Coding for Big Data: Error-Correction for Distributed Data Storage

P. Vijay Kumar

Professor,
Department of Electrical Communication Engineering
Indian Institute of Science, Bangalore

Bombay Information Theory Seminar (BITS'16)
IIT Bombay, Mumbai & TIFR, Mumbai

January 2, 2016

List of Collaborators

Presenting joint work with present and past IISc students:

1. Birenjith Sasidharan, Gaurav Agarwal, S. B. Balaji, K. P. Prasanth
2. Lalitha Vadlamani (now at IIIT Hyderabad)
3. Narayanamoorthy Prakash (now at MIT)

Data Centers and Storage

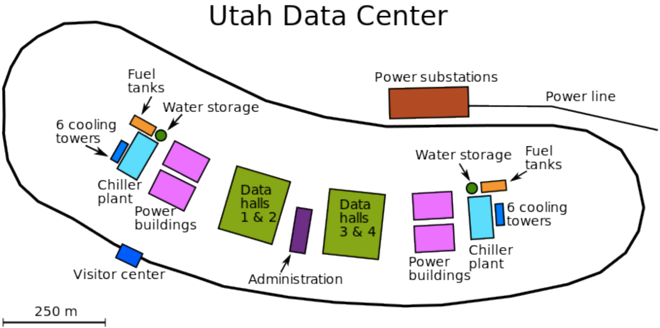
- Pictures from two different Data Centers..

Data Centers and Storage

Figure: The NSA Data Center in Utah.

- Estimated to store several between 3 to 12 Exabytes.

  GigaByte → TeraByte → PentaByte → ExaByte = One Billion GB!



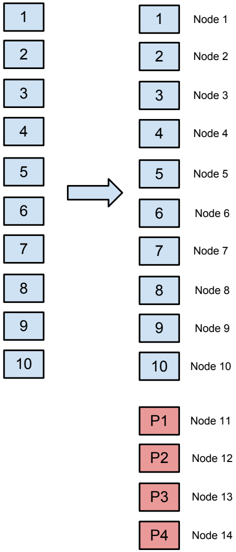
- Completed at an estimated cost of $1.5 billion..
- Another $2 billion for hardware, software, and maintenance
- 65 MW of power, costing about $40 million per year
- use 1.7 million gallons of water per day

A Data Center in Bengaluru

Figure: Bengaluru's Tulip Data City data center – estimated to be the third largest data center in the world at 900,000 square feet . (photo courtesy: http://www.forbes.com/pictures/eimh45mddm/tulip-data-center/).

Motivation

An Erasure-Correcting Code Employed by Facebook

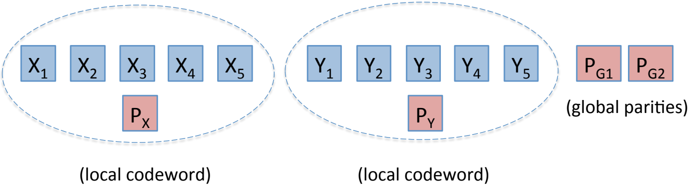


- $[14, 10]$ MDS code
- Can recover data by connecting to any 10 nodes
- Used in Facebook data centres
- If one disk is down, need contact (disturb!) 10 other disks to help fix it..
- Motivates the notion of locality of a code....

D. Borthakur, R. Schmit, R. Vadali, S. Chen, and P. Kling. "HDFS RAID." Tech talk. Yahoo Developer Network, Nov. 2010

Codes with Locality

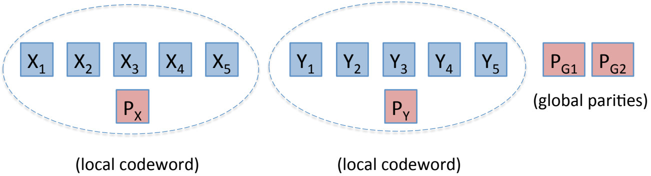
Codes with Information-Symbol Locality



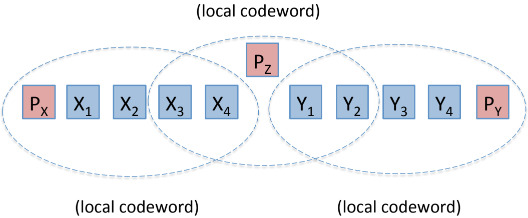
A code symbol has information-symbol locality r if all message symbols are protected by a single-parity-check code of

