 3)$, $n = 6$, $\mathbf{x} = 102022$, $\mathcal{S} = \{1, 3, 5\}$ and $\mathcal{S}' = \{1, 2\}$. Then $\sigma_{\mathcal{S}}(\mathbf{x}) = 112222$ and $\sigma_{\mathcal{S}'}(\mathbf{x}) = 110022$. The grain pattern $\mathcal{S}$ is nonoverlapping whereas the grain pattern $\mathcal{S}'$ has overlaps. □

For a positive integer t and $\mathbf{x} \in \Sigma^n$, let $\mathcal{R}_t(\mathbf{x})$ be defined as the set of all words $\mathbf{y} \in \Sigma^n$ such that there exist grain patterns $\mathcal{S}, \mathcal{S}'$ of size t at most for which $\sigma_{\mathcal{S}}(\mathbf{x}) = \sigma_{\mathcal{S}'}(\mathbf{y})$. We will refer to $\mathcal{R}_t(\mathbf{x})$ as the *risk set* of $\mathbf{x}$ of radius t. Two words $\mathbf{x}, \mathbf{y} \in \Sigma^n$ are t*-confusable* if $\mathbf{y} \in \mathcal{R}_t(\mathbf{x})$ (and therefore $\mathbf{x} \in \mathcal{R}_t(\mathbf{y})$). Words $\mathbf{x}$ and $\mathbf{y}$ are *finitely-confusable* if they are t-confusable for some finite t; otherwise, we say that they are ∞*-confusable*. A code $\mathcal{C}$ of length n over Σ (namely, a nonempty subset of Σ^n) is called t*-grain-correcting* if no two distinct codewords in $\mathcal{C}$ are t-confusable. Let $M_q(n, t)$ denote the largest size of any t-grain-correcting code of length n over Σ. For $\tau \in (0, 1)$, define the (asymptotic) *rate* of $\lceil \tau n \rceil$-grain-correcting codes over Σ as

$$R_q(\tau) = \limsup_{n \to \infty} \frac{1}{n} \log_q M_q(n, \lceil \tau n \rceil).$$

By the same token, two words $\mathbf{x}, \mathbf{y} \in \Sigma^n$ are t*-undetectable* if there exists a grain pattern $\mathcal{S}$ of size t at most for which either $\sigma_{\mathcal{S}}(\mathbf{x}) = \mathbf{y}$ or $\sigma_{\mathcal{S}}(\mathbf{y}) = \mathbf{x}$. Words $\mathbf{x}$ and $\mathbf{y}$ are *finitely-undetectable* if they are t-undetectable for some finite t; otherwise, we say that they are ∞*-undetectable*. A code $\mathcal{C}$ of length n over Σ is called t*-grain-detecting* if no two distinct codewords in $\mathcal{C}$ are

t-undetectable. Let $\mathsf{M}_q(n,t)$ denote the largest size of any t-grain-detecting code of length n over Σ. For $\tau \in (0,1)$, define the (asymptotic) *rate* of $\lceil \tau n \rceil$-grain-detecting codes over Σ as

$$\mathsf{R}_q(\tau) = \limsup_{n\to\infty} \frac{1}{n} \log_q \mathsf{M}_q(n, \lceil \tau n \rceil).$$

We will see in Section 2.5 that, in fact, $\mathsf{R}_q(\tau) = 1$ for any alphabet size q and any τ.

Finally, let $\mathsf{H}_q : [0,1] \to [0,1]$ denote the *q-ary entropy function*:

$$\mathsf{H}_q(p) = -p \log_q p - (1-p) \log_q (1-p) + p \log_q (q-1).$$

Using a combinatorial method based on the counts of runs, the following upper bound [26, Th. 2] on $M_2(n,t)$ was obtained for a fixed value of t.

Theorem 1.1. *Let $n, t \in \mathbb{Z}^+$ then*

$$M_2(n,t) \le \frac{2^{n+1} t!}{n^t} (1+o(1))$$

where $o(1) \to 0$ as $n \to \infty$. □

Applying the same method in the asymptotic context yields the next upper bound [26, Prop. 3] on the rate $R_2(\tau)$.

Proposition 1.2. Let $\tau \in (0, 0.072]$ and let $\mathsf{x} = \mathsf{x}(\tau)$ be the smallest positive solution of

$$\mathsf{H}_2(\mathsf{x}) + \frac{\mathsf{x}}{2} \cdot \mathsf{H}_2\left(\frac{2\tau}{\mathsf{x}}\right) = 1.$$

Then $R_2(\tau) \le \mathsf{H}_2(\mathsf{x})$. □

We will elaborate on the run-based method in Section 2.3.

A simple construction was pointed out [26, Sec. 2] for ∞-grain-correcting codes of any length n which we cite herein.

Construction 1.3. For an even positive n, the binary code

$$\mathcal{C}_n = \left\{ \mathbf{c} = (c_i)_{i\in[n]} : c_{2s} = c_{2s+1} \text{ for any } s \in [n/2] \right\}$$

is an ∞-grain-correcting code of length n and size $2^{n/2}$.

For an odd positive n, the code $\mathcal{C}_n = (0 \mid \mathcal{C}_{n-1}) \cup (1 \mid \mathcal{C}_{n-1})$ is a binary ∞-grain-correcting code of length n and size $2^{\lceil n/2 \rceil}$. □

In addition to the bound of Theorem 1.1, an additional general upper bound [26, Cor. 5] on $M_2(n,t)$ was obtained by investigating properties of the *confusability graph* of the code space. We cite its special case as the following lemma.

Lemma 1.4. For any $n \in \mathbb{Z}^+$, $M_2(n, \lfloor n/2 \rfloor) \leq 2^{\lceil n/2 \rceil}$. □

Construction 1.3 and Lemma 1.4 readily yield the following result.

Theorem 1.5. *For any* $n \in \mathbb{Z}^+$, $M_2(n, \lfloor n/2 \rfloor) = 2^{\lceil n/2 \rceil}$. □

1.2 Probabilistic error model

Iyengar *et al.* [13] modeled the granular medium with grains of length 2 as a channel by means of three infinite binary vectors $\mathbf{X} = (X_i)_{i=0}^{\infty}, \mathbf{Y} = (Y_i)_{i=0}^{\infty}, \mathbf{Z} = (Z_i)_{i=0}^{\infty}$ where $\mathbf{X}$ represents the input data, $\mathbf{Y}$ represents the data actually written on the medium and $\mathbf{Z}$ represents the error to the input, which is independent from $\mathbf{X}$ and $\mathbf{Y}$. The relation between the three is captured by the formula

$$Y_i = X_i \oplus (X_i \oplus X_{i-1}) \otimes Z_i \text{ for } i \in \mathbb{Z}^+ \cup \{0\}\,, \tag{1.1}$$

where $\oplus$ and $\otimes$ are binary addition and multiplication, respectively. Another formula expressing the same relation is $Y_i = X_{i-Z_i}$ for $i \in \mathbb{Z}^+ \cup \{0\}$.

Denote the random processes whose realizations are $\mathbf{X}$ and $\mathbf{Z}$ by $\mathcal{X}$ and $\mathcal{Z}$, respectively. Denote by $\mathtt{I}_{\mathrm{gr}}(\mathcal{X}, \mathcal{Z})$ the mutual information per bit for given probability distributions of the input process $\mathcal{X}$ and the error process $\mathcal{Z}$.

Iyengar *et al.* obtained lower and upper bounds on $\mathtt{I}_{\mathrm{gr}}(\mathcal{X}, \mathcal{Z})$ under different characterizations of $\mathcal{X}$ and $\mathcal{Z}$. If $\mathcal{Z}$ is an i.i.d. Bernoulli process with parameter p and $\mathcal{X}$ is an i.i.d. process with uniform distribution, then

$$\mathtt{I}_{\mathrm{gr}}(\mathcal{X}, \mathcal{Z}) \geq \frac{1{+}\mathsf{p}}{2}\mathsf{H}_2\left(\frac{1{+}\mathsf{p}^2}{2{+}2\mathsf{p}}\right) + \frac{1{-}\mathsf{p}}{2}\mathsf{H}_2\left(\frac{1{-}\mathsf{p}}{2}\right) - \frac{1}{2}\mathsf{H}_2(\mathsf{p})$$

and

$$\mathtt{I}_{\mathrm{gr}}(\mathcal{X}, \mathcal{Z}) \leq \frac{1{-}\mathsf{a}}{2}\mathsf{H}_2\left(\frac{1}{2{-}2\mathsf{a}}\right) + \frac{1{+}\mathsf{a}}{2}\mathsf{H}_2\left(\frac{1}{2{+}2\mathsf{a}}\right)\frac{1}{2}\mathsf{H}_2(\mathsf{p})$$

where $\mathsf{a} = \mathsf{p}(1{-}\mathsf{p})$.

If $\mathcal{Z}$ is an i.i.d. Bernoulli process with parameter p and $\mathcal{X}$ is a symmetric Markov process with memory 1 and transition parameter α (see [6, Sec. 2.8] and Section 2.1), then

$$\mathtt{I}_{\mathrm{gr}}(\mathcal{X}, \mathcal{Z}) \geq (1{-}\alpha(1{-}\mathsf{p}))\mathsf{H}_2\!\left(\frac{\alpha(1{-}\alpha(1{-}\mathsf{p}^2))}{1{-}\alpha(1{-}\mathsf{p})}\right) + \alpha(1{-}\mathsf{p})\mathsf{H}_2(\alpha(1{-}\mathsf{p})) - \alpha\mathsf{H}_2(\mathsf{p})$$

and

$$\begin{aligned}
\mathtt{I}_{\mathrm{gr}}(\mathcal{X},\mathcal{Z}) \quad \leq \quad & \left[\alpha^2(1-\alpha)(1-\mathsf{a})+\alpha^2(3-\alpha)\mathsf{a}+(1+\alpha)(1-\alpha)^2\right]\times \\
& \mathsf{H}_2\left(\frac{\alpha^2(1-\alpha)(1-\mathsf{a})+\alpha^3\mathsf{a}+\alpha(1-\alpha)^2}{\alpha^2(1-\alpha)(1-\mathsf{a})+\alpha^2(3-\alpha)\mathsf{a}+(1+\alpha)(1-\alpha)^2}\right)+ \\
& \left[2\alpha^2(1-\alpha)(1-\mathsf{a})+\alpha^3(1-2\mathsf{a})+\alpha(1-\alpha)^2\right]\times \\
& \mathsf{H}_2\left(\frac{\alpha^2(1-\alpha)(1-\mathsf{a})+\alpha^3\mathsf{a}+\alpha(1-\alpha)^2}{2\alpha^2(1-\alpha)(1-\mathsf{a})+\alpha^3(1-2\mathsf{a})+\alpha(1-\alpha)^2}\right)-\alpha\mathsf{H}_2(\mathsf{p}),
\end{aligned}$$

where, again, $\mathsf{a}=\mathsf{p}(1-\mathsf{p})$.

The *zero error capacity* of a noisy channel [31] is defined as the least upper bound of rates at which it is possible to transmit information with zero probability of error. Iyengar *et al.* proved that the zero error capacity of the discrete-output binary-input channel described by (1.1) is 0.5; this result is, in fact, an asymptotic version of Theorem 1.5.

1.3 Flash memories

1.3.1 WOM codes

Define an $\langle n,k,t\rangle$ *WOM code* as a coding scheme that uses n (write-only) memory cells to represent information of k bits that can be rewritten a total of t times. The naive way to write a 2-bit value twice into a WOM is by means of 4 memory cells (the first value is written into the first two cells and the second value — into the last two cells), thereby implying the existence of a $\langle 4,2,2\rangle$ WOM code. Rivest and Shamir showed [28, Lemma 1] a coding scheme that achieves the same goal with only 3 memory cells and proved [28, Lemma 4] that the obtained $\langle 3,2,2\rangle$ WOM code is optimal (that is, uses the minimal number of memory cells to "rewrite 2 bits twice").

Cohen *et al.* [5] and Godlewski [10] employed *coset coding* based on error-correcting codes to obtain many rewrites on a WOM. Let $\mathcal{C}$ be an $[n,k,d]$ binary linear code and let H be its parity-check matrix. The information word $\mathbf{u}$ of length $n-k$ is regarded as a syndrome (with respect to H), and the writing into the WOM is carried out by storing a *coset leader* $\mathbf{y}$ whose syndrome $H\mathbf{y}^\top$ equals $\mathbf{u}$. Interpreting the data $\mathbf{y}$ written to the WOM boils down to calculating its syndrome $H\mathbf{y}^\top$. To obtain multiple rewrites, the original code is shortened on the coordinates corresponding to the memory cells whose values were set to 1 in the previous runs. In this manner, a $\langle 23,11,3\rangle$ WOM code and $\left\langle 2^m-1,m,2^{m-2}+2^{m-4}+1\right\rangle$ WOM codes (for any $m\geq 4$) were obtained based on the well-known binary $[23,12,7]$ Golay code [22, Sec. 2.6] and the $[2^m-1,2^m-m-1,3]$ Hamming codes [22, Sec. 1.7], respectively. Since finding a coset leader necessitates a complete decoding algorithm for the code $\mathcal{C}$, it is crucial for the efficient operation of the WOM code that there exists a fast decoding algorithm for $\mathcal{C}$ and its shortened versions.

Zémor and Cohen [40] considered a broader problem of constructing WOM codes that are capable of error correction. Using two- and three-error-correcting BCH codes [22, Ch. 9], [30, Sec. 5.6, Sec. 8.4], they showed the existence of two families of single-error-correcting WOM codes for any $m \geq 3$:

$$\langle n = 2^m - 1, m, n/15.42 + o(n) \rangle \text{ and } \left\langle n = 2^{2m-1}, 2m, n/26.9 + o(n) \right\rangle . \quad (1.2)$$

Yaakobi *et al.* [38], based on their own construction of single-error-detecting WOM codes and the triple-error-correcting BCH-like codes due to Kasami [17] (also see [4]), constructed single-, double- and triple-error-correcting WOM codes that outperformed the codes of [40] in terms of the *WOM rate* kt/n. Specifically, the construction by Yaakobi *et al.* yielded [38, Ex. 2] a single-error-correcting WOM code of WOM rate

$$\frac{k(5 \cdot 2^{k-4} + 1)}{2 \cdot 2^k - 1} \geq \frac{\log_2(3.2(t-1))}{6.4},$$

which is an improvement on the WOM rates of WOM codes in (1.2).

The definition of a WOM code and a WOM rate slightly change, when the number of values stored in a WOM is allowed to be different on each write. Define an $\left\langle n, (M_i)_{i \in [t]}, t \right\rangle$ WOM code as a coding scheme using n (write-only) memory cells to represent an information that can be rewritten t times and the number of values that one can store on the i-th write is M_i for $i \in [t]$. The WOM rate of such a code is defined to be

$$\frac{\sum_{i \in [t]} \log_2 M_i}{n}.$$

The maximum achievable WOM rate of an $\left\langle n, (M_i)_{i \in [t]}, t \right\rangle$ WOM code was proved [8], [11] to be $\log_2(t+1)$, yet there is still a gap between this theoretical bound and the best known WOM rates of $\left\langle n, (M_i)_{i \in [t]}, t \right\rangle$ WOM codes. Even for $t = 2$, although it has been proved by Yaakobi *et al.* [37, Th. 4] that there exists a family of WOM codes whose WOM rate attains $\log_2 3 \approx 1.58$ as $n \to \infty$, the best known WOM rate of a specific $\langle n, M_0, M_1, 2 \rangle$ WOM code is 1.4928. This result is due to the linear code-based construction by the same authors [37, Sec. 3], an adaptation of which also yields a specific $\langle n, k, 2 \rangle$ WOM code of the WOM rate $2k/n \approx 1.4546$ that is short of the theoretical upper bound of ≈ 1.5458 due to [28, Sec. 4A].

