

# Coding for New Applications in Storage Media

## DOCTORAL THESIS PROPOSAL

Under supervision of Prof. Ron M. Roth

**Artyom Sharov**

*Department of Computer Science*

Technion – Israel Institute of Technology

# Introduction

- Conventional magnetic recording media: *grains*
- Wood *et al.* suggested a new approach for magnetizing areas as small as a single grain
- Mazumdar *et al.* considered a 1-D combinatorial error model with insertion/deletion errors and grains of length 1 and 2

# Combinatorial model

- $[s] = \{0, 1, \dots, s-1\}$
- An alphabet  $\Sigma = [q]$
- A *grain* (of length 2) ending at location  $e \in [n] \setminus \{0\}$  in a word  $\mathbf{x} = (x_i)_{i \in [n]}$  smears the value of cell  $e-1$  to cell  $e$ :  $x_e \leftarrow x_{e-1}$
- A *grain pattern*  $\mathcal{S} \subseteq [n] \setminus \{0\}$  contains all the grain locations and inflicts errors to  $\mathbf{x}$  by means of operator  $\sigma_{\mathcal{S}}$
- $\mathcal{S}$  has *overlaps* if there exist  $e, e' \in \mathcal{S}$  such that  $e' = e+1$ ; otherwise  $\mathcal{S}$  is *nonoverlapping*

# Example

- $\Sigma = [3]$ ,  $n = 6$ ,  $\mathbf{x} = 102022$ ,  $S = \{1, 3, 5\}$  and  $S' = \{1, 2\}$

$\mathbf{x}$ : 

1	0	2	0	2	2
---	---	---	---	---	---

$\mathbf{x}$ : 

1	0	2	0	2	2
---	---	---	---	---	---



$S$ : 

1	0	2	0	2	2
---	---	---	---	---	---

$S'$ : 

1	0	2	0	2	2
---	---	---	---	---	---



$\sigma_S(\mathbf{x})$ : 

1	1	2	2	2	2
---	---	---	---	---	---

$\sigma_{S'}(\mathbf{x})$ : 

1	1	0	0	2	2
---	---	---	---	---	---

## Combinatorial model (cont.)

- *Risk set*  $\mathcal{R}_t(\mathbf{x})$  is the set of words  $\mathbf{y} \in \Sigma^n$  such that there exist  $S, S'$  of size  $t$  at most for which  $\sigma_S(\mathbf{x}) = \sigma_{S'}(\mathbf{y})$
- If  $\mathbf{y} \in \mathcal{R}_t(\mathbf{x})$  then  $\mathbf{x}$  and  $\mathbf{y}$  are *t-confusable*
- If  $\mathbf{x}$  and  $\mathbf{y}$  are *t-confusable* for some  $t$  then  $\mathbf{x}$  and  $\mathbf{y}$  are *finitely-confusable*; otherwise  $\mathbf{x}$  and  $\mathbf{y}$  are  *$\infty$ -confusable*
- A code  $\mathcal{C} \subseteq \Sigma^n$  is *t-grain-correcting* if no two distinct codewords are *t-confusable*
- The largest size  $M_q(n, t)$
- The *rate*  $R_q(\tau) = \limsup_{n \rightarrow \infty} \frac{1}{n} \log_q M_q(n, t)$
- Asymmetry  $|\mathcal{R}_1(111)| = 3 \neq 4 = |\mathcal{R}_1(101)|$

## Lower bound on $M_q(n, t)$

- The number of ordered pairs of  $t$ -confusable words in  $\mathcal{X}$ :

$$W_t(\mathcal{X}) = \sum_{\mathbf{x} \in \mathcal{X}} |\mathcal{R}_t(\mathbf{x}) \cap \mathcal{X}|$$