- blocklength $\leq (r+1)$
- the present code has information-symbol locality $r = 5$

Information and All-Symbol Locality



(information-symbol locality)



(all-symbol locality)

Bound on Global Minimum Distance

If an $[n, \kappa, d_{\min}]$ code $\mathcal{C}$ has information symbol locality r, then

$$\boxed{d_{\min} \leq \underbrace{(n - \kappa + 1)}_{\text{Singleton bound}} - \underbrace{\left(\left\lceil \frac{\kappa}{r} \right\rceil - 1 \right)}_{\text{price for locality requirement}} .}$$

- Bound established by P. Gopalan et al.

- P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, “On the Locality of Codeword Symbols,” IT-Trans, Nov. 2012, IEEE ComSoc and IT Society Joint Paper Award, 2013.

Optimal Constructions of Codes with Locality

Explicit Constructions

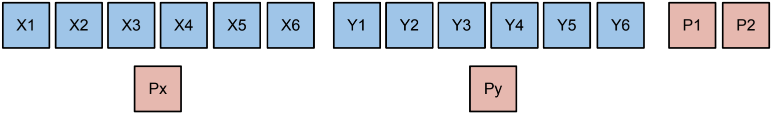
1. Pyramid Code construction for information locality.
2. Parity splitting construction for all symbol locality: $n = \left\lceil \frac{k}{r} \right\rceil (r + \delta - 1)$.
3. Rank-Distance based code with all-symbol locality : $\delta = 2$.
4. Tamo-Barg construction

Non-Explicit Construction All symbol locality codes can be constructed whenever

$$(r + \delta - 1) | n, \text{ provided } q > \binom{n-1}{k-1}$$

- C. Huang, M. Chen, and J. Li "Pyramid Codes: Flexible Schemes to Trade Space for Access Efficiency in Reliable Data Storage Systems," *NCA 2007*.
- J. Han, L. A. Lastras-Montano; , "Reliable Memories with Subline Accesses," *ISIT*- 2007.
- P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the Locality of Codeword Symbols," IT-Trans, Nov. 2012.
- N. Prakash, G. M. Kamath, V. Lalitha, and P. Vijay Kumar, "Optimal linear codes with a local-error-correction property," *ISIT*-2012.
- N. Silberstein, A. S. Rawat and S. Vishwanath, "Error Resilience in Distributed Storage via Rank-Metric Codes", *Allerton*, 2012.
- Itzhak Tamo and Alexander Barg, "A Family of Optimal Locally Recoverable Codes," ISIT 2014. (Subsequent journal paper won the IT-Trans. best paper award).

Windows Azure (Microsoft Azure) Storage Coding Solution

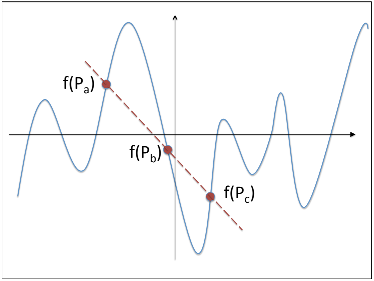


Employs a code related to the pyramid-code-type construction.

The code allows them to lower overhead without requiring a large number of disk I/O operations and network transfers to reconstruct the data, resulting in a saving to the company of millions of $$s!

C. Huang, H. Simitci, Y. Xu, A. Ogus, B. Calder, P. Gopalan, J. Li, and S. Yekhanin, "Erasure Coding in Windows Azure Storage," presented at the *USENIX Annu. Tech. Conf.*, Boston, MA, 2012.