1.3.2 Flash codes

Flash codes (or *floating codes*), first introduced in [14], can be viewed as a generalization of WOM codes to multilevel (q-ary flash) memory cells. An $\langle n, k, t\rangle_q$ flash code is a coding scheme for storing k bits in n memory cells with q levels, each enabling up to t rewrites between two consecutive erasures. Define the *write deficiency* of a flash code $\mathcal{C}$, denoted by $\mathsf{wd}(\mathcal{C})$, as $\mathsf{wd}(\mathcal{C}) = n(q-1)-t$; in other words, the write deficiency measures how efficient the code $\mathcal{C}$ is in terms of utilizing the available transitions between two consecutive erasures. Jiang *et al.* have found [14, Th. 2] a general lower bound (the best known lower bound up to date) on $\mathsf{wd}(\mathcal{C})$ for an $\langle n, k, t\rangle_q$ flash code $\mathcal{C}$:

$$\mathsf{wd}(\mathcal{C}) \geq \Omega(qk).$$

Yaakobi *et al.* [39] and later Mahdavifar *et al.* [23] reported upper bounds on the write deficiency which do not depend on the value of n. Specifically, the best known upper bound on $\mathsf{wd}(\mathcal{C})$ for an $\langle n, k, t\rangle_q$ flash code $\mathcal{C}$ to date is due to [23, Th. 3]:

$$\mathsf{wd}(\mathcal{C}) \leq O(\max\{q, \log k\}\, k \log k).$$

1.3.3 Rank-modulation

Recently, a new scheme for storing data in flash memories, called the *rank-modulation scheme*, was proposed in [15] and [16]. The *rank* of a memory cell in the flash memory is defined as the relative position of its charge level among the charge levels of all the memory cells, hence the ranks of n memory cells induce the symmetric group $\mathsf{Sym}(n)$ of permutations on $[n]$ which are used to store information. The main goal in the proposed scheme is to eliminate the risk of overprogramming and reducing the effect of asymmetric errors caused by the rigid setting of charge levels in memory cells. Jiang *et al.* considered the problem of error correction in this new scheme and constructed [16, Constr. 8] a family of single-error-correcting codes of size $0.5(n-1)!$ at least, that is, at least half the feasible maximum.

Arguably, the most natural metric on $\mathsf{Sym}(n)$ is the *Kendall tau distance* which is defined as the minimum number of transpositions of pairwise adjacent elements to turn one permutation into another. Let $\mathtt{M}(n, d)$ be the maximum size of the code over $\mathsf{Sym}(n)$ with (Kendall tau) distance at least d between any two permutations. Define the capacity of codes of distance d as

$$\mathtt{C}_{\text{rank}}(d) = \lim_{n\to\infty} \frac{\ln \mathtt{M}(n, d)}{\ln n!} \quad .$$

Barg and Mazumdar proved [3, Th. 3.1] that

$$\mathsf{C}_{\mathrm{rank}}(d) = \begin{cases} 1 & d = O(n) \\ 1-\epsilon & d = \Theta(n^{1+\epsilon}) \\ 0 & d = \Theta(n^2) \end{cases} .$$

1.3.4 Write-efficient memories

Associating a cost to transitions instead of disallowing some of them altogether gives rise to a different model called write-efficient memories (WEM, in short) coined by Ahlswede and Zhang [2]. In this model, just like in flash memories, the cell state is allowed to take on any of the values from $[q]$, and the cost assigned to the transition from state x to state x' is denoted by $\kappa(x' \mid x)$. Let X and X' be random variables representing the old and new contents of a memory cell, $\mathsf{P}_{X,X'}$ denote their joint probability distribution, P_X and $\mathsf{P}_{X'}$ be the marginal probability distributions of X and X', respectively, and $\mathsf{E}_{\mathsf{P}_{X,X'}}\{\cdot\}$ be the expectation operator under the probability distribution $\mathsf{P}_{X,X'}$. The information-theoretic result by Ahlswede and Zhang [2, Th. 1] states that the storage capacity $\mathsf{C}_{\mathrm{st}}(\xi)$ of the WEM, subject to average cost ξ, is given by

$$\mathsf{C}_{\mathrm{st}}(\xi) = \max_{\substack{\mathsf{P}_{X,X'}\colon \mathsf{P}_X = \mathsf{P}_{X'},\\ \mathsf{E}_{\mathsf{P}_{X,X'}}\{\kappa(X'|X)\} \le \xi}} \mathsf{H}_q(X' \mid X), \tag{1.3}$$

where $\mathsf{H}_q(X' \mid X) = -\sum_{x\in[q]}\sum_{x'\in[q]} \mathsf{P}_{X,X'}(x,x') \log_q \frac{\mathsf{P}_{X,X'}(x,x')}{\sum_{x\in[q]}\mathsf{P}_X(x)}$. This result bears a resemblance to Lemma 2.1 that we are about to see in Section 2.1.

Lastras-Montaño *et al.* [20] focused on the cost matrix

$$\kappa(x' \mid x) = \begin{cases} 1 & x' < x \\ 0 & \text{otherwise} \end{cases} ,$$

which assigned a cost of "one" to all the transitions to lower values and a cost of "zero" to the rest. In this case, the closed-form solution to (1.3) is

$$\mathsf{C}_{\mathrm{st}}(\xi) = \log \frac{1-\alpha^q}{1-\alpha} - \alpha \log \alpha \frac{1-q\alpha^{q-1}+(q-1)\alpha^q}{(1-\alpha)(1-\alpha^q)},$$

where $\alpha = \alpha(\xi) \in (0,1)$ is a parameter arising from the method of Lagrange multipliers. In the same paper [20, Sec. 4], codes (waterfall, multicell and hypercell) with low implementation complexity and good performance were presented, although there is a gap between their rates and $\mathsf{C}_{\mathrm{st}}(\xi)$ (see [20, Fig. 3] and [20, Fig. 4]).

Chapter 2

Preliminary results

This chapter is organized as follows. In Section 2.1, we list the definitions and the known results which will be of use later in the chapter. In Section 2.2, we compute asymptotic Gilbert–Varshamov-like lower bounds on $R_q(\tau)$ for different values of q, using the tools mentioned in Section 2.1. In Section 2.3, we find an upper bound on $M_2(n, t)$ using a general technique from [1]. In Section 2.4, we present constructions of binary t-grain-correcting codes of length n for some values of n and t and show the optimality and the uniqueness of some of these codes. In Section 2.5, we present constructions of ∞-grain-detecting codes of rates close to 1.

2.1 Useful tools

2.1.1 Markov chains

Let $G = (V_G, E_G)$ be a directed graph. Let $P : E_G \to [0, 1]$ be a probability distribution on E_G where $P(e)$ denotes the probability to traverse the edge $e \in E_G$ during a random walk on G. In other words, $P(e) \geq 0$ for $e \in E_G$ and $\sum_{e \in E_G} P(e) = 1$. A *stationary Markov chain* is a probability distribution P on E_G such that

$$\sum_{e=(v',v) \in E_G} P(e) = \sum_{e=(v,v') \in E_G} P(e)$$

for any $v \in V_G$; in other words, in stationary Markov chains there is a preservation of the probability flow: the probability that flows into a state is equal to the probability that flows out of the state. The (stationary) probability $\pi(v)$ to be in a state $v \in V_G$ in a random walk on G according to a stationary Markov chain P is

$$\pi(v) = \sum_{e=(v,v')\in E_G} P(e).$$

For a stationary Markov chain $P : E_G \to [0,1]$, a positive integer k and a vector function $f : E_G \to \mathbb{R}^k$, denote by $\mathsf{E}_P\{f\}$ the expected value of f with respect to P, that is,

$$\mathsf{E}_P\{f\} = \sum_{e\in E_G} P(e)f(e).$$

Finally, the *entropy rate* of a stationary Markov chain P is defined as[1]

$$\mathsf{H}_q(P) = -\sum_{v\in V_G}\sum_{e=(v,v')\in E_G} P(e)\log_q \frac{P(e)}{\pi(v)}\ .$$

2.1.2 Optimizing concave functions

We proceed by citing special cases of [24, Lemma 2] and [24, Lemma 5] which are consequences of well-known results on optimizing convex (concave) functions subject to linear equality and linear inequality constraints (also see [7, Lemma 2], [21, pp. 312–316], [27, Ch. 2, Th. 25] and [29, Sec. 28]). In both lemmas, $\mathcal{M}(f;U)$ denotes the set of all stationary Markov chains P on a graph G such that $\mathsf{E}_P\{f\} \in U \subseteq \mathbb{R}^k$.

Lemma 2.1. Let $G = (V_G, E_G)$ be a primitive graph and $f : E_G \to \mathbb{R}^k$ be a vector function. Let U be an open rectangular parallelepiped $\prod_{i\in[k]}(\widetilde{s}_i, s_i)$ and let Γ_n be a set of all cycles of length n in G. Then

$$\lim_{n\to\infty}\frac{1}{n}\log_q\left|\left\{\gamma\in\Gamma_n : \mathsf{E}_{\mathsf{P}_\gamma}\{f\}\in U\right\}\right| = \sup_{P\in\mathcal{M}(f;U)}\mathsf{H}_q(P).$$

□

Let k be a positive integer. For a graph $G = (V_G, E_G)$, a vector of indeterminates $\mathbf{z} = (z_i)_{i\in[k]} \in (0,\infty)^k$, and a function $f = (f_i)_{i\in[k]} : E_G \to \mathbb{R}^k$, define the matrix function $\mathsf{A}_G(\mathbf{z}) : (0,\infty)^k \to \mathbb{R}^{|V_G|\times|V_G|}$ (whose rows and columns are indexed by V_G) as

$$[\mathsf{A}_G(\mathbf{z})]_{v,v'\in V_G} = \begin{cases} \mathbf{z}^{f(e)} = \prod_{i\in[k]} z_i^{f_i(e)} & e\in E_G \\ 0 & \text{otherwise} \end{cases} \tag{2.1}$$

where $e = (v, v')$.

[1]In this definition we assume that $0\log_q 0 \triangleq 0$.

Lemma 2.2. Let $k, k' \in \mathbb{Z}^+$. Let $\mathbf{p} = (p_i)_{i\in[k']} \in [0,1]^{k'}$ and let $G = (V_G, E_G)$ be a graph. Let $f : E_G \to \mathbb{R}^k$ and $f' : E_G \to \mathbb{R}^{k'}$ be vector functions. Let U be a closed rectangular parallelepiped $\prod_{i\in[k]} [0, s_i]$. Then

$$\sup_{\substack{P\in\mathcal{M}(f;U)\\ \mathsf{E}_P\{f'\}=\mathbf{p}}} \mathsf{H}_q(P) = \inf_{\mathbf{z},\mathbf{h}} \{\log_q \lambda(\mathsf{A}_G(\mathbf{z}, \mathbf{h})) - \sum_{i\in[k]} s_i \log_q z_i - \sum_{i\in[k']} p_i \log_q h_i\},$$

where $\lambda(\cdot)$ denotes the spectral radius of a square real matrix, $\mathbf{z} = (z_i)_{i\in[k]}$ ranges over $(0,1]^k$ and $\mathbf{h} = (h_i)_{i\in[k']}$ ranges over $(0,\infty)^{k'}$. □

2.2 Gilbert–Varshamov-like bounds

For a subset $\mathcal{X} \subseteq \Sigma^n$, let $\mathsf{W}_t(\mathcal{X}) = \sum_{\mathbf{x}\in\mathcal{X}} |\mathcal{R}_t(\mathbf{x}) \cap \mathcal{X}|$. Namely, $\mathsf{W}_t(\mathcal{X})$ is the number of ordered pairs of t-confusable words in $\mathcal{X}$. The following lemma is essentially a reformulation of [19, Lemma 1] for grain-correcting codes.

Lemma 2.3. Let n, t be positive integers and let $\mathcal{X} \subseteq \Sigma^n$. Then

$$M_q(n,t) \geq \frac{|\mathcal{X}|^2}{4\mathsf{W}_t(\mathcal{X})}. \tag{2.2}$$

Proof. Let $\overline{\mathcal{R}}_t(\mathcal{X}) \triangleq \mathsf{W}_t(\mathcal{X})/|\mathcal{X}|$ for any positive integer t. At most half of the risk sets of radius t centered at a word from $\mathcal{X}$ have size greater than $2\overline{\mathcal{R}}_t(\mathcal{X})$, because otherwise the average risk set size exceeds $\overline{\mathcal{R}}_t(\mathcal{X})$. Therefore, there are at least $|\mathcal{X}|/2$ words in $\mathcal{X}$ whose risk sets of radius t contain at most $2\overline{\mathcal{R}}_t(\mathcal{X})$ words. Denote this set of words by $\mathcal{X}'$. For any $\mathbf{x} \in \mathcal{X}'$, $|\mathcal{R}_t(\mathbf{x}) \cap \mathcal{X}'| \leq 2\overline{\mathcal{R}}_t(\mathcal{X})$. Therefore, by picking up words of $\mathcal{X}'$ and erasing their risk sets of radius t in what is left of $\mathcal{X}'$, we can construct a t-grain-correcting code of size $|\mathcal{X}'|/(2\overline{\mathcal{R}}_t(\mathcal{X})) \geq |\mathcal{X}|/(4\overline{\mathcal{R}}_t(\mathcal{X}))$. □

Remark 2.1. We can get rid of the factor 4 in the denominator of the right-hand side of (2.2) by following the proof in [33], but it will have no effect on the asymptotical analysis we are about to do. □

2.2.1 General alphabet size

The forthcoming discussion is meant to evaluate $\mathsf{W}_t(\mathcal{X})$ for certain sets $\mathcal{X}$ of words with prescribed empirical distribution of transitions.

Define graphs $\mathcal{G}^{(\mathcal{N})} = (V^{(\mathcal{N})}, E^{(\mathcal{N})})$, $\mathcal{G}^{(\mathcal{O})} = (V^{(\mathcal{O})}, E^{(\mathcal{O})})$ corresponding to the scenarios without and with overlaps, respectively, as follows. Let $\overline{\Sigma} = \{\overline{a} : a \in \Sigma\}$ be a set where every element $\overline{a}$ designates a symbol whose original value $a \in \Sigma$ was smeared by a grain error. The set of states $V^{(\mathcal{N})} \subseteq (\Sigma \cup \overline{\Sigma})^2$ is defined as $V^{(\mathcal{N})} = V_0 \cup V_1 \cup V_2$ where $V_0 = \{aa : a \in \Sigma\}$, $V_1 = \left\{a\overline{b} : ab \in \Sigma^2,\ a \neq b\right\} \cup \left\{\overline{a}b : ab \in \Sigma^2,\ a \neq b\right\}$, and $V_2 = \left\{\overline{a}\overline{b} : ab \in \Sigma^2,\ a \neq b\right\}$. The set of states $V^{(\mathcal{O})} \subseteq \Sigma^2$ is defined as $V^{(\mathcal{O})} = V_0 \cup V_3$ where $V_3 = \left\{ab : ab \in \Sigma^2,\ a \neq b\right\}$. Specifically, for $q = 2$ (which will be our running example throughout most of this section), $V_0 = \{00, 11\}$, $V_1 = \left\{\overline{0}1, 0\overline{1}, \overline{1}0, 1\overline{0}\right\}$, $V_2 = \left\{\overline{0}\overline{1}, \overline{1}\overline{0}\right\}$, and $V_3 = \{01, 10\}$ (the states of the set V_2 will have no incoming edges for $q = 2$ in $\mathcal{G}^{(\mathcal{N})}$, so for $q = 2$, we will disregard V_2 altogether). For $\mathbf{x} = (x_i)_{i\in[n]} \in (\Sigma \cup \overline{\Sigma})^n$, define the operator $\partial(\mathbf{x}) = (\partial(x_i))_{i\in[n]}$ such that $\partial(a) = \partial(\overline{a}) = a$ for every $a \in \Sigma$.

There is an edge in $E^{(\mathcal{N})}$ from $v = \ell r$ to $v' = \ell' r'$ if

[N1] $v' \in V_0$, or

[N2] $\ell = \ell' \in \Sigma$ and $rr' \in \Sigma\overline{\Sigma}$, or $\ell\ell' \in \Sigma\overline{\Sigma}$ and $r = r' \in \Sigma$, or

[N3] $\ell = r' \in \Sigma$ and $\ell' r \in \overline{\Sigma}^2$, or $\ell' = r \in \Sigma$ and $\ell r' \in \overline{\Sigma}^2$, or

[N4] $v \in V_0$, $v' \in V_2$, $\ell \neq \partial(\ell')$, $r \neq \partial(r')$.