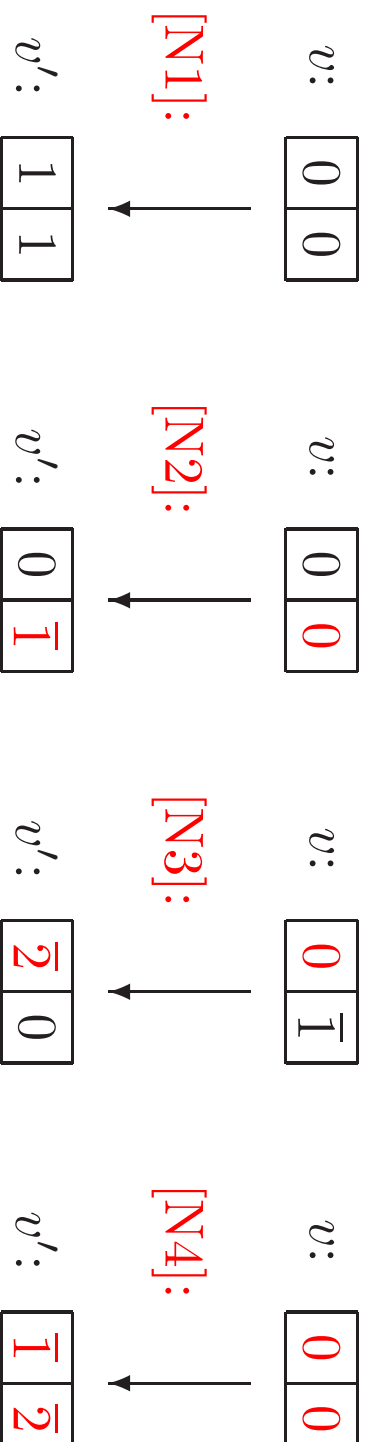
**Lemma.** *Let  $n, t$  be positive integers and let  $\mathcal{X} \subseteq \Sigma^n$ , then*

$$M_q(n, t) \geq \frac{|\mathcal{X}|^2}{4W_t(\mathcal{X})}.$$

- We will evaluate  $W_t(\mathcal{X})$  for certain sets  $\mathcal{X}$  with prescribed empirical distribution of transitions

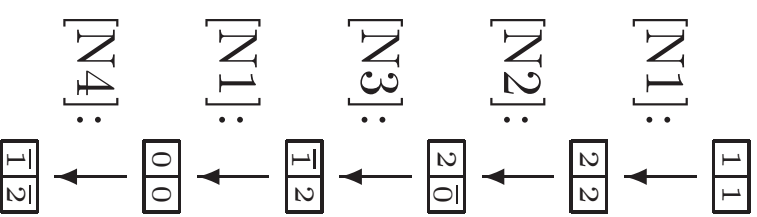
# Graph $\mathcal{G}(\mathcal{N}) = (V(\mathcal{N}), E(\mathcal{N}))$

- Set of states  $V(\mathcal{N}) = V_0 \cup V_1 \cup V_2$  where  $V_0 = \{aa : a \in \Sigma\}$ ,  $V_1 = \{a\bar{b} : ab \in \Sigma^2, a \neq b\}$  and  $V_2 = \{\bar{a}\bar{b} : ab \in \Sigma^2, a \neq b\}$
- For  $q = 2$ ,  $V_0 = \{00, 11\}$ ,  $V_1 = \{\bar{0}1, 0\bar{1}, \bar{1}0, 1\bar{0}\}$ ,  $V_2 = \{0\bar{1}, \bar{1}0\}$
- There is an edge in  $E(\mathcal{N})$  from  $v$  to  $v'$ :



# Example

- $\Sigma = [3]$ ,  $n = 6$ ,  $\gamma = (v_i)_{i \in [n]} = 11 \quad 22 \quad 2\bar{0} \quad \bar{1}2 \quad 00 \quad \bar{1}2$



- The patterns  $S = \{3, 5\}$  and  $S' = \{2, 5\}$  make  $\mathbf{x} = 122101$  (the left path) and  $\mathbf{y} = 120202$  (the right path) confusable