The Tamo-Barg Construction

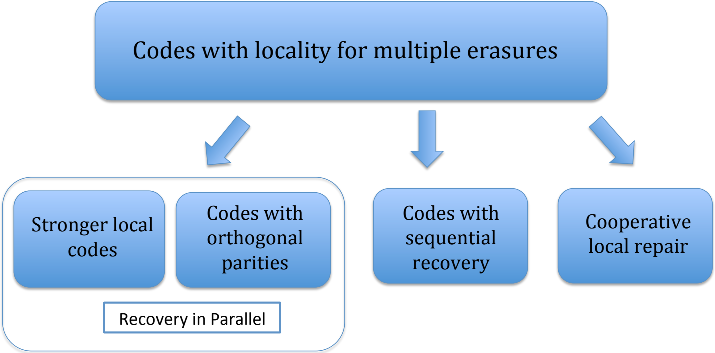


- Reed-Solomon (RS) codeword: $(f(P_1), f(P_2), \cdots, f(P_n))$, with $\deg(f) \leq (k-1)$
- authors employ subset of polynomials that restrict to a curve of degree $< (m-1)$ when evaluated at m points
- for example, to a line when evaluated at 3 points; this provides locality
- provides low-field-size constructions for many parameter sets
- There is also a Chinese Remainder Theorem interpretation

Codes with Locality for Multiple Erasures

Codes with Locality for Multiple Erasures

- Increasing trend towards low-cost commodity servers with higher failure rates
- Presence of "hot" nodes which are inaccessible during repair


Codes with locality for multiple erasures

Stronger local codes

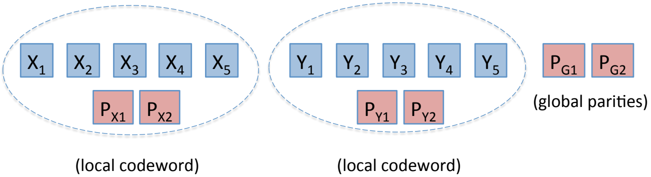
Codes with orthogonal parities

Recovery in Parallel

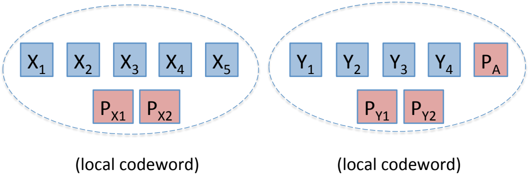
Codes with sequential recovery

Cooperative local repair

Handling Multiple Erasures: Stronger Local Codes Approach



(Information-Symbol locality)



More on 'Stronger Local Codes Approach'

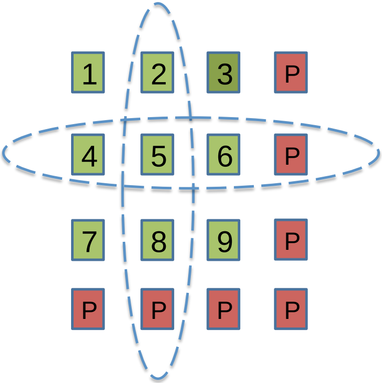
If an $[n, \kappa, d_{\min}]$ code $\mathcal{C}$ has information symbol locality r, then

$$\boxed{d_{\min} \leq \underbrace{(n - \kappa + 1)}_{\text{Singleton bound}} - \underbrace{\left(\left\lceil \frac{\kappa}{r} \right\rceil - 1\right)(\delta - 1)}_{\text{price for locality requirement}} .}$$

- Generalization of the Gopalan et al bound
- Pyramid code construction can be extended to this case as can the construction by Tamo and Barg
- More recent results by Wentu Song, Son Hoang Dau, Chau Yuen, and Tiffany Jing Li

- N. Prakash, G. Kamath, V. Lalitha, and P. Vijay Kumar, "Optimal linear codes with a local-error-correction property," in ISIT 2012.
- Optimal Locally Repairable Linear Codes, by Wentu Song, Son Hoang Dau, Chau Yuen, and Tiffany Jing Li.