There is an edge in $E^{(\mathcal{O})}$ from $v = \ell r$ to $v' = \ell' r'$ if

[O1] $v' \in V_0$, or

[O2] $v \in V_0$ and $v' \in V_3$, or

[O3] $v, v' \in V_3$, $\ell r \neq r'\ell'$ and either $\ell = r'$ or $r = \ell'$, or

[O4] $v, v' \in V_3$ and $\ell r = r'\ell'$.

Given a path[2] $\gamma = (\ell_i r_i)_{i\in[n]}$ of length $n-1$ in $\mathcal{G}^{(\mathcal{N})}$, define the sets $\beta_L(\gamma) = \left\{i : \ell_i \in \overline{\Sigma}\right\}$ and $\beta_R(\gamma) = \left\{i : r_i \in \overline{\Sigma}\right\}$. When the path γ is in $\mathcal{G}^{(\mathcal{O})}$, let $\beta_L(\gamma) = \{i : \ell_i \neq r_i,\ r_{i-1} \neq \ell_i\}$, $\beta_R(\gamma) = \{i : \ell_i \neq r_i,\ \ell_{i-1} \neq r_i\}$. In addition, for an edge $e \in (\ell r, \ell' r')$ in $\mathcal{G}^{(\mathcal{O})}$, define the function $\mu : E^{(\mathcal{O})} \to [2]$ that equals 1 if e satisfies Condition [O4] and 0 otherwise. Define

$$\mu(\gamma) = \sum_{i\in[n-1]} \mu(\ell_i r_i, \ell_{i+1} r_{i+1})$$

for a path $\gamma = (\ell_i r_i)_{i\in[n]}$ in $\mathcal{G}^{(\mathcal{O})}$. Note that $\mu(\gamma)$ equals the number of locations where (overlapping) grains, if switched from $\mathbf{x} = (\ell_i)_{i\in[n]}$ to the corresponding index of $\mathbf{y} = (r_i)_{i\in[n]}$ (and vice versa), will still confuse $\mathbf{x}$ and $\mathbf{y}$. For completeness, let $\mu(\gamma) = 0$ when γ is in $\mathcal{G}^{(\mathcal{N})}$.

[2]Since there are no parallel edges in $\mathcal{G}^{(\mathcal{N})}$ and $\mathcal{G}^{(\mathcal{O})}$, we can represent paths as sequences of states.

The path γ in $\mathcal{G}^{(\mathcal{N})}$ starting in V_0 represents a pair

$$\big(\mathbf{x} = (\partial(\ell_i))_{i\in[n]}, \mathbf{y} = (\partial(r_i))_{i\in[n]}\big)$$

of finitely-confusable words in Σ^n, as well as grain patterns $\mathcal{S} = \beta_L(\gamma), \mathcal{S}' = \beta_R(\gamma)$ that cause the corresponding smeared words, $\sigma_{\mathcal{S}}(\mathbf{x})$ and $\sigma_{\mathcal{S}'}(\mathbf{y})$, to be equal. As for the path γ in $\mathcal{G}^{(\mathcal{O})}$ starting in V_0, it represents a pair $(\mathbf{x} = (\ell_i)_{i\in[n]}, \mathbf{y} = (r_i)_{i\in[n]})$ of finitely-confusable words in Σ^n and $2^{\mu(\gamma)}$ confusing grain patterns $\mathcal{S} = \beta_L(\gamma) \cup \mu_L(\gamma), \mathcal{S}' = \beta_R(\gamma) \cup \mu_R(\gamma)$ where $\{\mu_L(\gamma), \mu_R(\gamma)\}$ is a partition of edges of γ that satisfy Condition [O4].

Example 2.1. For $\Sigma = [3]$ and $n = 6$, consider the path

$$\gamma = (v_i)_{i\in[n]} = 11 \quad 22 \quad 2\bar{0} \quad \bar{1}2 \quad 00 \quad \bar{1}\bar{2}$$

in $\mathcal{G}^{(\mathcal{N})}$. This path corresponds to the smeared words $122\bar{1}0\bar{1}$ and $12\bar{0}20\bar{2}$ (the grain-free words are $\mathbf{x} = 122101$ and $\mathbf{y} = 120202$) and the overlines indicate the grain patterns $\mathcal{S} = \beta_L(\gamma) = \{3, 5\}, \mathcal{S}' = \beta_R(\gamma) = \{2, 5\}$ that make $\mathbf{x}$ and $\mathbf{y}$ confusable. The edges (v_i, v_{i+1}) for $i = 0, 1, 2$ correspond to Conditions [N1]–[N3], respectively; the edge (v_4, v_5) corresponds to Condition [N4]. Now, for the same alphabet Σ and $n = 7$, consider the path

$$\gamma = (v_i)_{i\in[n]} = 11 \quad 22 \quad 20 \quad 12 \quad 00 \quad 12 \quad 21$$

in $\mathcal{G}^{(\mathcal{O})}$. Here $\beta_L(\gamma) = \{3, 5\}$, $\beta_R(\gamma) = \{2, 5\}$ and $\mu(\gamma) = 1$. The edges (v_i, v_{i+1}) for $i = 0, 1, 2$ correspond to Conditions [O1]–[O3], respectively; the edge (v_5, v_6) corresponds to Condition [O4]. □

For $q = 2$, the adjacency matrices $A_{\mathcal{G}}^{(j)}$ of $\mathcal{G}^{(j)}$ for $j \in \{\mathcal{N}, \mathcal{O}\}$, constructed as described above, are shown in Figure 2.1.

$A_{\mathcal{G}}^{(\mathcal{N})}$	00	$\bar{0}1$	$0\bar{1}$	$\bar{1}0$	$1\bar{0}$	11
00	1	0	1	1	0	1
$\bar{0}1$	1	0	0	0	1	1
$0\bar{1}$	1	0	0	1	0	1
$\bar{1}0$	1	0	1	0	0	1
$1\bar{0}$	1	1	0	0	0	1
11	1	1	0	0	1	1

and

$A_{\mathcal{G}}^{(\mathcal{O})}$	00	01	10	11
00	1	1	1	1
01	1	0	1	1
10	1	1	0	1
11	1	1	1	1

Figure 2.1: Adjacency matrices $A_{\mathcal{G}}^{(\mathcal{N})}$ and $A_{\mathcal{G}}^{(\mathcal{O})}$ for $q = 2$.

To make the presentation and the computation simpler, we will switch to a different criterion of confusability till the end of this section. Given a positive integer t, we will call two words $\mathbf{x}, \mathbf{y}$ *t-confusable in a wider sense* (or *t-cws*, in short) if there exist grain patterns $\mathcal{S}$ and $\mathcal{S}'$ such that $|\mathcal{S}| + |\mathcal{S}'| \leq 2t$ and $\sigma_{\mathcal{S}}(\mathbf{x}) = \sigma_{\mathcal{S}'}(\mathbf{y})$. Notice that any t-confusable pair of words is t-cws hence any t-grain-correcting code in a wider sense is also t-grain-correcting in the old sense. Our results will actually apply to the wider-sense notion of confusability.[3] The following lemma (with proof in Appendix A) establishes a correspondence between ordered pairs of t-cws words and paths in $\mathcal{G}^{(\mathcal{N})}$ or $\mathcal{G}^{(\mathcal{O})}$.

Lemma 2.4. For $j \in \{\mathcal{N}, \mathcal{O}\}$, let $\mathcal{W}_t^{(j)}$ denote the set of all t-cws (ordered) pairs $(\mathbf{x}, \mathbf{y}) \in \Sigma^n \times \Sigma^n$ and let $\mathcal{P}_t^{(j)}$ be the following set of paths in $\mathcal{G}^{(j)}$:

$$\mathcal{P}_t^{(j)} = \left\{ \gamma = (v_i)_{i \in [n]} : v_0 \in V_0,\ |\beta_L(\gamma)| + |\beta_R(\gamma)| + \mu(\gamma) \leq 2t \right\}. \tag{2.3}$$

Then for each pair $(\mathbf{x}, \mathbf{y})$ in $\mathcal{W}_t^{(j)}$, there is exactly one path $\gamma = (\ell_i r_i)_{i \in [n]}$ in $\mathcal{P}_t^{(j)}$ such that $\mathbf{x} = (\partial(\ell_i))_{i \in [n]}$ and $\mathbf{y} = (\partial(r_i))_{i \in [n]}$ for $j \in \{\mathcal{N}, \mathcal{O}\}$. $\square$

For $j \in \{\mathcal{N}, \mathcal{O}\}$, let $\Gamma_n^{(j)}$ denote the set of all the cycles in $\mathcal{G}^{(j)}$ of length n that start and terminate in the same state of V_0. Define the functions $f^{(\mathcal{N})} : E^{(\mathcal{N})} \to [3]^2$, $f^{(\mathcal{O})} : E^{(\mathcal{O})} \to [3]^2 \times [2]$ such that for any edge e,

$$f^{(\mathcal{N})}(e) = (\nu(e), \chi(e)) \text{ and } f^{(\mathcal{O})}(e) = (\omega(e), \chi(e), \mu(e))$$

where the function μ is defined before Example 2.1 and the functions $\nu : E^{(\mathcal{N})} \to [3]$, $\omega : E^{(\mathcal{O})} \to [3]$, $\chi : E^{(\mathcal{N})} \cup E^{(\mathcal{O})} \to [3]$ are defined next. For an edge $e = (\ell r, \ell' r')$:

- $\nu(e)$ equals 2 if e satisfies Condition [N4], 1 if e satisfies Conditions [N2] or [N3], and 0 otherwise.
- $\omega(e)$ equals 2 if e satisfies Condition [O2] for $\ell = r \notin \{\ell', r'\}$, 1 if e satisfies either Condition [O2] for $\ell = r \in \{\ell', r'\}$ or Condition [O3], and 0 otherwise.
- $\chi(e)$ equals 2 if $\partial(\ell) \neq \partial(\ell')$ and $\partial(r) \neq \partial(r')$, 0 if $\partial(\ell) = \partial(\ell')$ and $\partial(r) = \partial(r')$, and 1 otherwise.

The function $\nu(e)$ counts the smallest number of grains making $\ell\ell'$ and rr' confusable for $j = \mathcal{N}$. The function $\omega(e)$ counts the smallest number of

[3] Though we do not currently have a general proof for this phenomenon, our empirical results show that the relaxation of the notion of confusability does not result in worse bounds while using our technique.

nonoverlapping grains making $\ell\ell'$ and rr' confusable (these grains, unlike those counted by function $\mu(\cdot)$, cannot be interchanged arbitrarily) for $j = \mathcal{O}$. The function $\chi(e)$ counts the number of crossovers in $\ell\ell'$ or rr', namely, we add 1 if $\partial(\ell) \neq \partial(\ell')$ and another 1 if $\partial(r) \neq \partial(r')$.

For a cycle $\gamma = (v_i)_{i\in[n+1]}$ (where $v_0 = v_n$) of length n in $\mathcal{G}$, let $\mathsf{P}_\gamma : E \to [0,1]$ be the *empirical probability distribution* of γ, namely, for $e \in E$,

$$\mathsf{P}_\gamma(e) = \frac{1}{n}|\{i \in [n] : (v_i, v_{i+1}) = e\}|.$$

Now, set $\tau, p \in (0,1)$, let $\epsilon > 0$ and define

$$U^{(\mathcal{N})}_{\tau,p,\epsilon} = \{(u_1, u_2) : -\epsilon < u_1 < 2\tau + \epsilon, |u_2 - 2p| < 2\epsilon\}$$

and

$$U^{(\mathcal{O})}_{\tau,p,\epsilon} = \{(u_1, u_2, u_3) : u_1, u_3 > -\epsilon,\ u_1 + u_3 < 2\tau + \epsilon,\ |u_2 - 2p| < 2\epsilon\}.$$

Sets $U^{(\mathcal{N})}_{\tau,p,\epsilon}$ and $U^{(\mathcal{O})}_{\tau,p,\epsilon}$ contain all the values of parameters we are interested in, namely, the ratio of the number of grains of length 2 to the word length n, controlled by τ, and the ratio of the number of crossovers to the word length n, controlled by p.

For $j \in \{\mathcal{N}, \mathcal{O}\}$, let

$$\Gamma^{(j)}_{\tau,p,\epsilon} = \left\{\gamma \in \Gamma^{(j)}_n : \mathsf{E}_{\mathsf{P}_\gamma}\{f^{(j)}\} \in U^{(j)}_{\tau,p,\epsilon}\right\}.$$

The set $\Gamma^{(j)}_{\tau,p,\epsilon}$ for $j \in \{\mathcal{N}, \mathcal{O}\}$ stands for all the cycles of length n of $\mathcal{G}^{(j)}$ representing pairs of words $(\mathbf{x}, \mathbf{y})$ that can be confused by at most $\sim 2\tau$ grains (either overlapping or not, depending on the context) and whose collective ratio of crossovers is $\sim 2p$. Additionally, for $j \in \{\mathcal{N}, \mathcal{O}\}$ and the same τ, p, ϵ, let

$$\mathcal{P}^{(j)}_{\tau,p,\epsilon} = \left\{\gamma \in \mathcal{P}^{(j)}_{\lceil\tau(n-1)\rceil} : \left|\mathsf{E}_{\mathsf{P}_\gamma}\{\chi\} - 2p\right| \leq \epsilon\right\}.$$

The set $\mathcal{P}^{(j)}_{\tau,p,\epsilon}$ possesses properties similar to those of the set $\Gamma^{(j)}_{\tau,p,\epsilon}$. The following lemma characterizes the relation between the sizes of these two sets for sufficiently large values of n.

Lemma 2.5. Let $\tau, p \in (0,1)$ and $\epsilon > 0$. Then

$$\left|\mathcal{P}^{(j)}_{\tau,p,\epsilon}\right| \leq \left|\Gamma^{(j)}_{\tau,p,\epsilon}\right|$$

for $j \in \{\mathcal{N}, \mathcal{O}\}$.

Proof. We will prove for the case without overlaps; when overlaps are allowed, the proof is similar. For a path $\gamma \in \mathcal{P}^{(\mathcal{N})}_{\tau,p,\epsilon}$, one has

$$|\beta_L(\gamma)| + |\beta_R(\gamma)| \le 2\lceil \tau(n-1) \rceil \text{ and } \left|\mathsf{E}_{\mathsf{P}_\gamma}\left\{\chi\right\} - 2p\right| \le \epsilon.$$

We can draw an edge from the last state of γ to the first one (by the construction, there is an edge to a state of V_0 from any state of $\mathcal{G}^{(\mathcal{N})}$) to create a cycle γ' of length n. Since

$$|\beta_L(\gamma)| + |\beta_R(\gamma)| = \left|\beta_L(\gamma')\right| + \left|\beta_R(\gamma')\right|,$$

we have $\mathsf{E}_{\mathsf{P}_{\gamma'}}\left\{f^{(\mathcal{N})}\right\} \in [0, 2\tau] \times [2p-2\epsilon, 2p+2\epsilon]$ for sufficiently large n. Hence $\left|\mathcal{P}^{(\mathcal{N})}_{\tau,p,\epsilon}\right| \le \left|\Gamma^{(\mathcal{N})}_{\tau,p,\epsilon}\right|$. □

For $z \in (0,1]$ and $h, m \in (0,\infty)$, let the matrix functions $\mathsf{A}^{(\mathcal{N})}_{\mathcal{G}}(z,h)$ and $\mathsf{A}^{(\mathcal{O})}_{\mathcal{G}}(z,h,m)$, with rows and columns indexed by the sets of states $V^{(\mathcal{N})}$ and $V^{(\mathcal{O})}$, respectively, be defined as a special case of (2.1):

$$\left[\mathsf{A}^{(\mathcal{N})}_{\mathcal{G}}(z,h)\right]_{v,v'\in V} = \begin{cases} z^{\nu(e)}h^{\chi(e)} & e = (v,v') \in E^{(\mathcal{N})} \\ 0 & \text{otherwise} \end{cases}$$

and

$$\left[\mathsf{A}^{(\mathcal{O})}_{\mathcal{G}}(z,h,m)\right]_{v,v'\in V} = \begin{cases} z^{\omega(e)}h^{\chi(e)}m^{\mu(e)} & e = (v,v') \in E^{(\mathcal{O})} \\ 0 & \text{otherwise} \end{cases}.$$