# Adjacency matrix $A_G^{(\mathcal{N})}$ for $q = 2$

$$A_G^{(\mathcal{N})} = \begin{array}{c|cccccc} & 00 & \bar{0}1 & 0\bar{1} & \bar{1}0 & 1\bar{0} & 11 \\ \hline 00 & 1 & 0 & 1 & 1 & 0 & 1 \\ \bar{0}1 & 1 & 0 & 0 & 0 & 1 & 1 \\ 0\bar{1} & 1 & 0 & 0 & 1 & 0 & 1 \\ \bar{1}0 & 1 & 0 & 1 & 0 & 0 & 1 \\ 1\bar{0} & 1 & 1 & 0 & 0 & 0 & 1 \\ 11 & 1 & 1 & 0 & 0 & 1 & 1 \end{array}$$

## Wider-sense confusability

- Words  $\mathbf{x}$ ,  $\mathbf{y}$  are *t-confusable in a wider sense* (or *t-cws*) if there exist grain patterns  $S$  and  $S'$  such that  $|S| + |S'| \leq 2t$  and  $\sigma_S(\mathbf{x}) = \sigma_{S'}(\mathbf{y})$
- Any *t*-grain-correcting code in a wider sense is also a *t*-grain-correcting in the ordinary sense
- The forthcoming results apply to this notion of confusability
- Empirically, this relaxation does not result in worse bounds while using our technique

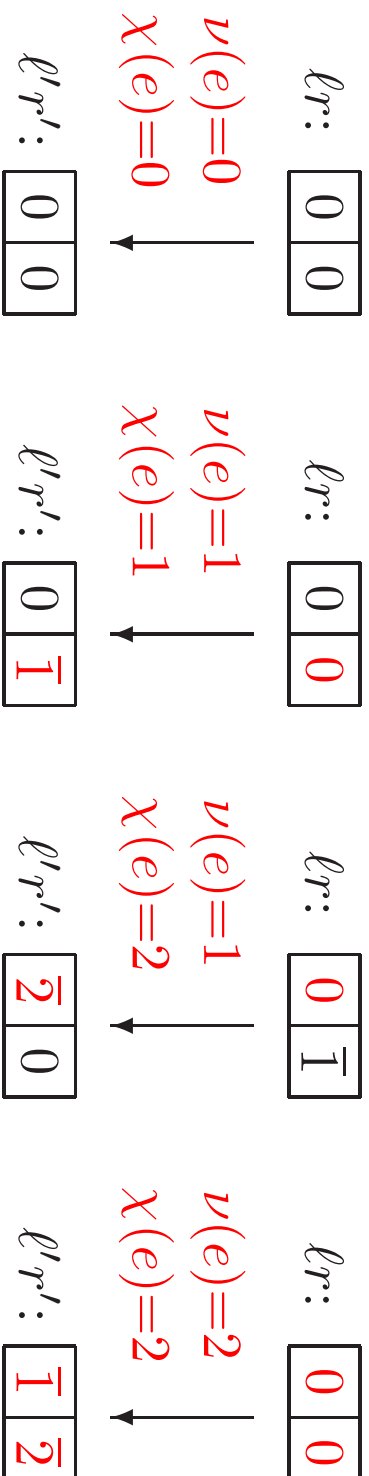
# Correspondence between pairs of words and paths

**Lemma.** For each  $t$ -cws (ordered) pair  $(\mathbf{x}, \mathbf{y}) \in \Sigma^n \times \Sigma^n$  there is exactly one path  $\gamma = (v_i)_{i \in [n]}$  in  $\mathcal{G}(\mathcal{N})$  such that

1.  $v_0 \in V_0$ ,
2.  $\mathbf{x} = (\partial(\ell_i))_{i \in [n]}$ ,
3.  $\mathbf{y} = (\partial(r_i))_{i \in [n]}$  and
4. The total number of grains that confuse  $\mathbf{x}$  and  $\mathbf{y}$  is at most  $2t$ .