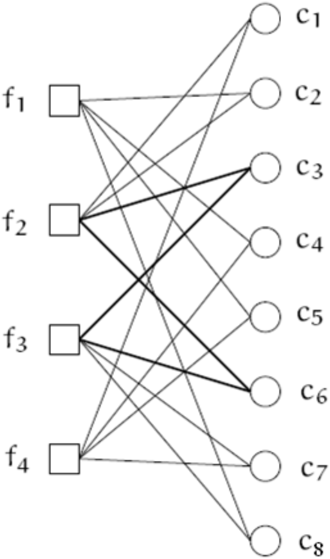
The Orthogonal Parity-Check Approach



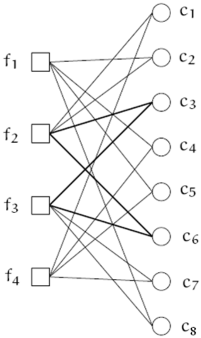
- Each data symbol is protected by two local codes with disjoint support
- All local codes are single-parity-check codes

A Computation Graph Approach to Codes with Locality for Multiple Erasures
- Falls into Orthogonal Parities Class

An Example (d_v, d_c)-Regular LDPC Code



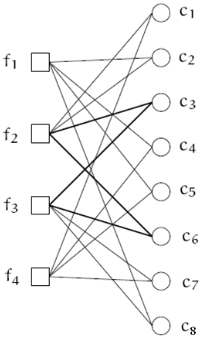
An Example (d_v, d_c)-Regular LDPC Code



Our interest is in those codes where

- each variable node has degree t
- each check node has degree $(r + 1)$
- there are no cycles of length 4

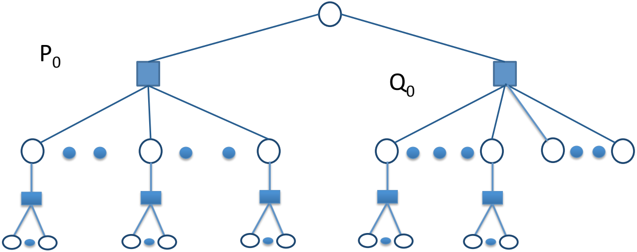
An Example (d_v, d_c)-Regular LDPC Code



This ensures that:

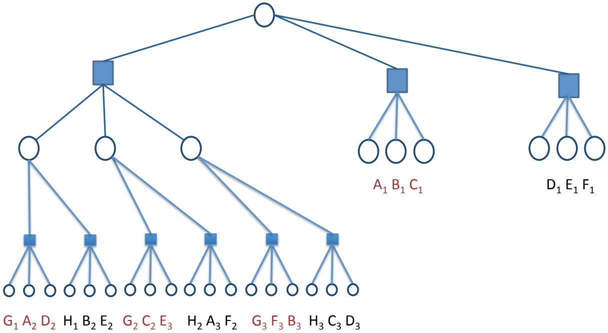
- each code symbol has locally r
- Each code symbol is protected by t orthogonal parity checks

A Canonical Form for the Computation Graph



- No repetition of parity nodes
- Variable nodes are permitted to repeat
- Expansion of the graph via top-to-bottom, left-to-right descent

Minimal-Length Code



Let

- m = number of parity checks
- $d_v = t$ the degree of each variable node
- $d_c = r + 1$ the degree of each parity node

Can be seen that

$$m \quad \geq \quad (t-1)(r+1)+1$$

Minimal-Length Code

Turns out that code symbols form a balanced-incomplete block design:

- (b, v, k, r, λ) – blocks are the code symbols and elements are parities
$$b \equiv n, \quad v \equiv m, \quad k \equiv t, \quad r \equiv (r+1), \quad \lambda \equiv 1.$$
- in the example,
$$b = 12, \quad v = 9, \quad k = 3, \quad r = 4, \quad \lambda = 1.$$
- note: all column inner products $= 1$ as required

		9 parity checks (elements)								
		1	2	3	4	5	6	7	8	9
12 code symbols (blocks)	0	1	1	1						
	1	1			1	1				
	2	1					1	1		
	3	1							1	1
	A		1		1			1		
	B		1			1			1	
	C		1				1			1
	D			1	1					1
	E			1		1	1			
	F			1				1	1	
	G				1		1		1	
	H					1		1		1