Applying Lemma 2.1 with $G = \mathcal{G}^{(\mathcal{N})}$, $U = U^{(\mathcal{N})}_{\tau,p,\epsilon}$ and $f = f^{(\mathcal{N})}$ and combining the result with Lemma 2.5, we conclude that

$$\lim_{n\to\infty} \frac{1}{n} \log_q \left|\mathcal{P}^{(\mathcal{N})}_{\tau,p,\epsilon}\right| \le \sup_{P \in \mathcal{M}(f^{(\mathcal{N})}; U^{(\mathcal{N})}_{\tau,p,\epsilon})} \mathsf{H}_q(P).$$

By the continuity of the functions $P \mapsto \mathsf{E}_P(f^{(\mathcal{N})})$ and $P \mapsto \mathsf{H}_q(P)$,

$$\lim_{\epsilon\to 0} \lim_{n\to\infty} \frac{1}{n} \log_q \left|\mathcal{P}^{(\mathcal{N})}_{\tau,p,\epsilon}\right| \le \sup_{P \in \mathcal{M}(f^{(\mathcal{N})}; U^{(\mathcal{N})}_{\tau,p})} \mathsf{H}_q(P)$$

where $U^{(\mathcal{N})}_{\tau,p} = \{(u, 2p) : 0 \le u \le 2\tau\}$. Applying Lemma 2.2 with $G = \mathcal{G}^{(\mathcal{N})}$, $f = \nu$, $f' = \chi$, $U = \{u : 0 \le u \le 2\tau\}$ and $\mathbf{p} = (2p)$ yields

$$\lim_{\epsilon\to 0} \lim_{n\to\infty} \frac{1}{n} \log_q \left|\mathcal{P}^{(\mathcal{N})}_{\tau,p,\epsilon}\right| \le K^{(\mathcal{N})} = \inf_{z\in(0,1],\, h\in(0,\infty)} \left\{\log_q \lambda(\mathsf{A}^{(\mathcal{N})}_{\mathcal{G}}(z,h)) - 2\tau \log_q z - 2p \log_q h\right\}. \tag{2.4}$$

Turning now to $j = \mathcal{O}$, we define for $\eta \in [0, 2\tau]$,

$$U^{(\mathcal{O})}_{\tau,p,\epsilon,\eta} = \{(u_1, u_2, u_3) : u_1, u_3 \geq 0,\\ u_1 \leq 2\tau - \eta, |u_3 - \eta| \leq 2\epsilon, |u_2 - 2p| \leq 2\epsilon\}.$$

Since $\mu(\gamma) \in [n+1]$ for any $\gamma \in \Gamma^{(\mathcal{O})}_{\tau,p,\epsilon}$, there is a polynomial (in n) number of types of cycles in $\Gamma^{(\mathcal{O})}_{\tau,p,\epsilon}$ characterized by the same value of $\mathsf{E}_{\mathsf{P}_\gamma}\{\mu\}$, therefore

$$\lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log_q\left|\Gamma^{(\mathcal{O})}_{\tau,p,\epsilon}\right| = \sup_{0\leq\eta\leq 2\tau}\lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log_q\left|\Gamma^{(\mathcal{O})}_{\tau,p,\epsilon,\eta}\right|$$

where

$$\Gamma^{(\mathcal{O})}_{\tau,p,\epsilon,\eta} = \left\{\gamma \in \Gamma^{(\mathcal{O})}_{\tau,p,\epsilon} : \left|\mathsf{E}_{\mathsf{P}_\gamma}\{\mu\} - \eta\right| \leq 2\epsilon\right\}.$$

Applying Lemma 2.1 with $G = \mathcal{G}^{(\mathcal{O})}$, $U = U^{(\mathcal{O})}_{\tau,p,\epsilon,\eta}$ and $f = f^{(\mathcal{O})}$, then applying Lemma 2.2 with $G = \mathcal{G}^{(\mathcal{O})}$, $f = \omega$, $f' = (\chi, \mu)$, $U = \{u : 0 \leq u \leq 2\tau - \eta\}$ and $\mathbf{p} = (2p, \eta)$, and then combining the result with Lemma 2.5 yields

$$\lim_{\epsilon\to 0}\lim_{n\to\infty}\frac{1}{n}\log_q\left|\mathcal{P}^{(\mathcal{O})}_{\tau,p,\epsilon}\right| \leq K^{(\mathcal{O})} = \\ \sup_{0\leq\eta\leq 2\tau}\inf_{z\in(0,1],\, h,m\in(0,\infty)}\{\log_q \lambda(\mathsf{A}^{(\mathcal{O})}_{\mathcal{G}}(z, h, m)) \\ -(2\tau - \eta)\log_q z - \eta\log_q m - 2p\log_q h\}. \tag{2.5}$$

For $j \in \{\mathcal{N}, \mathcal{O}\}$, let $\mathcal{W}^{(j)}_{\tau,p,\epsilon}$ be the set of pairs $(\mathbf{x}, \mathbf{y}) \in \mathcal{W}^{(j)}_{\lceil\tau(n-1)\rceil}$ such that the average number of crossovers in both $\mathbf{x}$ and $\mathbf{y}$ is within $2p\pm\epsilon$. By Lemma 2.4, the inequalities in (2.4) and (2.5) hold if we replace $\mathcal{P}^{(j)}_{\tau,p,\epsilon}$ therein by $\mathcal{W}^{(j)}_{\tau,p,\epsilon}$.

We are now in the position to prove the main theorem of this chapter.

Theorem 2.6. *Let $\tau \in (0, 1)$. Then for $j \in \{\mathcal{N}, \mathcal{O}\}$,*

$$R_q(\tau) \geq \varrho^{(j)}_q(\tau) = \sup_{p\in[0,1]}\left\{2\mathsf{H}_q(p) - K^{(j)}\right\}, \tag{2.6}$$

where $K^{(\mathcal{N})}$ and $K^{(\mathcal{O})}$ are defined in (2.4) and (2.5).

Proof. For a word $\mathbf{x} = (x_i)_{i\in[n+1]} \in \Sigma^{n+1}$ and symbols $a \neq a' \in \Sigma$, let $\kappa_1(\mathbf{x}; a)$ denote the number of locations $i \in [n]$ such that $x_i = a$ and $\kappa_2(\mathbf{x}; a, a')$ denote the number of locations $i \in [n]$ such that $(x_i, x_{i+1}) = (a, a')$. Let $\mathcal{X}$ be a subset of words in Σ^{n+1} such that $\kappa_1(\mathbf{x}; a) = n/q$ and $\kappa_2(\mathbf{x}; a, a') = np/[q(q-1)]$ for any $\mathbf{x} \in \mathcal{X}$ and any $a \neq a' \in \Sigma$.

The exponential growth rate of $\mathcal{X}$ as n approaches infinity is

$$\lim_{n\to\infty}\frac{1}{n}\log_q|\mathcal{X}| = \mathsf{H}_q(p).$$

The inequality (2.6) now follows from the logarithmic version of (2.2), inequalities (2.4) and (2.5) (with $\mathcal{P}^{(j)}_{\tau,p,\epsilon}$ replaced therein by $\mathcal{W}^{(j)}_{\tau,p,\epsilon}$ for $j \in \{\mathcal{N}, \mathcal{O}\}$) and the fact that

$$\lim_{n\to\infty} \frac{1}{n} \log_q \mathsf{W}_{\lceil \tau n \rceil}(\mathcal{X}) = \lim_{\epsilon\to 0} \lim_{n\to\infty} \frac{1}{n} \log_q \left| \mathcal{W}^{(j)}_{\tau,p,\epsilon} \right|$$

for $j \in \{\mathcal{N}, \mathcal{O}\}$. □

To alleviate the computations, we can now merge states in $\mathcal{G}^{(j)}$ to reduce the order of the matrix $\mathsf{A}^{(j)}_{\mathcal{G}}$ while preserving its spectral radius for $j \in \{\mathcal{N}, \mathcal{O}\}$, as described in [25, Sec. 4.6]. This is similar to the standard procedure for reducing the number of states in a presentation of a constrained system $\mathbb{S}$ by merging states whose outgoing paths generate the same sets of words in $\mathbb{S}$.

The states of V_0 can be merged into superstate 0, states of V_1 in $\mathcal{G}^{(\mathcal{N})}$ and states of V_3 in $\mathcal{G}^{(\mathcal{O})}$ — into superstate 1, whereas states of V_2 — into superstate 2. Specifically, for $q = 2$, the merging ends up with reduced matrices $\mathcal{A}^{(\mathcal{N})}_{\mathcal{G}}$ and $\mathcal{A}^{(\mathcal{O})}_{\mathcal{G}}$ whose spectral radii equal those of $\mathsf{A}^{(\mathcal{N})}_{\mathcal{G}}$ and $\mathsf{A}^{(\mathcal{O})}_{\mathcal{G}}$, respectively, as shown in Figure 2.2.

$\mathcal{A}^{(\mathcal{N})}_{\mathcal{G}}$	0	1
0	$1+h^2$	$2hz$
1	$2h$	$h^2 z$

and

$\mathcal{A}^{(\mathcal{O})}_{\mathcal{G}}$	0	1
0	$1+h^2$	$2hz$
1	$2h$	$h^2 m$

Figure 2.2: Reduced matrices $\mathcal{A}^{(\mathcal{N})}_{\mathcal{G}}$ and $\mathcal{A}^{(\mathcal{O})}_{\mathcal{G}}$ for $q = 2$.

It turns out that for $q = 2$, $\varrho^{(\mathcal{N})}_2(\tau) = \varrho^{(\mathcal{O})}_2(\tau)$ for any $\tau \in [0, 1]$. This phenomenon is due to the fact that when $q = 2$, for any path $\gamma' \in \mathcal{P}^{(\mathcal{O})}_t$ that confuses the t-cws pair of words $(\mathbf{x}, \mathbf{y})$ there exists a path $\gamma \in \mathcal{P}^{(\mathcal{N})}_t$ that confuses the same pair of words. The path γ is obtained by moving overlapping grains from $\beta_L(\gamma')$ to $\beta_R(\gamma')$ and vice versa until $\beta_L(\gamma')$ and $\beta_R(\gamma')$ do not contain consecutive numbers.

Lower bound $\varrho^{(\mathcal{N})}_2(\tau) = \varrho^{(\mathcal{O})}_2(\tau)$ of Theorem 2.6 attains its maximum when

$$\tau = \frac{z}{2\sqrt{z^2+12z+4}} \frac{\sqrt{z^2+12z+4}+z+6}{\sqrt{z^2+12z+4}+z+2}.$$

Theorem 2.6 strictly improves on the traditional Gilbert–Varshamov bound (in the Hamming metric)

$$\varrho^{(\mathrm{GV})}_2(\tau) = 1 - \mathsf{H}_2(2\tau)$$

on the entire interval $(0, 0.25]$, however, on the interval $[0.0566, 0.25]$ it falls short of the simple lower bound of 0.5 which is realized by Construction 1.3 from Section 1.1. The difference between $\varrho_2^{(\mathcal{N})}(\tau) = \varrho_2^{(\mathcal{O})}(\tau)$ and $\varrho_2^{(\mathrm{GV})}(\tau)$ on the interval $(0, 0.0566]$ does not exceed 0.012.

For $q > 2$, $\mathcal{A}_{\mathcal{G}}^{(\mathcal{N})}$ and $\mathcal{A}_{\mathcal{G}}^{(\mathcal{O})}$ are, respectively,

	0	1	2
0	$1+(q-1)h^2$	$2(q-1)hz$	$(q-1)(q-2)h^2z^2$
1	$2h+(q-2)h^2$	$(q-1)h^2z$	0
2	$2h+(q-2)h^2$	0	0

and

	0	1
0	$1+(q-1)h^2$	$2(q-1)hz+(q-1)(q-2)h^2z^2$
1	$2h+(q-2)h^2$	$h^2m+2(q-2)h^2z$

.

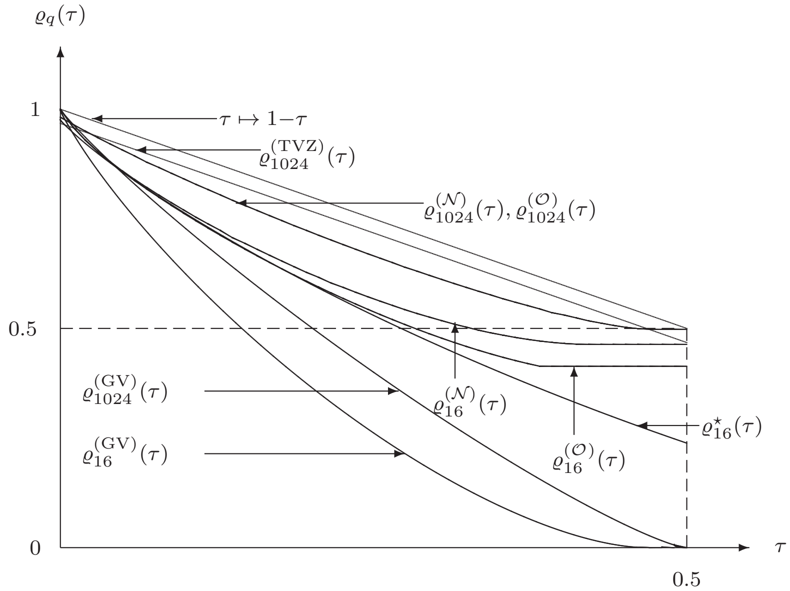

Figure 2.3: Functions $\varrho_q^{(\mathcal{N})}(\tau)$, $\varrho_q^{(\mathcal{O})}(\tau)$ and $\varrho_q^{(\mathrm{GV})}(\tau)$ for $q \in \{16, 1024\}$.

Figure 2.3 depicts the functions $\tau \mapsto \varrho_q^{(\mathcal{N})}(\tau)$ and $\tau \mapsto \varrho_q^{(\mathcal{O})}(\tau)$ for $q \in \{16, 1024\}$ along with the corresponding traditional Gilbert–Varshamov bounds $\varrho_q^{(\mathrm{GV})}(\tau) : \tau \mapsto 1-\mathsf{H}_q(2\tau)$. Both $\varrho_q^{(\mathcal{N})}(\tau)$ and $\varrho_q^{(\mathcal{O})}(\tau)$ strictly improve on $\varrho_q^{(\mathrm{GV})}(\tau)$ on the entire interval $(0, 0.5]$ (and $\varrho_q^{(\mathcal{N})}(\tau)$ is strictly above $\varrho_q^{(\mathcal{O})}(\tau)$); besides, both $\varrho_q^{(\mathcal{N})}(\tau)$ and $\varrho_q^{(\mathcal{O})}(\tau)$ converge to the function $\tau \mapsto 1-\tau$ on that interval when $q \to \infty$. The last statement readily follows from substituting $z = m = 1/\sqrt{q}$, $h = 1$ into (2.4), (2.5) and noticing that $\lambda(\mathcal{A}_{\mathcal{G}}^{(\mathcal{N})}(z,h))$ and $\lambda(\mathcal{A}_{\mathcal{G}}^{(\mathcal{O})}(z,h,m))$ are bounded [9, Ch. 13] by the minimal $(q(1+o(1)))$ and the maximal $(2q(1+o(1)))$ sum of rows in their matrices ($o(1)$ goes to 0 for $q \to \infty$).