# Function $f(\mathcal{N})$

- Function  $f(\mathcal{N}) : E(\mathcal{N}) \rightarrow [3]^2$
- For an edge  $e = (\ell r, \ell' r') \in E(\mathcal{N})$ ,  $f(\mathcal{N})(e) = (\nu(e), \chi(e))$
- $\nu(e)$  counts the smallest number of grains confusing  $\ell'$  and  $r r'$ ;
- $\chi(e)$  counts the number of crossovers



# Matrix function $A_G^{(\mathcal{N})}$

- $[A_G^{(\mathcal{N})}(z, h)]_{v, v' \in V} = \begin{cases} z^{\nu(e)} h^{\chi(e)} & e = (v, v') \in E^{(\mathcal{N})} \\ 0 & \text{otherwise} \end{cases}$
- For  $q = 2$ ,

$$A_G^{(\mathcal{N})}(z, h) = \begin{array}{c|cccccc} & 00 & \bar{0}1 & 0\bar{1} & \bar{1}0 & 1\bar{0} & 11 \\ \hline 00 & 1 & 0 & hz & hz & 0 & h^2 \\ \bar{0}1 & h & 0 & 0 & 0 & h^2z & h \\ 0\bar{1} & h & 0 & 0 & h^2z & 0 & h \\ \bar{1}0 & h & 0 & h^2z & 0 & 0 & h \\ 1\bar{0} & h & h^2z & 0 & 0 & 0 & h \\ 11 & h^2 & hz & 0 & 0 & hz & 1 \end{array}$$

# Main theorem

- Applying special cases of lemmas from [MR92]: optimizing convex functions subject to linear equality and inequality constraints
- Asymptotic upper bound on the number of paths with average number of crossovers  $\sim 2p$  and number of confusing grains  $\leq 2\tau$

$$K^{(\mathcal{N})} = \inf_{z \in (0,1], h \in (0,\infty)} \{ \log_q \lambda(A_G^{(\mathcal{N})}(z, h)) - 2\tau \log_q z - 2p \log_q h \}$$

**Theorem.** *Let  $\tau \in (0, 1)$ , then<sup>a</sup>*

$$R_q(\tau) \geq \varrho_q^{(\mathcal{N})}(\tau) = \sup_{p \in [0,1]} \{ 2H_q(p) - K^{(\mathcal{N})} \}$$

---

<sup>a</sup>Asymptotic version of  $M_q(n, t) \geq \frac{|\mathcal{X}|^2}{4W_t(\mathcal{X})}$

# Merging states of $G^{(\mathcal{N})}$

- Similar to the standard procedure for reducing the number of states in a presentation of a constrained system while preserving its spectral radius
- The states of  $V_0$  can be merged into superstate 0,  $V_1$  — into superstate 1,  $V_2$  — into superstate 2
- Reduced matrix  $A_G^{(\mathcal{N})}$ :

$$A_G^{(\mathcal{N})} = \begin{array}{c|ccc} & 0 & 1 & 2 \\ \hline 0 & 1+(q-1)h^2 & 2(q-1)hz & (q-1)(q-2)h^2z^2 \\ 1 & 2h+(q-2)h^2 & (q-1)h^2z & 0 \\ 2 & 2h+(q-2)h^2 & 0 & 0 \end{array}$$