Minimal-Length Code

- allows us to provide constructions that are optimal (with unique parameters for certain values of r, t, n)

		9 parity checks (elements)								
		1	2	3	4	5	6	7	8	9
12 code symbols (blocks)	0	1	1	1						
	1	1			1	1				
	2	1					1	1		
	3	1							1	1
	A		1		1			1		
	B		1			1			1	
	C		1				1			1
	D			1	1					1
	E			1		1	1			
	F			1				1	1	
	G				1		1		1	
	H					1		1		1

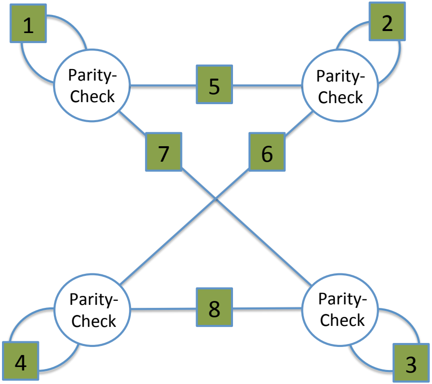
Codes based on designs

1. $t \mid r(r+1)$ is a necessary condition for the existence of designs.
2. For $t = r + 1 = 2^s + 1$, the code can be obtained from the point-line incidence matrix of $PG(2, 2^s)$. The code has parameters: $[2^{2s} + 2^s + 1, 2^{2s} + 2^s - 3^s, 2^s + 2]$ (This was observed by others as well). The code has rate $\geq \frac{r}{r+t}$.
3. For $t = r = 2^s$, the code can be obtained from an affine plane of order r. The code has parameters: $[2^{2s} + 2^s, 2^{2s} - 3^s, \geq 2^s + 1]$.

Codes based on designs

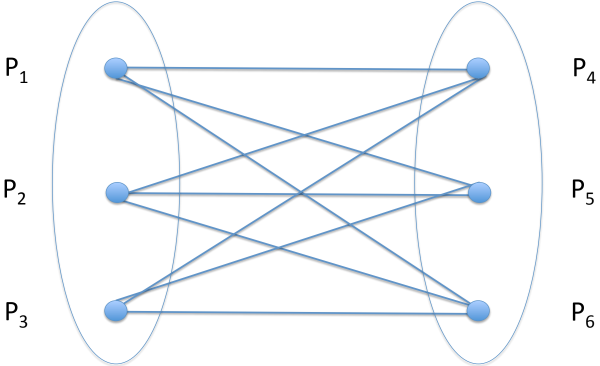
1. For $t = 3$, the code can be obtained from a Steiner triple system. In particular, the code can be obtained from the point-line incidence matrix of $PG(s-1,2)$. The code has parameters: Let $m = 2^s - 1$, $r + 1 = \frac{m-1}{2}$, $t = 3$,$[\frac{m(m-1)}{6}, \frac{m(m-1)}{6} - m + s, 4]$. This code also has rate $\geq \frac{r}{r+3}$
2. For the designs given above for $t = r+1, t = r, t = 3$, Hamada's conjecture has been proven, so the codes are rate-optimal for the given parameters.

The Sequential-Recovery Approach - An Example



The Sequential-Recovery Approach - A More General Turan-Graph Framework

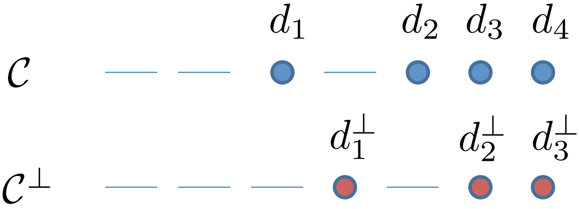
Turan Graph



(9 edges form remaining code symbols)
(n=15)

The Sequential-Recovery Approach - A More General Turan-Graph Framework

- The Turan graph construction has an additional feature that it leads to optimal solutions for smaller rates than the rate that arises from the constraints
- This is related to generalized Hamming weights