For large values of q when overlaps are not allowed, the lower bound $\varrho_q^{(\mathcal{N})}(\tau)$ is worse on nearly the entire interval $(0, 0.5)$ than the following construction based on the family of linear $[n, nR, \lceil \tau n \rceil + 1]$ codes by Tsfasman *et al.* [34] with rate

$$R \geq 1 - \frac{1}{\sqrt{q}-1} - \tau - o(1)$$

where $o(1)$ goes to 0 for $n \to \infty$. By an averaging argument, there exists at least one coset of an $[n, nR, \lceil \tau n \rceil + 1]$ code of this family whose intersection $\mathcal{C}^{(\mathrm{TVZ})}$ with the code

$$\left\{ \mathbf{c} = (c_i)_{i \in [n]} \in \Sigma^n : c_i \neq c_{i+1} \text{ for any } i \in [n-1] \right\} \tag{2.7}$$

is of rate at least $R - 1 + \log_q (q-1)$. Since adjacent symbols in each codeword in $\mathcal{C}^{(\mathrm{TVZ})}$ are different, grain errors become erasures, hence $\mathcal{C}^{(\mathrm{TVZ})}$ is a $\lceil \tau n \rceil$-grain-correcting code of rate at least

$$\varrho_q^{(\mathrm{TVZ})}(\tau) = \log_q (q-1) - \frac{1}{\sqrt{q}-1} - \tau.$$

Remark 2.2. By the same token, one can construct a family of $\lceil \tau n \rceil$-grain-correcting codes of length n and rate at least $\log_q(q/2) - 1/(\sqrt{q}-1) - \tau$ when overlaps are allowed. Specifically, instead of the code in (2.7) one can choose the code $(\Sigma_1 \Sigma_2)^{n/2}$ where $\Sigma_1 = [q/2]$, $\Sigma_2 = [q] \setminus [q/2]$ and both n and q are even. □

A similar reasoning applied to the family of linear codes guaranteed by the Gilbert–Varshamov bound yields $\lceil \tau n \rceil$-grain-correcting codes of rate at least

$$\varrho_q^{\star}(\tau) = \log_q (q-1) - \mathsf{H}_q(\tau).$$

The functions $\tau \mapsto \varrho_{1024}^{(\mathrm{TVZ})}(\tau)$ and $\tau \mapsto \varrho_{16}^{\star}(\tau)$ are shown in Figure 2.3 alongside the other bounds. It can be observed that $\varrho_{16}^{(\mathcal{N})}(\tau)$ and $\varrho_{16}^{(\mathcal{O})}(\tau)$ are strictly above $\varrho_{16}^{\star}(\tau)$, whereas $\varrho_{1024}^{(\mathcal{N})}(\tau)$ is above $\varrho_{1024}^{(\mathrm{TVZ})}(\tau)$ only in the interval $[0, 0.06] \cup [0.44, 0.5]$.

2.2.2 General grain length

We conclude this section by considering the case where grains of any length up to g are allowed and $\Sigma = [2]$. In this case, since the grains are allowed to be of different lengths, it is reasonable to assume that longer grains inflict more errors, specifically, a grain of length $j \in [g+1] \setminus \{0,1\}$ inflicts $j-1$ errors. This case necessitates a change in the definition of a grain pattern. Refine the previous definition (see Section 1.1) of a *grain pattern* as a set $\mathcal{S}(g) \subset [n-1] \times ([n] \setminus \{0\})$ such that any pair $(b,e) \in \mathcal{S}(g)$, denoting the beginning and the end of a grain, satisfies $b < e < b+g$. We say that a granular structure has *overlaps* if there exist two pairs $(b_1,e_1),(b_2,e_2) \in \mathcal{S}(g)$ such that $b_1 \leq e_2$ and $b_2 \leq e_1$. In a *nonoverlapping* granular structure for any two pairs $(b_1,e_1),(b_2,e_2) \in \mathcal{S}(g)$ one has either $e_1 < b_2$ or $e_2 < b_1$. We prohibit grains from being included in one another, viz., we disallow[4] the existence of grains (b_1,e_1), (b_2,e_2) such that $b_1 < b_2 < e_2 < e_1$. A granular pattern $\mathcal{S}(g)$ inflicts errors to a codeword $\mathbf{c} = (c_i)_{i\in[n]}$ over alphabet Σ of size q by means of the smearing operator $\sigma = \sigma(\mathcal{S}(g))$ that yields an output word $\mathbf{y} = (y_i)_{i\in[n]} = \sigma(\mathbf{c})$ over Σ in the following way. For any index $i \in [n]$ if no pair $(b,e) \in \mathcal{S}(g)$ satisfies $b < i \leq e$, then $y_i = c_i$, otherwise $y_i = c_{\mathsf{b}}$ where

$$\mathsf{b} = \arg\max \{b \mid \exists e : (b,e) \in \mathcal{S}(g) \text{ and } b < i \leq e\}. \tag{2.8}$$

Let $\mathtt{v}_s(\Sigma^*) = \{a_{(s)} \mid a \in \Sigma^*\}$ for any $s \in [g]$ and any alphabet Σ^*. The set of states $V^{(\mathcal{N})}$ of graph $\mathcal{G}^{(\mathcal{N})}$ will now take its values from alphabet $\bigcup_{s\in[g]} \mathtt{v}_s(\Sigma)$ whereas the set of states $V^{(\mathcal{O})}$ of graph $\mathcal{G}^{(\mathcal{O})}$ will take its values from $\bigcup_{s\in[g]} \mathtt{v}_s(\Sigma^2)$. The subindex (s) in an alphabet symbol $a_{(s)}$ will denote the distance from the beginning of the grain that smeared symbol a at this location. For brevity, we will write a instead of $a_{(0)}$ and $\overline{a}$ instead of $a_{(1)}$ for any $a \in \Sigma^*$; in this notation $\Sigma = \mathtt{v}_0(\Sigma)$ and $\overline{\Sigma} = \mathtt{v}_1(\Sigma)$ (compare with Section 1.1). More specifically, $V^{(\mathcal{N})} = \bigcup_{s\in[g]} V_s^{(\mathcal{N})}$ and $V^{(\mathcal{O})} = \bigcup_{s\in[g]} V_s^{(\mathcal{O})}$ where $V_0^{(\mathcal{N})} = V_0^{(\mathcal{O})} = \{00, 11\}$,

$$V_s^{(\mathcal{N})} = \{01_{(s)}\} \cup \{0_{(s)}1\}$$

and

$$V_s^{(\mathcal{O})} = \{(01)_{(s)}, (10)_{(s)}\}.$$

The new definition of the operator $\partial(\cdot)$ is $\partial(\mathbf{x}) = (\partial(x_i))_{i\in[n]}$ where $\partial(a_{(s)}) = a$ for every $a \in \Sigma^*$ and every $s \in [g]$; that is, the operator $\partial(\cdot)$ strips off the subindex from a symbol returning the symbol back to its original alphabet. There is an edge in $E^{(\mathcal{N})}$ between $v = \ell r$ and $v' = \ell' r'$ if

[N1'] $v' \in V_0^{(\mathcal{N})}$, or

[4] The reason behind this restriction is the fact that otherwise we would have to encapsulate the hierarchy of the included grains in states of the graph which would quickly render the computation impractical, as then the number of states would grow exponentially in n.

[N2'] $\ell = \ell' \in \Sigma$, $rr' \in \mathtt{v}_s(\Sigma)\mathtt{v}_{s+1}(\Sigma)$, or $\ell\ell' \in \mathtt{v}_s(\Sigma)\mathtt{v}_{s+1}(\Sigma)$, $r = r' \in \Sigma$ for $s \in [g-1]$, or

[N3'] $\ell = r' \in \Sigma$, $\ell' r \in \mathtt{v}_1(\Sigma)\mathtt{v}_s(\Sigma)$, $\partial(\ell') = \partial(r)$, or $\ell' = r \in \Sigma$, $\ell r' \in \mathtt{v}_s(\Sigma)\mathtt{v}_1(\Sigma)$, $\partial(\ell) = \partial(r')$ for $s \in [g] \setminus \{0\}$.

There is an edge in $E^{(\mathcal{O})}$ between v and v' if

[O1'] $v' \in V_0^{(\mathcal{O})}$, or

[O2'] $v \in V_s^{(\mathcal{O})}$, $v' \in V_{s+1}^{(\mathcal{O})}$ for $s \in [g-1]$ and $\partial(v) = \partial(v')$, or

[O3'] $v \in V_s^{(\mathcal{O})}$, $v' \in V_1^{(\mathcal{O})}$ for $s \in [g] \setminus \{0\}$ and $\{\partial(v), \partial(v')\} = \{01, 10\}$.

It is easy to verify that the counterparts of Lemmas 2.4, 2.5 and Theorem 2.6 hold in this case. The reduced adjacency matrices $\mathcal{A}_{\mathcal{G}}^{(\mathcal{N})}$ and $\mathcal{A}_{\mathcal{G}}^{(\mathcal{O})}$ are of size $g \times g$ and are obtained after merging states of $V_s^{(j)}$ into superstate s for $j \in \{\mathcal{N}, \mathcal{O}\}$ and every $s \in [g]$:

$$\left[\mathcal{A}_{\mathcal{G}}^{(\mathcal{N})}(z,h)\right]_{j,j'} = \begin{cases} 1{+}h^2 & (j,j') = (0,0) \\ 2h & j \neq 0 \text{ and } j' = 0 \\ 2hz & (j,j') = (0,1) \\ h^2 z & j \neq 0 \text{ and } j' = 1 \\ z & j \notin \{0, g{-}1\} \text{ and } j' = j{+}1 \\ 0 & \text{otherwise} \end{cases}$$

and

$$\left[\mathcal{A}_{\mathcal{G}}^{(\mathcal{O})}(z,h,m)\right]_{j,j'} = \begin{cases} 1{+}h^2 & (j,j') = (0,0) \\ 2h & j \neq 0 \text{ and } j' = 0 \\ 2hz & (j,j') = (0,1) \\ h^2 m & j \neq 0 \text{ and } j' = 1 \\ z & j \notin \{0, g{-}1\} \text{ and } j' = j{+}1 \\ 0 & \text{otherwise} \end{cases}.$$

Example 2.2. For $g = 4$,

$$\mathcal{A}_{\mathcal{G}}^{(\mathcal{N})}(z,h) = \begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & 1{+}h^2 & 2hz & 0 & 0 \\ 1 & 2h & h^2 z & z & 0 \\ 2 & 2h & h^2 z & 0 & z \\ 3 & 2h & h^2 z & 0 & 0 \end{array}$$

and

$$\mathcal{A}_{\mathcal{G}}^{(\mathcal{O})}(z,h,m) = \begin{array}{c|cccc} & 0 & 1 & 2 & 3 \\ \hline 0 & 1{+}h^2 & 2hz & 0 & 0 \\ 1 & 2h & h^2m & z & 0 \\ 2 & 2h & h^2m & 0 & z \\ 3 & 2h & h^2m & 0 & 0 \end{array}.$$

□

The characteristic polynomial $\chi^{(\mathcal{N})}(x)$ of $\mathcal{A}_{\mathcal{G}}^{(\mathcal{N})}$ is

$$\chi^{(\mathcal{N})}(x) = x^{g-2}(x^2-(2+z)x-2z) - (1-(z/x)^{g-2})\frac{z^2(x+2)}{x(x-z)}.$$

For[5] $g \to \infty$, $\chi^{(\mathcal{N})}(x)$ converges to $x^g(x-2z-2)/(x-z)$ therefore $\lambda(\mathcal{A}_{\mathcal{G}}^{(\mathcal{N})})$ converges to $2(z{+}1)$. Plugging this into (2.4) yields

$$\lim_{g\to\infty} \varrho_2^{(\mathcal{N})}(\tau) = 1{-}\mathsf{H}_2(\tau) = \varrho_2^{(\mathrm{GV})}(\tau).$$

In other words, for $g \to \infty$, both $\varrho_2^{(\mathcal{N})}(\tau)$ and $\varrho_2^{(\mathcal{O})}(\tau)$, which satisfy the inequality $\varrho_2^{(\mathrm{GV})}(\tau) \le \varrho_2^{(\mathcal{O})}(\tau) \le \varrho_2^{(\mathcal{N})}(\tau)$, converge to the traditional Gilbert–Varshamov bound.

Remark 2.3. We leave it for the reader to verify that it is also possible to obtain Gilbert–Varshamov-like bounds in the general case when the values of q and g are arbitrary positive integers greater than 1. Both reduced adjacency matrices $\mathcal{A}_{\mathcal{G}}^{(\mathcal{N})}$ and $\mathcal{A}_{\mathcal{G}}^{(\mathcal{O})}$ are of size $\binom{g+1}{2} \times \binom{g+1}{2}$ when $q \ge 3$. □

2.3 Upper bound

Henceforth, we will restrict the discussion to $q = g = 2$ only. Let $\mathcal{C}$ be a binary t-grain-correcting code of length n. Let $\mathcal{B}_t(\mathbf{x})$ be the set of all words $\mathbf{w} \in [2]^n$ for which there exists a grain pattern $\mathcal{S}$ of size $|\mathcal{S}| \le t$ such that $\sigma_{\mathcal{S}}(\mathbf{x}) = \mathbf{w}$. (Since grains of such a grain pattern can be placed only between two successive runs (subvectors of consecutive identical symbols) of $\mathbf{x}$, one also has $|\mathcal{S}| \le \mathsf{r}(\mathbf{x}){-}1$ where $\mathsf{r}(\mathbf{x})$ is the number of runs in $\mathbf{x}$.) Let $\mathcal{B}_t(\mathcal{C};h)$ be the h-th smallest such set in $\mathcal{C}$. Readily, for any positive integer t there exists a positive integer u such that

$$\sum_{h=1}^{u} |\mathcal{B}_t(\mathcal{C};h)| \le |\Sigma^n| = 2^n \text{ and } \sum_{h=1}^{u+1} |\mathcal{B}_t(\mathcal{C};h)| > |\Sigma^n| = 2^n.$$

[5]We are abusing the notation here, implying by $g \to \infty$ that $n \to \infty$, yet the ratio of the inflicted errors to the word length cannot exceed τ.

Since $\sum_{h=1}^{|\mathcal{C}|} |\mathcal{B}_t(\mathcal{C};h)| \leq \sum_{\mathbf{c}\in\mathcal{C}} |\mathcal{B}_t(\mathbf{c})| \leq 2^n$, we obtain $|\mathcal{C}| \leq u$.

Suppose now that for any $\mathbf{c} \in \mathcal{C}$ we have a lower bound $\psi_t(r)$ on $|\mathcal{B}_t(\mathbf{c})|$ that depends only on the number of runs $r = \mathsf{r}(\mathbf{c})$ in $\mathbf{c}$ and that, as a function of r, it is nondecreasing. Hence, by sphere-packing arguments,

$$\Psi = \sum_{\mathbf{c}\in\mathcal{C}} \psi_t(\mathsf{r}(\mathbf{c})) \leq \sum_{\mathbf{c}\in\mathcal{C}} |\mathcal{B}_t(\mathbf{c})| \leq 2^n.$$

Let $\mathsf{N}(r)$ be the number of words $\mathbf{x} \in [2]^n$ with r runs. Let $\mathcal{U}$ be a largest set of words in Σ^n such that $\psi_t(\mathsf{r}(\mathbf{x})) \leq \psi_t(\mathsf{r}(\mathbf{x}'))$ for any $\mathbf{x} \in \mathcal{U}$, $\mathbf{x}' \in \mathcal{C} \setminus \mathcal{U}$ and $\sum_{\mathbf{x}\in\mathcal{U}} \psi_t(\mathsf{r}(\mathbf{x})) \leq \Psi$. Clearly, $|\mathcal{C}| \leq u \leq |\mathcal{U}|$, and the set $\mathcal{U}$ contains $|\mathcal{U}|$ smallest expressions $\psi_t(\mathsf{r}(\mathbf{x}))$ of words $\mathbf{x}$ in Σ^n. Since $\psi_t(r)$ is nondecreasing in r, we can group words with the same number of the runs till the largest integer ρ such that

$$\sum_{r=1}^{\rho} \mathsf{N}(r)\psi_t(r) \leq \Psi, \tag{2.9}$$

and thus,

$$|\mathcal{C}| \leq |\mathcal{U}| = \sum_{r=1}^{\rho} \mathsf{N}(r) + \left\lfloor \frac{\Psi - \sum_{r=1}^{\rho} \mathsf{N}(r)\psi_t(r)}{\psi_t(\rho+1)} \right\rfloor. \tag{2.10}$$

By replacing Ψ with 2^n in (2.9) and (2.10) we overcount in comparison with the right-hand side of (2.10). Hence

$$|\mathcal{C}| \leq \textstyle\sum_{r=1}^{\rho} \mathsf{N}(r) + \left\lfloor \frac{2^n - \sum_{r=1}^{\rho} \mathsf{N}(r)\psi_t(r)}{\psi_t(\rho+1)} \right\rfloor$$

where ρ is the largest integer such that $\sum_{r=1}^{\rho} \mathsf{N}(r)\psi_t(r) \leq 2^n$.