# Overlaps allowed

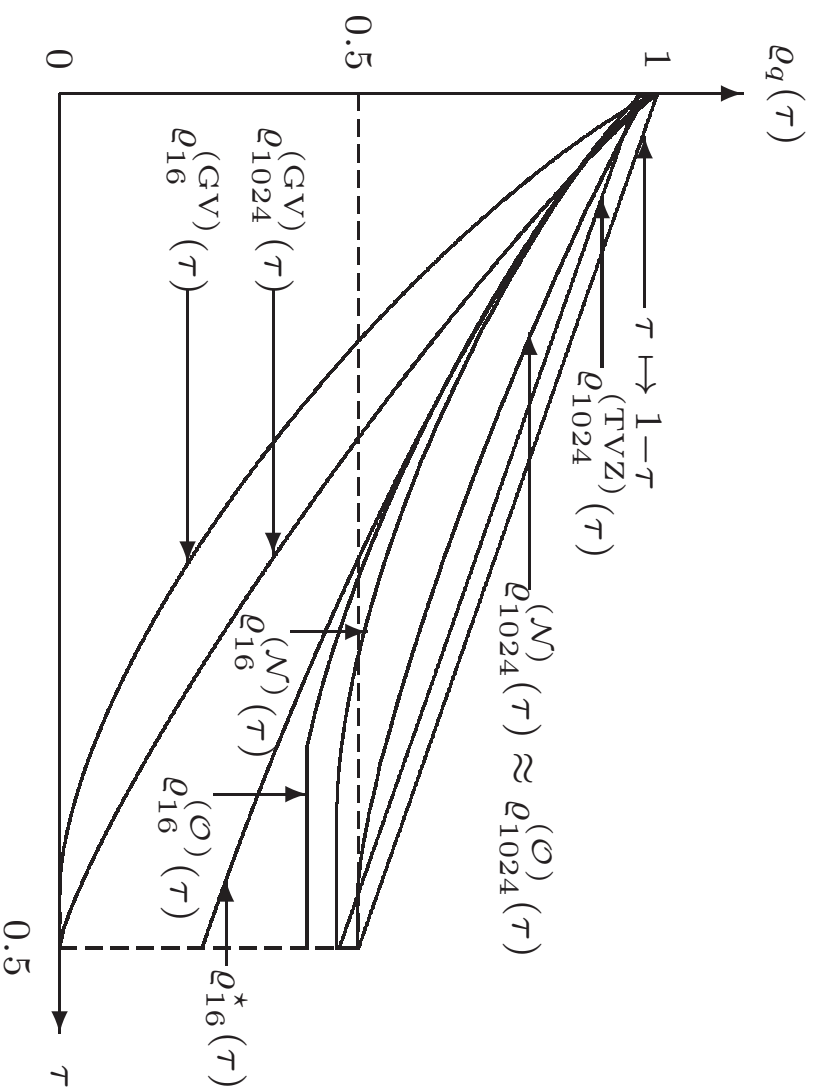
$$[A_g^{(\mathcal{O})}(z, h, m)]_{v, v' \in V} = \begin{cases} z^{\omega(e)} h^{\chi(e)} m^{\mu(e)} & e = (v, v') \in E^{(\mathcal{O})} \\ 0 & \text{otherwise} \end{cases}$$

$$K^{(\mathcal{O})} = \sup_{0 \leq \eta \leq 2\tau} \inf_{z \in (0, 1], h, m \in (0, \infty)} \{ \log_q \lambda(A_g^{(\mathcal{O})}(z, h, m)) - (2\tau - \eta) \log_q z - \eta \log_q m - 2p \log_q h \}$$

Lower bound on the rate  $\varrho_q^{(\mathcal{O})}(\tau) = \sup_{p \in [0, 1]} \{ 2H_q(p) - K^{(\mathcal{O})} \}$

	0	1
$A_g^{(\mathcal{O})} =$	$\begin{matrix} 0 & 1 + (q-1)h^2 & 2(q-1)hz + (q-1)(q-2)h^2z^2 \\ 1 & 2h + (q-2)h^2 & h^2m + 2(q-2)h^2z \end{matrix}$	

# Lower bounds $\varrho_q^{(j)}(\tau)$



Bounds  $\varrho_{1024}^{(TVZ)}(\tau)$  and  $\varrho_{16}^*(\tau)$  rely on the code  $\{\mathbf{c} = (c_i)_{i \in [n]} \in \Sigma^n : c_i \neq c_{i+1} \text{ for any } i \in [n-1]\}$