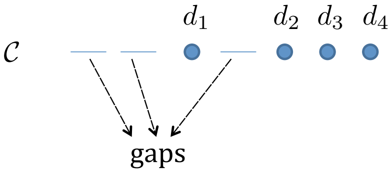


Generalized Hamming Weights of a Code

- d_i is the smallest support-size of an i dimensional subcode of $\mathcal{C}$
- $d_{\min} = d_1 < d_2 < \ldots < d_k = n$
- Example: Consider the $[7, 4, 3]$ Hamming code

$$G = \begin{bmatrix} 1 & & & & 0 & 1 & 1 \\ & 1 & & & 1 & 0 & 1 \\ & & 1 & & 1 & 1 & 0 \\ & & & 1 & 1 & 1 & 1 \end{bmatrix},$$

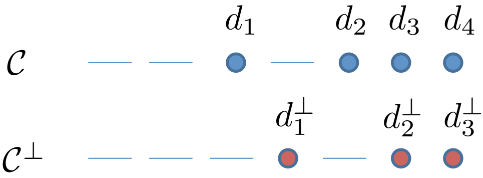
- Hamming code GHW : $d_1 = 3$, $d_2 = 5$, $d_3 = 6$, $d_4 = 7$



- k GHWs implies $n - k$ gaps

k^{th} Gap and Minimum Distance

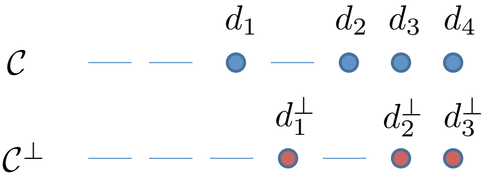
- $d_{min}(\mathcal{C}) \;=\; d_1 \;=\; n+1-g_k^{\perp}$, $g_k^{\perp}$ is largest gap of $\mathcal{C}^{\perp}$



- V.K. Wei, "Generalized Hamming Weights for Linear Codes," IEEE Trans. Inform. Th, 1991.

k^{th} Gap and Minimum Distance

- $d_{min}(\mathcal{C}) \;=\; d_1 \;=\; n+1-g_k^{\perp}$, $g_k^{\perp}$ is largest gap of $\mathcal{C}^{\perp}$

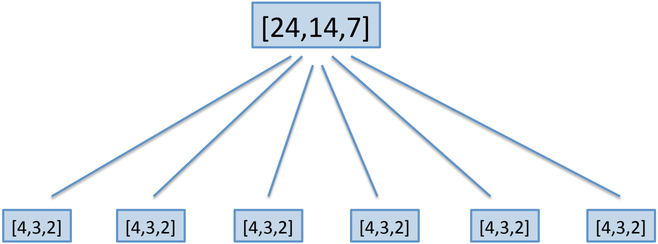


- V.K. Wei, "Generalized Hamming Weights for Linear Codes," IEEE Trans. Inform. Th, 1991.

Codes with Hierarchical Locality

Birenjith Sasidharan, Gaurav Kumar Agarwal, P. Vijay Kumar, "Codes With Hierarchical Locality," submitted to ISIT 2015, see also arXiv:1501.06683 [cs.IT]

Codes with Locality

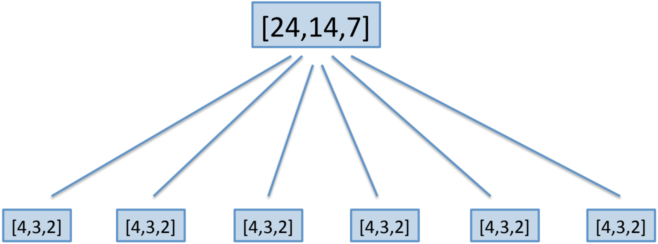


$$d \leq \underbrace{(n-k+1)}_{\text{Singleton bound}} - \underbrace{\left(\lceil \frac{k}{r} \rceil - 1\right)(\delta - 1)}_{\text{loss due to locality}}$$