When overlaps are not allowed, we can bound $|\mathcal{B}_t(\mathbf{c})|$ with $\mathsf{r}(\mathbf{c}) = r$ from below using [26, Lemma 1]:

$$\psi_t^{(\mathcal{N})}(r) = 1 + \sum_{s=1}^{\min\{t,\lfloor (r-1)/3 \rfloor\}} \frac{1}{s!} \prod_{s'=0}^{s-1} (r-1-3s') \,. \tag{2.11}$$

When overlaps are allowed, we are able to calculate the size of $\mathcal{B}_t(\mathbf{c})$ with $\mathsf{r}(\mathbf{c}) = r$ precisely, namely,

$$\psi_t^{(\mathcal{O})}(r) = \sum_{s=0}^{\min\{t,r-1\}} \binom{r-1}{s} \,. \tag{2.12}$$

Both (2.11) and (2.12) are clearly nondecreasing functions on r. In both cases, $\mathsf{N}(r) = 2\binom{n-1}{r-1}$.

The next theorem summarizes the above discussion and, in fact, reformulates the sphere-packing bound that Abdel-Ghaffar and Weber [1, Th. 5] first used in a different context (also see [25, Sec. 7.3]).

Theorem 2.7. *For $j \in \{\mathcal{N}, \mathcal{O}\}$, let $\mathcal{C}$ be a binary t-grain-correcting code of length n. Then $|\mathcal{C}| \le \Delta^{(j)}(n,t)$ where*

$$\Delta^{(j)}(n,t) = 2\sum_{r=1}^{\rho}\binom{n-1}{r-1} + \left\lfloor \frac{2^n - 2\sum_{r=1}^{\rho}\binom{n-1}{r-1}\psi_t^{(j)}(r)}{\psi_t^{(j)}(\rho+1)} \right\rfloor$$

and ρ is the largest integer such that $\sum_{r=1}^{\rho}\binom{n-1}{r-1}\psi_t^{(j)}(r) \le 2^{n-1}$. The formulas for $\psi_t^{(\mathcal{N})}(r)$ and $\psi_t^{(\mathcal{O})}(r)$ are given in (2.11) and (2.12), respectively. □

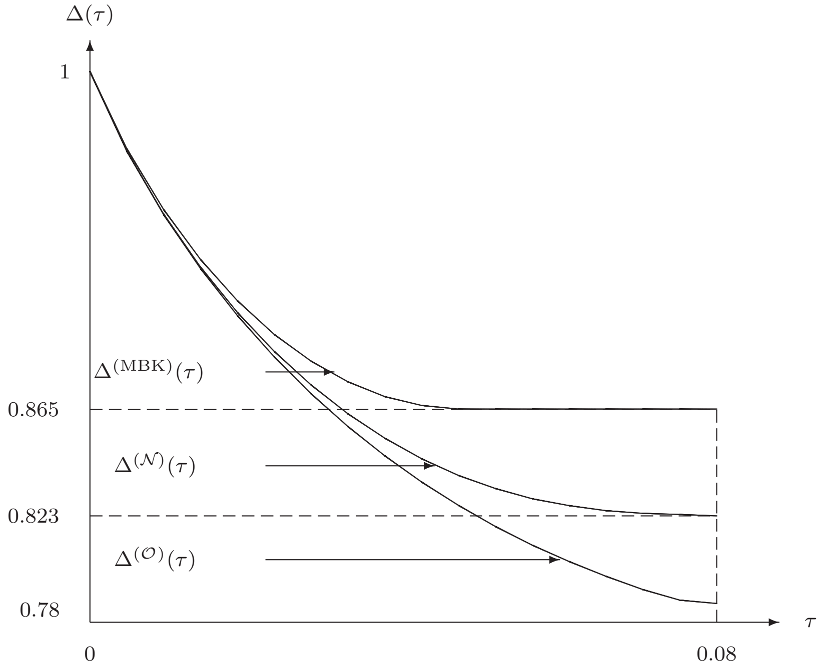

Figure 2.4: Functions $\Delta^{(\mathcal{N})}(\tau)$, $\Delta^{(\mathcal{O})}(\tau)$ and $\Delta^{(\mathrm{MBK})}(\tau)$.

Set $\tau = t/n$ and let $\Delta^{(j)}(\tau) = \log_2 \Delta^{(j)}(n, \tau n)/n$. Figure 2.4 depicts the functions $\tau \mapsto \Delta^{(j)}(\tau)$ for $j \in \{\mathcal{N}, \mathcal{O}\}$ calculated for $t \in \{1, 2, \ldots, 16\}$ and $n = 200$.

As we have mentioned in Section 1.1, Mazumdar *et al.* [26, Th. 2] obtained an upper bound on $M_2(n,t)$ (for $j = \mathcal{N}$) using a similar technique by considering a t-grain-correcting code $\mathcal{C}$ and dividing it into two subcodes

$$\mathcal{C}_1 = \{\mathbf{c} \in \mathcal{C} : |\mathsf{r}(\mathbf{c}) - n/2| \le n/2 - g(n,t)\} \text{ and } \mathcal{C}_2 = \mathcal{C} \setminus \mathcal{C}_1$$

where $g(n,t) = n/2 - \sqrt{nt\log_2 n}$. The sizes of $\mathcal{B}_t(\mathbf{c})$ for $\mathbf{c} \in \mathcal{C}_1$ and $\mathbf{c} \in \mathcal{C}_2$ were then bounded from below by $\psi_t^{(\mathcal{N})}(g(n,t))$ and 1, respectively. The obtained upper bound on $M_2(n,t)$ is

$$\Delta^{(\mathrm{MBK})}(n,t) = \frac{2^n t!}{(g(n,t)-1-3(t-1))^t} + 4\sum_{i=0}^{g(n,t)}\binom{n-1}{i}. \tag{2.13}$$

For specific values of n and t, this bound can be tightened by taking $g(n,t)$ to minimize the right-hand side of (2.13). Let

$$\Delta^{(\mathrm{MBK})}(\tau) = \frac{1}{n}\log_2\left[\min_{3t-2<g(n,t)<n/2}\Delta^{(\mathrm{MBK})}(n,t)\right].$$

The function $\tau \mapsto \Delta^{(\mathrm{MBK})}(\tau)$ is depicted in Figure 2.4 alongside the other bounds calculated for $t \in \{1,2,\ldots,16\}$ and $n = 200$, as well. It can be seen that the functions $\Delta^{(j)}(\tau)$ for $j \in \{\mathcal{N},\mathcal{O}\}$ improve on $\Delta^{(\mathrm{MBK})}(\tau)$ on the entire interval $(0, 0.08]$.

The asymptotic growth rate of $\Delta^{(\mathcal{N})}(\tau)$ as $n \to \infty$ is identical to that of $\Delta^{(\mathrm{MBK})}(\tau)$ (characterized by Proposition 1.2). The asymptotic growth rate of $\Delta^{(\mathcal{O})}(\tau)$, obtained in a similar vein, for $\tau \in [0, 0.114]$ equals $\mathsf{H}_2(\mathsf{x})$ where x is the smallest positive solution of

$$\mathsf{H}_2(\mathsf{x}) + \mathsf{x}\cdot\mathsf{H}_2\left(\frac{\tau}{\mathsf{x}}\right) = 1.$$

2.4 Constructions of grain-correcting codes

In this section, we restrict the discussion to $q = g = 2$ and disallow overlaps between grains.

It turns out that Construction 1.3 is the only way to construct binary ∞-grain-correcting codes of odd length n and size $2^{\lceil n/2\rceil}$.

Theorem 2.8. *Let n be an odd positive integer. The binary ∞-grain-correcting code of length n and size $2^{\lceil n/2\rceil}$ is unique.*

Proof. The proof is by induction on n. For $n = 1$, there is clearly only one binary ∞-grain-correcting code of size 2, which is $[2]$.

Let now $\mathcal{C}$ be a largest binary ∞-grain-correcting code of odd length n. Let $\mathcal{C}_{0*} \subseteq \mathcal{C}$ include codewords of $\mathcal{C}$ ending in either 00 or 01. Any two different codewords of $\mathcal{C}_{0*}$ are ∞-confusable, thus their prefixes of length $n-2$ are ∞-confusable as well. Therefore, the punctured code

$$\mathcal{C}'_{0*} = \{\mathbf{c} : \mathbf{c}00 \in \mathcal{C}_{0*} \text{ or } \mathbf{c}01 \in \mathcal{C}_{0*}\},$$

of length $n-2$, is ∞-grain-correcting. By the same token, the similarly defined punctured code $\mathcal{C}'_{1*}$ is ∞-grain-correcting. The sum $|\mathcal{C}_{1*}|+|\mathcal{C}_{0*}| = |\mathcal{C}'_{1*}|+|\mathcal{C}'_{0*}|$ clearly equals $|\mathcal{C}| = 2^{\lceil n/2\rceil}$, therefore both $\mathcal{C}'_{1*}$ and $\mathcal{C}'_{0*}$ are the largest possible ∞-grain-correcting codes of length $n-2$. By the induction hypothesis, the largest binary ∞-grain-correcting code of length $n-2$ is unique, hence $\mathcal{C}'_{1*} = \mathcal{C}'_{0*}$.

The only way a codeword of $\mathcal{C}'_{1*} = \mathcal{C}'_{0*}$ can be a prefix of two different codewords in $\mathcal{C}$ is when their suffixes are 00 and 11. This implies the uniqueness of the code $\mathcal{C}$ as well. The unique optimal ∞-grain-correcting code of length n is in fact obtained from Construction 1.3. $\square$

Remark 2.4. For even lengths, Construction 1.3 is not unique. For an even $n \geq 4$, at least four different largest constructions can be obtained by prepending the prefixes $\{00, 11\}$, $\{00, 10\}$, $\{01, 10\}$ or $\{01, 11\}$ to all the codewords of $\mathcal{C}_{n-2}$. $\square$

Construction 1.3 trivially yields $(\lfloor n/2 \rfloor - 1)$-grain-correcting codes of odd length n and size $2^{\lceil n/2 \rceil}$. We prove next that this size is optimal for $t = \lfloor n/2 \rfloor - 1$.

Theorem 2.9. *Let $n \geq 5$ be an odd integer. Then $M_2(n, \lfloor n/2 \rfloor - 1) = 2^{\lceil n/2 \rceil}$.*

Proof. The proof is similar to that of Theorem 2.8. The induction basis ($M_2(5, 1) = 8$) can be verified by a computer-based exhaustive search (also see Table 2.1).

Let now $\mathcal{C}$ be a largest binary $(\lfloor n/2 \rfloor - 1)$-grain-correcting code of odd length n. Let $\mathcal{C}'_{0*}$ and $\mathcal{C}'_{1*}$ be defined as in the proof of Theorem 2.8. These are $(\lfloor n/2 \rfloor - 2)$-grain-correcting codes of length $n-2$ such that $|\mathcal{C}'_{1*}| + |\mathcal{C}'_{0*}| = |\mathcal{C}|$. By the induction hypothesis,

$$|\mathcal{C}| = |\mathcal{C}'_{1*}| + |\mathcal{C}'_{0*}| \leq 2 \cdot 2^{\lceil n/2 \rceil - 1} = 2^{\lceil n/2 \rceil}. \tag{2.14}$$

The inequality in (2.14) cannot be strict since otherwise $\mathcal{C}$ will be of smaller size than the code $\mathcal{C}_n$ of Construction 1.3. Hence there is an equality, and the optimality of Construction 1.3 ensues. $\square$

Remark 2.5. Notice that the code $\mathcal{C}_n$ of Construction 1.3 is also ∞-grain-correcting in a wider sense (for the definition of the wider-sense confusability see the discussion preceding Lemma 2.4), hence Theorems 1.5 and 2.9 hold in this wider sense as well. $\square$

As for the binary $(n/2-1)$-grain-correcting codes of even length n, the value of $M_2(n, n/2-1)$ is realized by the augmentation of $\mathcal{C}_n$ with the words $(0110)^s(01)^{n-4s}$ and $(1001)^s(10)^{n-4s}$ for $s = \lfloor n/4 \rfloor$ (for the proof see Appendix B).

Theorem 2.10. *Let $n \geq 4$ be an even integer. Then $M_2(n, n/2-1) = 2^{n/2} + 2$.* $\square$

However, if we switch to the wider-sense confusability criterion, the size of the largest $(n/2-1)$-grain-correcting codes of even length n is $2^{n/2}$. The code $\mathcal{C}_n$ is again one of these largest codes.

Theorem 2.11. *Let $n \geq 4$ be an even integer. Then under the wider-sense confusability criterion, $M_2(n, n/2-1) = 2^{n/2}$.*

Proof. It is easy to verify the correctness of the claim for $n = 4$, hence we will assume further in the proof that $n \geq 6$.

W.l.o.g. we can assume that a largest code $\mathcal{C}$ is symmetric under bit complementation. The only words in $0[2]^{n-1}$ that are $n/2$-cws yet not $(n/2-1)$-cws are $\mathbf{c}_1 = (01)^{n/2}$ and $\mathbf{c}_2 = 00(10)^{n/2-1}$, therefore if $\{\mathbf{c}_1, \mathbf{c}_2\} \not\subseteq \mathcal{C}$, then $\mathcal{C}$ is an ∞-grain-correcting code (in a wider sense), and the result is implied by Theorem 1.5 and Remark 2.5.

If $\{\mathbf{c}_1, \mathbf{c}_2\} \subseteq \mathcal{C}$, then since any word $\mathbf{x} \in 0[2]^{n-1} \setminus \{\mathbf{c}_2\}$ is $(n/2-1)$-cws with $\mathbf{c}_1$, we have $\mathcal{C} = \{\mathbf{c}_1, \mathbf{c}_2, \overline{\mathbf{c}}_1, \overline{\mathbf{c}}_2\}$. Yet this contradicts the maximality of $\mathcal{C}$, since the size $2^{n/2}$ of the ∞-grain-correcting code (in a wider sense) $\mathcal{C}_n$ is greater than $|\mathcal{C}| = 4$ for $n \geq 6$. □

Using an inductive argument similar to that of Theorem 2.8, one can also prove that for $n \geq 5$, the binary $((n-3)/2)$-grain-correcting (in a wider sense) code of length n is unique.

t \ n	2	3	4	5	6	7	8	9
1	**2**	***4***	**6**	**8**	16	26*	44	
2			**4**	***8***	**10**	**16**	22	
3					**8**	***16***	**18**	**32**

Table 2.1: Sizes $M_2(n, t)$ of largest t-grain-correcting codes of length n.

t \ n	2	3	4	5	6	7	8	9
1	**2**	***4***	**4**	***8***	12	24*	32	
2			**4**	***8***	**8**	***16***	16	32
3					**8**	***16***	**16**	***32***

Table 2.2: Sizes $M_2(n, t)$ of largest t-grain-correcting codes in a wider sense of length n.

An interesting (yet not provably optimal) construction of binary 1-grain-correcting codes can be obtained by the augmentation of a Hamming code with a subset of $\mathcal{C}_n$. We cite this result in the following lemma.

Lemma 2.12. Let $m \geq 2$ be an integer and let $n = 2^m{-}1$. Then $M_2(n, 1) \geq 2^{n-m}{+}2^{(n-1)/2}$.