# Upper bound

- $\mathcal{C}$  is a binary  $t$ -grain-correcting code of length  $n$
- Set  $\mathcal{B}_t(\mathbf{x})$  of all words  $\mathbf{w} \in [2]^n$  for which there is  $S$  of size  $t$  at most such that  $\sigma_S(\mathbf{x}) = \mathbf{w}$
- Lower bound  $\psi_t(r)$  on  $|\mathcal{B}_t(\mathbf{c})|$  that depends only on the number of runs  $r = r(\mathbf{c})$  in  $\mathbf{c}$ ;  $N(r)$  is the number of words with  $r$  runs
- By sphere-packing arguments,

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

- $|\mathcal{C}| \leq \sum_{r=1}^{\rho} N(r) + \left\lfloor \frac{\Psi - \sum_{r=1}^{\rho} N(r) \psi_t(r)}{\psi_t(\rho+1)} \right\rfloor$  where  $\rho$  is the largest such that  $\sum_{r=1}^{\rho} N(r) \psi_t(r) \leq 2^n$

## Upper bound (cont.)

- Lower bounds on  $|\mathcal{B}_t(\mathbf{c})|$  without overlaps

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

with overlaps

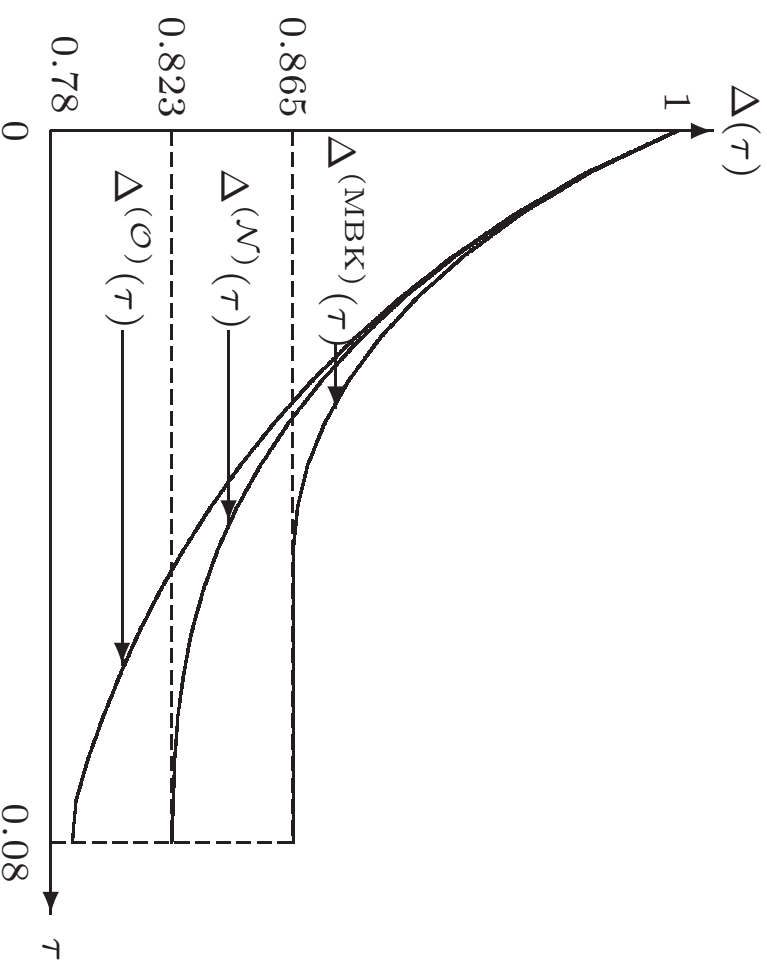
$$\psi_t^{(\mathcal{O})}(r) = \sum_{s=0}^{\min\{t, r-1\}} \binom{r-1}{s}$$