r = locality

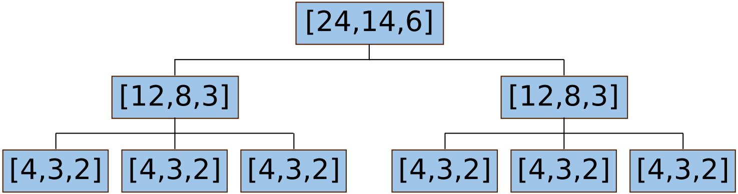
δ = minimum distance of the local code

Codes with Locality do not Scale



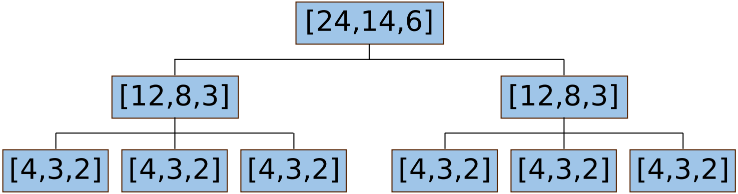
- If the local code is overwhelmed, then one has to appeal to the overall code which means contacting all 14 nodes for node repair.
- Is it possible to build a code where the repair degree increases gradually as opposed to in a single jump ?

Codes with Hierarchical Locality



- Codes with hierarchical locality do exactly that by calling for help from an intermediate layer of codes when the local code fails.
- These codes may be regarded as the “middle codes”.

Codes with Hierarchical Locality - Parameters

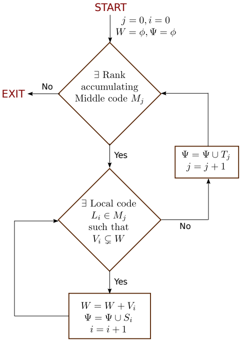


$$d \quad \leq \quad \underbrace{n - k + 1 - \left(\left\lceil \frac{k}{r_2} \right\rceil - 1 \right) (\delta_2 - 1)}_{\text{bound for codes with locality}} - \quad \underbrace{\left(\left\lceil \frac{k}{r_1} \right\rceil - 1 \right) (\delta_1 - \delta_2)}_{\text{additional loss for 2nd locality layer}}$$

Derivation of the Bound on Minimum Distance

- Proceeds along the lines of the original paper on codes with locality, shown below.

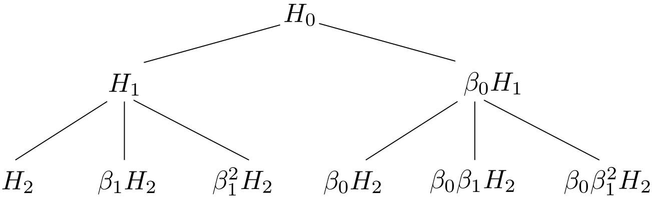
- Based on a recursive algorithm that searches for a large $(k \times \ell)$ sub-matrix of the generator matrix whose rank is $\leq (k-1)$.



- P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, “On the Locality of Codeword Symbols,” *IEEE Trans. Inf. Theory*, Nov. 2012.

All-symbol Local Optimal Construction: An Example

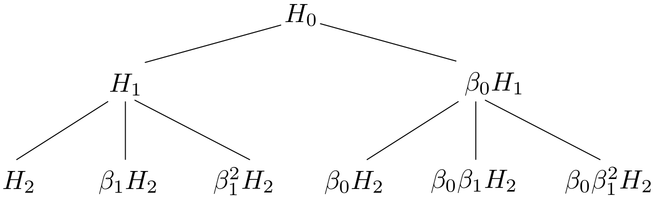
- Need to satisfy a divisibility condition $n_2 \mid n_1 \mid n$
- Example: $[24, 14]$, $[12, 8]$, $[4, 3]$.



1. Choose finite field $\mathbb{F}_{25}$.
2. The nonzero elements of $\mathbb{F}_{25}$ form cyclic subgroup of order 24.
3. Identify a cyclic subgroup chain $H_2 < H_1 < H_0 \ = \mathbb{F}_{25}^{(*)}$
4. Carry out a coset decomposition (as shown)
5. Each coset is now the support of a local code.