Proof. Consider a Hamming code $\mathcal{C}$ of length n with the parity-check matrix whose columns range over all the nonzero vectors in $[2]^m$ in the lexicographic order. Let

$$\mathcal{C}' = \left\{\mathbf{x}' = 0(x_i'x_i')_{i\in[(n-1)/2]} : \mathsf{w}\left(\mathbf{x}'\right) \in 4\mathbb{Z}{+}2\right\}$$

and

$$\mathcal{C}'' = \left\{\mathbf{x}'' = 1(x_i''x_i'')_{i\in[(n-1)/2]} : \mathsf{w}\left(\mathbf{x}''\right) \in 4\mathbb{Z}{+}1\right\}$$

where $\mathsf{w}(\mathbf{x})$ is the Hamming weight of word $\mathbf{x}$. Denote $\mathcal{C}^* = \mathcal{C}' \cup \mathcal{C}''$. Any codeword $\mathbf{c} \in \mathcal{C}$ is at Hamming distance 1, 2, or 3 and higher from a word $\mathbf{x} \in \mathcal{C}^*$. The only codeword of $\mathcal{C}$ at distance 1 from $\mathbf{x}$ is $\mathbf{x}{+}10^{n-1}$, but it is ∞-confusable with $\mathbf{x}$. Codewords of $\mathcal{C}$ at distance 2 from $\mathbf{x}$ differ from $\mathbf{x}$ in coordinates $1{+}2i$, $2{+}2i$ for $i \in [(n{-}1)/2]$, yet this makes codewords of $\mathcal{C}$ at distance 2 ∞-confusable with $\mathbf{x}$. Codewords of $\mathcal{C}$ at distance 3 from $\mathbf{x}$ are not 1-confusable with $\mathbf{x}$ merely because 1-confusable words are at most at Hamming distance 2 from one another. Moreover, the code $\mathcal{C}^* \subseteq \mathcal{C}_n$ is ∞-grain-correcting therefore $\mathcal{C} \cup \mathcal{C}^*$ is a 1-grain-correcting code of size $2^{n-m}{+}2^{(n-1)/2}$. □

Remark 2.6. Notice that the code $\mathcal{C} \cup \mathcal{C}^*$ is also 1-grain-correcting in a wider sense. □

Tables 2.1 and 2.2 contain the values of $M_2(n, t)$ (for the standard and the wider-sense notions of confusability, respectively) for small n and t obtained using computer search. Values marked in bold are guaranteed by Theorems 1.5, 2.9, 2.10 or 2.11; values marked in italics are attained by unique codes due to Theorem 2.8 (and its variations). One can also observe that for $(n, t) = (7, 1)$, marked by stars in both tables, the construction in Lemma 2.12 gives a code of size 24 which is close to the optimum $M_2(7, 1) = 26$. Under the wider-sense confusability criterion, the very same code is a largest 1-grain-correcting code of length 7 (by Remark 2.6).

2.5 Error detection

In this section we will prove that there exist ∞-grain-detecting codes with rates going to 1 as $n \to \infty$ for every fixed value of q (≥ 2). Consider the binary ($q = 2$) case first. For $\mathbf{x} \in \Sigma^n$, let $\mathsf{s}(\mathbf{x})$ denote the sum of indexes of the beginnings of runs of $\mathbf{x}$. Take all the binary words whose number of runs is either $\lfloor n/2 \rfloor$ or $\lfloor n/2 \rfloor + 1$ and categorize this set of words $\mathcal{F}$ by the value of $\mathsf{s}(\cdot)$ modulo $\lceil (n+1)/2 \rceil$ (modulo n when overlaps are allowed). By the pigeonhole principle, there has to be a category $\mathcal{C}_\mathcal{F}$, that is, a subset of $\mathcal{F}$, of size at least $|\mathcal{F}| / \lceil (n+1)/2 \rceil$ (at least $|\mathcal{F}| / n$ when overlaps are allowed) with the property that $\mathsf{s}(\mathbf{x})$ modulo $\lceil (n+1)/2 \rceil$ ($\mathsf{s}(\mathbf{x})$ modulo n when overlaps are allowed) is the same for any $\mathbf{x} \in \mathcal{C}_\mathcal{F}$.

Proposition 2.13. The code $\mathcal{C}_\mathcal{F}$ is an ∞-grain-detecting code with rate attaining 1 as $n \to \infty$ (either when overlaps are allowed or not).

Proof. A grain pattern $\mathcal{S}$ applied to a word $\mathbf{x} \in \mathcal{F}$ can either leave the number of runs $\mathsf{r}(\mathbf{y})$ in $\mathbf{y} = \sigma_\mathcal{S}(\mathbf{x})$ intact or decrease it by at least 2. If $\mathsf{r}(\mathbf{y})$ decreases by 2 (or more), then

$$\mathsf{r}(\mathbf{y}) \le \lfloor n/2 \rfloor - 1 < \lfloor n/2 \rfloor \le \mathsf{r}(\mathbf{x}),$$

and the error is detected. In particular, we will be able to detect such an error when words of $\mathcal{C}_\mathcal{F} \subseteq \mathcal{F}$ are transmitted.

Now, when the transmitted word $\mathbf{x}$ is from $\mathcal{C}_\mathcal{F}$ and $\mathsf{r}(\mathbf{y})$ remains intact, we can detect the inflicted errors by comparing $\mathsf{s}(\mathbf{y})$ with that of any word in $\mathcal{C}_\mathcal{F}$. Since the maximal size of $\mathcal{S}$ is $\lfloor n/2 \rfloor$ ($n-1$ when overlaps are allowed), and any single grain increases the value of $\mathsf{s}(\mathbf{x})$ by 1, the maximal difference between $\mathsf{s}(\mathbf{y})$ and $\mathsf{s}(\mathbf{x})$ is $\lfloor n/2 \rfloor$ ($n-1$ when overlaps are allowed), hence $\mathbf{y}$ and $\mathbf{x}$ are in different categories, viz., $\mathbf{y} \notin \mathcal{C}_\mathcal{F}$.

The size of $\mathcal{C}_\mathcal{F}$ is at least

$$\frac{2\left(\binom{n-1}{\lfloor n/2 \rfloor - 1} + \binom{n-1}{\lfloor n/2 \rfloor}\right)}{\lceil (n+1)/2 \rceil}$$

when overlaps are disallowed and at least

$$\frac{2\left(\binom{n-1}{\lfloor n/2 \rfloor - 1} + \binom{n-1}{\lfloor n/2 \rfloor}\right)}{n}$$

when overlaps are allowed, therefore the rate of $\mathcal{C}_\mathcal{F}$ is at least

$$1 - \frac{1.5 \log n}{n} + \frac{O(1)}{n},$$

in both scenarios and it attains 1 as $n \to \infty$. $\square$

An immediate corollary of Proposition 2.13 is that $\mathsf{R}_2(\tau) = 1$ for any $\tau \in [0, 1]$ when overlaps are allowed and for any $\tau \in [0, 0.5]$ when they are not. This is a rather surprising result as compared to the Hamming metric where it is known that the asymptotic rate of error-detecting codes is bounded away from 1 (for instance, due to the Singleton bound [22, Sec. 1.10]).

The code $\mathcal{C}_{\mathcal{F}}$ is easily generalized to larger alphabets, yielding a construction of rate

$$1 - \frac{1.5 \log_q n}{n} + \frac{O(1)}{n}$$

that attains 1 as $n \to \infty$ (the set $\mathcal{F}$ contains all the words with $\sim n(q-1)/q$ runs), which gives rise to the following theorem.

Theorem 2.14. *Let $q \geq 2$ be a positive integer. Then $\mathsf{R}_q(\tau) = 1$ for any τ.* □

Chapter 3

Objectives

- Trying to derive an explicit construction of a family of $\lceil \tau n \rceil$-grain-correcting codes of length n that attains (or exceeds) the lower bound on $R_q(\tau)$ that was presented in Section 2.2 for any value of q.

- Finding a practical extension of the model from Section 1.1 and the bounding technique of Section 2.2 to multi-dimensionsional media.

- There is a certain similarity between the model of grains and the binary asymmetric channel (the *Z-channel*) [18, Def. 2.1]. The theory of asymmetric binary channels might be a source of inspiration for discovering new constructions and/or bounding techniques. We will try to improve on the upper bounds on $R_q(\tau)$ discovered by Mazumdar *et al.* (Theorem 1.1, Proposition 1.2 and [26, Cor. 6]) employing the similarity to the Z-channel.

- As we can observe by comparing Tables 2.1 and 2.2, the relaxation of the confusability criterion does affect the largest code size $M_2(n, t)$. It would be interesting to find out whether the relaxation has an affect on the code rate $R_q(\tau)$ as well.

- The values $M_2(n, t)$ presented in Table 2.1 appear to be the same both when overlaps are allowed and disallowed. While it is easily explained under the wider-sense confusability criterion, understanding of this phenomenon under the standard confusability is currently missing. We will try to prove it or find a counter-example where allowing overlaps reduces the value of $M_2(n, t)$.

- At the end of Section 2.4, we have seen that the construction guaranteed by Lemma 2.12 yields the largest 1-grain-correcting code in a wider sense for $m = 3$. It is also easy to verify (with the aid of Table 2.2) that for $m = 2$ the construction gives the (unique) largest 1-grain-correcting code in a wider sense. It would be interesting to generalize this claim for any $m \geq 2$.

- The ∞-grain-detecting q-ary code $\mathcal{C}_{\mathcal{F}}$ suggested in Section 2.5 has redundancy $\propto 1.5 \log_q n$. It would be interesting to find out whether this is the minimum order of redundancy that an ∞-grain-detecting code can have or at least to give a lower bound on the redundancy. There is at least one scenario where it is easy to obtain this lower bound: when overlapping grains of any length are allowed, and, in addition, we assume that grains can be applied cyclically (i.e. the first cell in the array is not different from others), and the grain errors are inflicted in the following way. For any index $i \in [n]$ if no pair $(b, e) \in \mathcal{S}(g)$ satisfies $b < i \leq e$, then $y_i = c_i$, otherwise $y_i = c_{\mathsf{b}}$ where[1]

  $$\mathsf{b} = \arg\min \{b \mid \exists e : (b, e) \in \mathcal{S}(g) \text{ and } b < i \leq e\}$$

  (compare with (2.8)). In this case, a word and any of its cyclic shifts are undetectable, yielding a lower bound of $\propto \log_q n$ on the redundancy. Sperner's theorem [32] may be of some inspiration in this regard.

- The tools of Section 2.1 can be used to find the capacity of the flash memory where the memory cells are worn off evenly with a prescribed ratio of block erasures in the total number of writes (the results in [20, Sec. 3] were obtained in a similar vein). We hope to find a coding scheme that attains this capacity while leveling the wear of the cells.

- It would be interesting to find a specific family of $\langle n, M_0, M_1, 2\rangle$ WOM codes with WOM rates between the currently best known WOM rate of 1.4928 and the theoretical limit of $\log_2 3 \approx 1.58$.

[1] In the case of grains that wrap around, the same logic applies: the value of y_i is determined by the grain whose beginning is the "cyclically leftmost" among all the grains that contain y_i.

Appendix A

Proof of Lemma 2.4

Let $\mathbf{x} = (x_i)_{i\in[n]}, \mathbf{y} = (y_i)_{i\in[n]} \in \Sigma^n$ be a pair of t-cws words from $\mathcal{W}_t^{(\mathcal{N})}$. Then there exist grain patterns $\mathcal{S}, \mathcal{S}' \subset [n] \setminus \{0\}$ such that $\sigma_{\mathcal{S}}(\mathbf{x}) = \sigma_{\mathcal{S}'}(\mathbf{y})$ and $|\mathcal{S}| + |\mathcal{S}'|$ $(\leq 2t)$ is the minimal possible sum of sizes of grain patterns that confuse $\mathbf{x}$ and $\mathbf{y}$. Construct the path $\gamma = (v_i = \ell_i r_i)_{i\in[n]}$ corresponding to $(\mathbf{x}, \mathbf{y})$ as follows. Let ℓ_i be equal to $\overline{x_i} \in \overline{\Sigma}$ if $i \in \mathcal{S}$ and $x_i \in \Sigma$ otherwise (r_i's are defined in a similar way with y_i's and $\mathcal{S}'$). Clearly, $|\beta_L(\gamma)| = |\mathcal{S}|$, $|\beta_R(\gamma)| = |\mathcal{S}'|$ and thus $|\beta_L(\gamma)| + |\beta_R(\gamma)| \leq 2t$. Every v_i constructed this way is indeed a state in $V^{(\mathcal{N})}$:

- It cannot be of the form $\ell_i r_i \in \Sigma\Sigma$ where $\ell_i \neq r_i$ because then $\sigma_{\mathcal{S}}(\mathbf{x}) \neq \sigma_{\mathcal{S}'}(\mathbf{y})$.
- It cannot be of the form $\ell_i r_i \in \overline{\Sigma\Sigma}$ where $\partial(\ell_i) = \partial(r_i)$ because then $\sigma_{\mathcal{S}\setminus\{i\}}(\mathbf{x}) = \sigma_{\mathcal{S}'\setminus\{i\}}(\mathbf{y})$ contradicting the minimality of $|\mathcal{S}| + |\mathcal{S}'|$.
- It cannot be of the form $\ell_i r_i \in \Sigma\overline{\Sigma}$ (or $\in \overline{\Sigma}\Sigma$) where $\partial(\ell_i) = \partial(r_i)$ because then $\sigma_{\mathcal{S}}(\mathbf{x}) \neq \sigma_{\mathcal{S}'}(\mathbf{y})$ or grain i is redundant in $\mathcal{S}'$ (or in $\mathcal{S}$), contradicting the minimality of $|\mathcal{S}| + |\mathcal{S}'|$.

To verify that γ is indeed a path in $\mathcal{G}^{(\mathcal{N})}$, it is left to show that there are edges between the constructed $v_i = \ell_i r_i$ and $v_{i+1} = \ell_{i+1} r_{i+1}$ for any $i \in [n-1]$. Indeed, if $v_{i+1} \in V_0$ then $\ell_{i+1} = r_{i+1}$ and by Condition [N1], $(v_i, v_{i+1}) \in E^{(\mathcal{N})}$. If $v_{i+1} \in V_1$ and $v_{i+1} \in \Sigma\overline{\Sigma}$ (the case when $v_{i+1} \in \overline{\Sigma}\Sigma$ is similar), then $\partial(r_i) = \ell_{i+1}$ because otherwise $\sigma_{\mathcal{S}}(\mathbf{x}) \neq \sigma_{\mathcal{S}'}(\mathbf{y})$. Hence v_i can be only of the following forms:

- $v_i \in \Sigma^2$ where $\ell_i = r_i = \ell_{i+1}$. This corresponds to Condition [N2].
- $v_i \in \overline{\Sigma}\Sigma$ where $r_i = \ell_{i+1}$. This corresponds to Condition [N3].

If $v_{i+1} \in V_2$ then $v_i \in V_0$ because otherwise $\sigma_{\mathcal{S}}(\mathbf{x}) \neq \sigma_{\mathcal{S}'}(\mathbf{y})$. Moreover, $\ell_i = r_i \notin \{\partial(\ell_{i+1}), \partial(r_{i+1})\}$ since otherwise grain $i+1$ is redundant in $\mathcal{S}$ or in $\mathcal{S}'$. This corresponds to Condition [N4].

From the above discussion we conclude that $\gamma \in \mathcal{P}_t^{(\mathcal{N})}$.

To prove that there is at least one path in $\mathcal{P}_t^{(\mathcal{N})}$ for every ordered pair in $\mathcal{W}_t^{(\mathcal{N})}$, it remains to show that the above construction creates different paths for two different ordered pairs of t-cws words $(\mathbf{x} = (x_i)_{i\in[n]}, \mathbf{y} = (y_i)_{i\in[n]}) \neq (\mathbf{x}' = (x'_i)_{i\in[n]}, \mathbf{y}' = (y'_i)_{i\in[n]})$. W.l.o.g. assume that there exists $s \in [n] \setminus \{0\}$ such that $x_s \neq x'_s$ then state v_s in path γ in $\mathcal{G}^{(\mathcal{N})}$ constructed from $(\mathbf{x}, \mathbf{y})$ is different from state v'_s in path γ' in $\mathcal{G}^{(\mathcal{N})}$ constructed from $(\mathbf{x}', \mathbf{y}')$. Therefore $\gamma \neq \gamma'$.