**Theorem.** *Let  $\mathcal{C}$  be a binary  $t$ -grain-correcting code of length  $n$ , then  $|\mathcal{C}| \leq \Delta^{(j)}(n, t)$  for  $j \in \{\mathcal{N}, \mathcal{O}\}$  where*

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

# Upper bounds $\Delta^{(j)}(\tau)$

$n = 200, t \in \{1, 2, \dots, 16\}, \tau = t/n, \Delta^{(j)}(\tau) = \Delta^{(j)}(n, \tau n)$



Asymptotically,  $\Delta^{(\mathcal{N}')}(\tau)$  and  $\Delta^{(\text{MBK})}(\tau)$  are the same;  $\Delta^{(\emptyset)}(\tau)$  grows like  $H_2(x)$  where  $x$  is the smallest positive solution of  $H_2(x) + x \cdot H_2(\tau/x) = 1$

# Constructions of grain-correcting codes

**Theorem** (due to Mazumdar *et al.*). For any  $n \in \mathbb{Z}^+$ ,

$$M_2(n, \lfloor n/2 \rfloor) = 2^{\lceil n/2 \rceil}$$

For even  $n$ ,

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

For odd  $n$ ,  $\mathcal{C}_n = (0 \mid \mathcal{C}_{n-1}) \cup (1 \mid \mathcal{C}_{n-1})$

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

**Theorem.** Let  $n \geq 5$  be an odd integer. Then

$$M_2(n, \lfloor n/2 \rfloor - 1) = 2^{\lceil n/2 \rceil}$$

## Constructions of grain-correcting codes (cont.)

**Theorem.** *Let  $n \geq 4$  be an even integer. Then*

$$M_2(n, n/2 - 1) = 2^{n/2} + 2.$$

This result is attained by augmenting  $\mathcal{C}_n$  with  $(0110)^s(01)^{n-4s}$  and  $(1001)^s(10)^{n-4s}$  for  $s = \lfloor n/4 \rfloor$

**Theorem.** *Let  $m \geq 2$  be an integer and let  $n = 2m - 1$ . Then*

$$M_2(n, 1) \geq 2^{n-m} + 2^{(n-1)/2}$$

This result is obtained by the augmentation of a Hamming code with a subset of  $\mathcal{C}_n$

## Sizes $M_2(n, t)$ of largest $t$ -grain-correcting codes of length $n$

$t \backslash n$	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

**Sizes  $M_2(n, t)$  of largest  $t$ -grain-correcting codes in a wider sense of length  $n$**

$n \backslash t$	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

# Grain detection

- Sum of indexes  $\mathbf{s}(\mathbf{x})$  of the beginnings of runs of  $\mathbf{x}$
- Set of words  $\mathcal{F}$  with number of runs either  $\lfloor n/2 \rfloor$  or  $\lfloor n/2 \rfloor + 1$
- Categorize  $\mathcal{F}$  by  $\mathbf{s}(\cdot)$  modulo  $\lceil (n+1)/2 \rceil$  (modulo  $n$  when overlaps are allowed)

- There is at least one category  $\mathcal{C}_{\mathcal{F}} \subseteq \mathcal{F}$  of size  $\frac{2^{\left(\binom{n-1}{\lfloor n/2 \rfloor - 1} + \binom{n-1}{\lfloor n/2 \rfloor}\right)}}{\lceil (n+1)/2 \rceil}$  (of size  $\frac{2^{\left(\binom{n-1}{\lfloor n/2 \rfloor - 1} + \binom{n-1}{\lfloor n/2 \rfloor}\right)}}{n}$  when overlaps are allowed)

**Proposition.** *The code  $\mathcal{C}_{\mathcal{F}}$  is a binary  $\infty$ -grain-detecting code with rate  $1 - \frac{1.5 \log n}{n} + \frac{O(1)}{n}$*

- Easily generalized to larger alphabets ( $\mathcal{F}$  contains all the words with  $\sim n(q-1)/q$  runs)