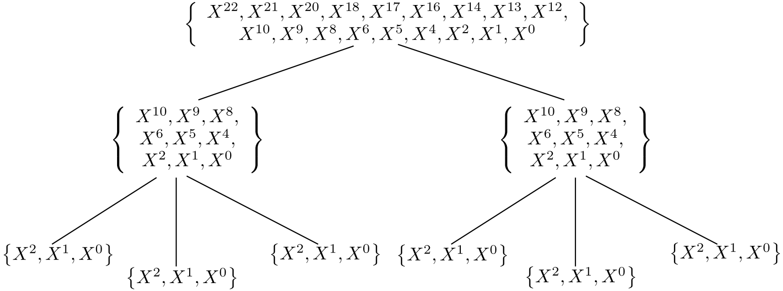
All-symbol Local Optimal Construction: An Example

- Need to satisfy a divisibility condition $n_2 \mid n_1 \mid n$
- Example: $[24, 14]$, $[12, 8]$, $[4, 3]$.



1. Choose $\mathbb{F}_{25}$.
2. Identify subgroup chain $H_2 \subseteq H_1 \subseteq H = \mathbb{F}_{25}^*$
3. Coset decomposition - supports of local codes

All-symbol Local Optimal Construction: An Example



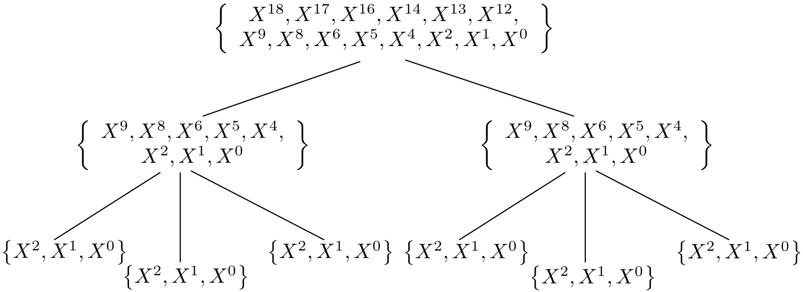
$\left\{ \begin{matrix} X^{22}, X^{21}, X^{20}, X^{18}, X^{17}, X^{16}, X^{14}, X^{13}, X^{12}, \\ X^{10}, X^{9}, X^{8}, X^{6}, X^{5}, X^{4}, X^{2}, X^{1}, X^{0} \end{matrix} \right\}$

$\left\{ \begin{matrix} X^{10}, X^{9}, X^{8}, \\ X^{6}, X^{5}, X^{4}, \\ X^{2}, X^{1}, X^{0} \end{matrix} \right\}$ $\left\{ \begin{matrix} X^{10}, X^{9}, X^{8}, \\ X^{6}, X^{5}, X^{4}, \\ X^{2}, X^{1}, X^{0} \end{matrix} \right\}$

$\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$

- The tree above shows the monomials appearing in the restriction of the code polynomial (its monomials appear on top) to each local code.

All-symbol Local Optimal Construction: An Example (continued)



$\left\{ \begin{array}{c} X^{18}, X^{17}, X^{16}, X^{14}, X^{13}, X^{12}, \\ X^{9}, X^{8}, X^{6}, X^{5}, X^{4}, X^{2}, X^{1}, X^{0} \end{array} \right\}$

$\left\{ \begin{array}{c} X^{9}, X^{8}, X^{6}, X^{5}, X^{4}, \\ X^{2}, X^{1}, X^{0} \end{array} \right\}$

$\left\{ \begin{array}{c} X^{9}, X^{8}, X^{6}, X^{5}, X^{4}, \\ X^{2}, X^{1}, X^{0} \end{array} \right\}$

$\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$ $\{X^2, X^1, X^0\}$

- The local codes can be tied together using an overall global code by simply restricting the set of code polynomials at the top. Here we do not allow the maximum degree to exceed 18. (The maximum was previously 22).

Thanks!