We turn to prove now that there is exactly one path in $\mathcal{P}_t^{(\mathcal{N})}$ for every ordered pair in $\mathcal{W}_t^{(\mathcal{N})}$ by establishing the following correspondence. Let path $\gamma = (v_i = \ell_i r_i)_{i\in[n]} \in \mathcal{P}_t^{(\mathcal{N})}$ be matched with the ordered pair $(\mathbf{x}, \mathbf{y}) = ((x_i)_{i\in[n]}, (y_i)_{i\in[n]}) = (\partial((\ell_i)_{i\in[n]}), \partial((r_i)_{i\in[n]}))$. Then $|\beta_L(\gamma)| + |\beta_R(\gamma)| \leq 2t$, and by the definition of $\mathcal{G}^{(\mathcal{N})}$, $\beta_L(\gamma)$ and $\beta_R(\gamma)$ do not contain consecutive indexes. To prove that $\mathbf{x}$ and $\mathbf{y}$ are t-cws, we will show that the grain patterns $\mathcal{S} = \beta_L(\gamma)$ and $\mathcal{S}' = \beta_R(\gamma)$ satisfy $\sigma_{\mathcal{S}}(\mathbf{x}) = \sigma_{\mathcal{S}'}(\mathbf{y})$. Since $v_0 \in V_0$, one has $x_0 = y_0$. Let $\beta_j^{(s)} = \beta_j \cap [s]$ for $s \in [n] \setminus \{0\}$ and $j \in \{L, R\}$. We assume by induction that for $s \in [n] \setminus \{0\}$, $\sigma_{\beta_L^{(s)}}(\mathbf{x})$ is identical to $\sigma_{\beta_R^{(s)}}(\mathbf{y})$ in the first s symbols and prove this claim for $s+1$. If $v_s \in V_0$ then $\ell_s = r_s \in \Sigma$, $\beta_L^{(s+1)} = \beta_L^{(s)}$, and $\beta_R^{(s+1)} = \beta_R^{(s)}$, hence $\sigma_{\beta_L^{(s+1)}}(\mathbf{x})$ is identical to $\sigma_{\beta_R^{(s+1)}}(\mathbf{y})$ in the first $s+1$ symbols. If $v_s \in V_1$ and w.l.o.g. $v_s \in \Sigma\overline{\Sigma}$ (the case when $v_s \in \overline{\Sigma}\Sigma$ is similar) then $r_{s-1} = \ell_s$, implying that $\beta_L^{(s+1)} = \beta_L^{(s)}$ and $\beta_R^{(s+1)} = \beta_R^{(s)} \cup \{s\}$, and, again, $\sigma_{\beta_L^{(s+1)}}(\mathbf{x})$ is identical to $\sigma_{\beta_R^{(s+1)}}(\mathbf{y})$ in the first $s+1$ symbols. If $v_s \in V_2$ then $v_{s-1} \in V_0$ and $\ell_{s-1} = r_{s-1} \notin \{\partial(\ell_s), \partial(r_s)\}$ implying that $\beta_L^{(s+1)} = \beta_L^{(s)} \cup \{s\}$ and $\beta_R^{(s+1)} = \beta_R^{(s)} \cup \{s\}$, and, again, $\sigma_{\beta_L^{(s+1)}}(\mathbf{x})$ is identical to $\sigma_{\beta_R^{(s+1)}}(\mathbf{y})$ in the first $s+1$ symbols. This proves the existence of the desired grain patterns $\mathcal{S} = \beta_L, \mathcal{S}' = \beta_R$.

There is no other path $\gamma'{=}(\ell'_i r'_i)_{i\in[n]}{\neq}\gamma$ in $\mathcal{G}^{(\mathcal{N})}$ such that $(\mathbf{x}, \mathbf{y}) = (\partial((\ell'_i)_{i\in[n]}), \partial((r'_i)_{i\in[n]}))$: this implies the existence of an index $s \in [n{-}1]$ such that $\ell_s r_s = \ell'_s r'_s$, $\ell_{s+1} r_{s+1} \neq \ell'_{s+1} r'_{s+1}$, and $\partial(\ell_{s+1} r_{s+1}) = \partial(\ell'_{s+1} r'_{s+1})$. Yet this is impossible since the columns in $A_{\mathcal{G}}^{(\mathcal{N})}$ that are indexed by $\ell_{s+1} r_{s+1}$ and ℓ'_{s+1}, r'_{s+1} cannot both have a 1 in the same position.

In a similar vein, one can prove the one-to-one correspondence between $\mathcal{W}_t^{(\mathcal{O})}$ and $\mathcal{P}_t^{(\mathcal{O})}$.

Appendix B

Proof of Theorem 2.10

It is easy to verify that the augmentation of $\mathcal{C}_n$ with

$$\mathbf{y}_{(n)} = (0110)^{\lfloor n/4 \rfloor}(01)^{n-4\lfloor n/4 \rfloor}$$

and its binary complement $\overline{\mathbf{y}}_{(n)}$ yields an $(n/2-1)$-grain-correcting code, implying the lower bound $M_2(n, n/2-1) \geq 2^{n/2}+2$.

Assume that we have proved that

$$\mathbf{c}_{(n)} = (0011)^{\lfloor n/4 \rfloor}(00)^{n-4\lfloor n/4 \rfloor}$$

and the binary complement $\overline{\mathbf{c}}_{(n)}$ of $\mathbf{c}_{(n)}$ are in a code $\mathcal{C}$ of size $M_2(n, n/2-1)$. Define

$$\mathcal{T} = \left\{ \mathbf{x} = (x_i)_{i \in [n]} : x_{2s} = \overline{x}_{2s+1} \text{ for } s \in [n/2] \right\}.$$

There is only one word in $\mathcal{T}$, $\mathbf{y}_{(n)}$, that is $n/2$-confusable yet not $(n/2-1)$-confusable with $\mathbf{c}_{(n)}$; the other words in $\mathcal{T} \setminus \{\mathbf{y}_{(n)}\}$ starting with 0 are $(n/2-1)$-confusable with $\mathbf{c}_{(n)}$. Similarly, the only word of $\mathcal{T}$ that is $n/2$-confusable yet not $(n/2-1)$-confusable with $\overline{\mathbf{c}}_{(n)}$ is $\overline{\mathbf{y}}_{(n)}$, and the other words in $\mathcal{T} \setminus \{\overline{\mathbf{y}}_{(n)}\}$ starting with 1 are $(n/2-1)$-confusable with $\overline{\mathbf{c}}_{(n)}$. Therefore $\mathcal{C} \cap \mathcal{T} \subseteq \{\mathbf{y}_{(n)}, \overline{\mathbf{y}}_{(n)}\}$ and $\mathcal{C} \setminus \{\mathbf{y}_{(n)}, \overline{\mathbf{y}}_{(n)}\}$ is ∞-grain-correcting, hence by Theorem 1.5, $|\mathcal{C} \setminus \{\mathbf{y}_{(n)}, \overline{\mathbf{y}}_{(n)}\}| \leq 2^{n/2}$ implying $|\mathcal{C}| \leq 2^{n/2} + 2$. Notice that in the case of equality, $\{\mathbf{y}_{(n)}, \overline{\mathbf{y}}_{(n)}\} \subseteq \mathcal{C}$ due to the maximality of $\mathcal{C}$.

We prove that $\mathbf{c}_{(n)} \in \mathcal{C}$ by induction on the even values of $n \geq 4$. For $n = 4$, the claim is easily verified by the exhaustive search: there is only one largest code of size $M_2(4, 1) = 6$, and it contains both $\mathbf{c}_{(4)}$ and $\overline{\mathbf{c}}_{(4)}$, namely,

$$\{0000, \mathit{0011}, 0110, 1001, \mathit{1100}, 1111\}$$

($\mathbf{c}_{(4)}$ and $\overline{\mathbf{c}}_{(4)}$ are marked in italics). As for the inductive step, let us assume by contradiction that there exists a $(n/2-1)$-grain-correcting code $\widetilde{\mathcal{C}}$

of length n of size at least $2^{n/2}+4$ not containing $\{\mathbf{c}_{(n)}, \overline{\mathbf{c}}_{(n)}\}$. Therefore (w.l.o.g.) the number of words in $\widetilde{\mathcal{C}}$ starting with 0 is at least $2^{n/2-1}+2$. If we puncture the first two coordinates in all the words of $\widetilde{\mathcal{C}}$ starting with 0, the number of words in the obtained punctured code $\mathcal{C}^*$ will not change as otherwise there would have existed two 1-confusable words in $\widetilde{\mathcal{C}}$. The code $\mathcal{C}^*$ is of length $n-2$ and is $(n/2-2)$-grain-correcting hence by the induction hypothesis, $|\mathcal{C}^*| = 2^{n/2-1}+2$ and $\overline{\mathbf{c}}_{(n-2)}$, $\overline{\mathbf{y}}_{(n-2)} \in \mathcal{C}^*$.

Consequently, $01\overline{\mathbf{c}}_{n-2} \in \widetilde{\mathcal{C}}$, since we have assumed that $\mathbf{c}_{(n)} \notin \widetilde{\mathcal{C}}$. Yet $01\overline{\mathbf{c}}_{(n-2)}$ and $01\overline{\mathbf{y}}_{(n-2)}$, as well as $01\overline{\mathbf{c}}_{(n-2)}$ and $00\overline{\mathbf{y}}_{(n-2)}$, are $(n/2-1)$-confusable; this is a contradiction since either $00\overline{\mathbf{y}}_{(n-2)}$ or $01\overline{\mathbf{y}}_{(n-2)}$ is in $\widetilde{\mathcal{C}}$. Thus, any largest $(n/2-1)$-grain-correcting code of length n and size $M_2(n, n/2-1)$ contains $\mathbf{c}_{(n)}$, hence by the previous proof, it is of size $2^{n/2}+2$.

Bibliography

[1] K.A.S. Abdel-Ghaffar, J.H. Weber, *Bounds and constructions for runlength limited error-control block codes, IEEE Trans. Inform. Theory,* vol. 37, pp. 789–800, 1991.

[2] R. Ahlswede, Z. Zhang, *Coding for write-efficient memory, Inform. and Comput.,* vol. 83, pp. 80–97, 1989.

[3] A. Barg, A. Mazumdar, *Codes in permutations and error correction for rank modulation, Proc. IEEE Symp. Inform. Theory (ISIT 2010),* Austin, TX, 2010, pp. 854–858.

[4] C. Bracken, T. Helleseth, *Triple-error-correcting BCH-like codes, Proc. IEEE Symp. Inform. Theory (ISIT 2009),* Seoul, Korea, 2009, pp. 1723–1725.

[5] G.D. Cohen, P. Godlewski, F. Merkx, *Linear binary code for write-once memories, IEEE Trans. Inform. Theory*, vol. 32, pp. 697–700, 1986.

[6] T.M. Cover, J.A. Thomas, *Elements of Information Theory,* Wiley, New York, NY, 1991.

[7] I. Csiszár, T.M. Cover, B.-S. Choi, *Conditional limit theorems under Markov conditioning, IEEE Trans. Inform. Theory,* vol. 33, pp. 788–801, 1987.

[8] F. Fu, A.J.H. Vinck, *On the capacity of generalized write-once memory with state transitions described by an arbitrary directed acyclic graph, IEEE Trans. Inform. Theory,* vol. 45, pp. 308–313, 1999.

[9] F.R. Gantmacher, *Matrix Theory, Volume II,* Chelsea Publishing Company, New York, 1960.

[10] P. Godlewski, *WOM-codes construits à partir des codes de Hamming, Discrete Math.,* vol. 65, pp. 237–243, 1987.

[11] C. Heegard, *On the capacity of permanent memory, IEEE Trans. Inform. Theory,* vol. 30, pp. 470–480, 1984.

[12] K.A.S. Immink, *Codes for Mass Data Storage Systems,* Shannon Foundation Publishers, Eindhoven, Netherlands, 2004.

[13] A.R. Iyengar, P.H. Siegel, J.K. Wolf, *Write channel model for bit-patterned media recording, IEEE Trans. Magn.*, vol. 47, pp. 35–45, 2011.

[14] A. Jiang, V. Bohossian, J. Bruck, *Floating codes for joint information storage in write asymmetric memories, Proc. IEEE Symp. Inform. Theory (ISIT 2007),* Nice, France, 2007, pp. 1166–1170.

[15] A. Jiang, R. Matescu, M. Schwartz, J. Bruck, *Rank modulation for flash memories, Proc. IEEE Symp. Inform. Theory (ISIT 2008),* Toronto, ON, 2008, pp. 1731–1735.

[16] A. Jiang, M. Schwartz, J. Bruck, *Error-correcting codes for rank modulation, Proc. IEEE Symp. Inform. Theory (ISIT 2008),* Toronto, ON, 2008, pp. 1736–1740.

[17] T. Kasami, *The weight enumerators for several classes of subcodes of the second order binary Reed–Muller codes, Inform. and Contr.,* vol. 18, pp. 369–394, 1971.

[18] T. Kløve, *Error correcting codes for the asymmetric channel, Tech. Rep. 18–09–07–81,* Dep. Inform., Univ. Bergen, 1981.

[19] V.D. Kolesnik, V.Yu. Krachkovsky, *Generating functions and lower bounds on rates for limiting error-correcting codes, IEEE Trans. Inform. Theory,* vol. 37, pp. 778–788, 1991.

[20] L.A. Lastras-Montaño, M. Franceschini, T. Mittelholzer, J. Karidis, M. Wegman, *On the lifetime of multilevel memories, Proc. IEEE Symp. Inform. Theory (ISIT 2009),* Seoul, Korea, 2009, pp. 1224–1228.

[21] D.G. Luenberger, *Introduction to Linear and Nonlinear Programming,* Addison-Wesley, Reading, MA, 1973.

[22] F.J. MacWilliams, N.J.A. Sloane, *The Theory of Error-Correcting Codes,* North-Holland, Amsterdam, 1977.

[23] H. Mahdavifar, P.H. Siegel, A. Vardy, J.K. Wolf, E. Yaakobi, *A nearly optimal construction of flash codes, Proc. IEEE Symp. Inform. Theory (ISIT 2009),* Seoul, Korea, 2009, pp. 1239–1243.

[24] B.H. Marcus, R.M. Roth, *Improved Gilbert–Varshamov bound for constrained systems, IEEE Trans. Inform. Theory,* vol. 38, pp. 1213–1221, 1992.

[25] B.H. Marcus, R.M. Roth, P.H. Siegel, *Constrained systems and coding for recording channels,* in *Handbook of Coding Theory,* V.S. Pless and W.C. Huffman (Editors), Elsevier Scientific Publishers, Amsterdam, Netherlands, 1998.

[26] A. Mazumdar, A. Barg, N. Kashyap, *Coding for high-density magnetic recording, Proc. IEEE Symp. Inform. Theory (ISIT 2010),* Austin, TX, 2010, pp. 978–982.

[27] W. Parry, S. Tuncel, *Classification Problems in Ergodic Theory, London Math. Soc. Lecture Note Series,* vol. 67, Cambridge University Press, 1982.

[28] R.L. Rivest, A. Shamir, *How to reuse a "write-once" memory, Inform. Contr.,* vol. 55, pp. 1–19, 1982.

[29] T. Rockafellar, *Convex Analysis,* Princeton University Press, 1970

[30] R.M. Roth, *Introduction to Coding Theory,* Cambridge University Press, 2006.

[31] C.E. Shannon, *The zero error capacity of a noisy channel, IRE Trans. Inform. Theory,* vol. 2, pp. 8–19, 1956.

[32] E. Sperner, *Ein Satz über Untermengen einer endlichen Menge, Mathematische Zeitschrift,* vol. 27, pp. 544–548, 1928.

[33] L.M.G.M. Tolhuizen, *The generalized Gilbert–Varshamov bound is implied by Turáns theorem, IEEE Trans. Inform. Theory,* vol. 43, pp. 1605–1606, 1997.

[34] M.A. Tsfasman, S.G. Vlădut, T. Zink, *Modular curves, Shimura curves, and Goppa codes, better than Varshamov–Gilbert bound, Math. Nachr.,* vol. 109, pp. 21–28, 1982.

[35] S.X. Wang, A.M. Taratorin, *Magnetic Information Storage Technology,* Academic Press, 1999.

[36] R. Wood, M. Williams, A. Kavcic, J. Miles, *The feasibility of magnetic recording at 10 terabits per square inch on conventional media, IEEE Trans. Magn.,* vol. 45, pp. 917–923, 2009.

[37] E. Yaakobi, S. Kayser, P.H. Siegel, A. Vardy, J.K. Wolf, *Efficient two-write WOM-codes, Proc. IEEE Inform. Theory Workshop,* Dublin, Ireland, 2010.

[38] E. Yaakobi, P.H. Siegel, A. Vardy, J.K. Wolf, *Multiple error-correcting WOM-Codes, Proc. IEEE Symp. Inform. Theory (ISIT 2010),* Austin, TX, 2010, pp. 1933–1937.

[39] E. Yaakobi, A. Vardy, P.H. Siegel, J.K. Wolf, *Multidimensional flash codes, Proc. Allerton Conf. Commun., Control and Comp.,* Monticello, IL, 2008, pp. 392–399.

[40] G. Zémor, G.D. Cohen, *Error-correcting WOM-codes, IEEE Trans. Inform. Theory,* vol. 37, pp. 730–734, 1991.

צפינה ליישומים חדשים בהתקני אחסון מידע

ארטיום שרוב