On the Locality of Codeword Symbols in Non-Linear Codes

Michael Forbes *
MIT
miforbes@mit.edu

Sergey Yekhanin
Microsoft Research
yekhanin@microsoft.com

February 6, 2014

Abstract

Coordinate i of an error-correcting code has locality r if its value is determined by some r other coordinates. Recently an optimal trade-off between information locality of linear codes, code distance, and redundancy has been obtained. Furthermore, for linear codes meeting this trade-off, structure theorems were derived. In this work we generalize the trade-off and structure theorems to non-linear codes.

1 Introduction

We say that a certain coordinate of an error-correcting code has locality r if, when erased, the value at this coordinate can be recovered by accessing at most r other coordinates. Motivated by applications to data storage [4] the authors of [3] introduced (r, d)-codes, which are systematic codes that have distance d and thus tolerate up to $d-1$ erasures, but also have the property that any information coordinate has locality r or less. They established that in all *linear* $[n, k, d]_q$ codes with the (r, d)-property

$$n \geq k + \left\lceil \frac{k}{r} \right\rceil + d - 2. \tag{1.1}$$

In what follows we refer to codes that meet (1.1) with equality as optimal. A construction of [2] implies that optimal codes exist for all values of parameters.

While locality of data symbols and code distance are the two primary considerations in the design of codes for data storage applications, locality of parity coordinates is also important. Parity locality (in the class of optimal (r, d)-codes) has been considered in [3]. In the natural setting of $r|k$, the lower-bound argument of [3] yields structure theorems for optimal linear codes. These theorems are particularly strong when $d < r + 3$. In that case they imply tight lower bounds for parity locality.

*Research was done while an intern at Microsoft Research.

Coding theory knows many examples of problems where non-linear codes improve upon the best available constructions of linear codes, e.g., [7]. While there is currently no evidence that non-linearity facilitates better (r, d)-codes, the novelty of this regime suggests that further study is required. In particular, it is natural to ask whether non-linearity can help reduce redundancy of (r, d)-codes or parity locality of optimal (r, d)-codes. The first question has been addressed in [5, 6] where the inequality (1.1) was generalized to non-linear setting under the stronger assumption that every code coordinate (and not just information coordinates) has locality r.

In this paper we strengthen the results of [5, 6] and establish the inequality (1.1) for general non-linear (r, d)-codes. We then use our lower-bound argument to derive structure theorems for optimal non-linear codes. Our results imply that lower bounds for parity locality of optimal (r, d)-codes that were derived in [3] in the linear setting also apply to non-linear codes. Therefore the answers to the two questions above are both negative.

1.1 Our techniques

Our new proof of the bound (1.1) follows the same high level strategy as the proofs in [3] and [6]. We assume that the bound (1.1) is violated and use an iterative argument to arrive at a code that violates the distance bound. Unlike [6], our iterative steps use elementary coordinate restrictions instead of entropy inequalities. This makes it easier to use our argument as a basis for structure theorems.

The main technical problem that we have to address going from the lower bound to structure theorems is that of reversibility of the local constraints. In linear codes, any local constraint on coordinates in the code must be a linear constraint, and linear constraints are trivially reversible, in that knowing all but 1 coordinate in the constraint always determines that coordinate, regardless of the identity of that 1 coordinate. However, for non-linear codes it is possible to have local constraints that are not reversible. For example, it is possible for the coordinates $\{i, i'\}$ to determine the coordinate i'', but for the coordinates $\{i', i''\}$ to not determine the coordinate i. However, we show that for optimal (r, d)-codes, even in the non-linear case, all locality constraints must be reversible. Once this is established, the structural results of [3] (and thus the parity locality lower bounds) can then be extended to the non-linear case.

2 Preliminaries

We will first fix some notation, then define the objects we will be considering.

2.1 Notation

Throughout, we consider codes which may be non-linear over an arbitrary alphabet Σ, where $|\Sigma| = q \geq 2$ is an arbitrary integer. Given two vectors $\vec{x}, \vec{y} \in \Sigma^n$, $\Delta(\vec{x}, \vec{y})$ will denote the unnormalized Hamming distance between $\vec{x}$ and $\vec{y}$. For an integer $n \geq 0$, $[n]$ denotes the set $\{1, \ldots, n\}$, where $[0]$ will be understood as the empty-set. For $S \subseteq [n]$, we will denote $\vec{x}|_S$

for the sequence of symbols in $\vec{x}$ with coordinates in S. When $S = \{i\}$ we will just write $\vec{x}|_i$. For disjoint sets A and B, we write $A \sqcup B$ to denote their disjoint union.

2.2 Definitions

Recall the definition of a code, which we do not assume to be linear.

Definition 2.1. *An $(n, K, d)_q$ code is a subset $\mathcal{C} \subseteq \Sigma^n$ with size $|\mathcal{C}| = K$, such that for any $\vec{x} \neq \vec{y} \in \mathcal{C}$, $\Delta(\vec{x}, \vec{y}) \geq d$. If $\mathcal{C}' \subseteq \mathcal{C}$ then $\mathcal{C}'$ is a **sub-code** of $\mathcal{C}$. The parameter n will be referred to as the **block-length**, $k = \log_q K$ the **dimension** and d the **distance**.*

*The code is **systematic** if $k \in \mathbb{Z}$, and there is an encoding function $\mathrm{Enc} : \Sigma^k \to \Sigma^n$ such that for $\vec{x} \in \Sigma^k$, $\mathrm{Enc}(\vec{x})|_i = \vec{x}|_i$, for $i \in [k]$.*

*A code is **maximum distance separable (MDS)** if $n = \log_q K + d - 1$.*

A systematic code takes on all q^k values in its first k coordinates, and the values of these coordinates determine the rest of the codeword. The first k coordinates of the codewords are thus referred to as the information symbols, other coordinates will be called parity symbols. This work will be interested in codes with local constraints on the information symbols.

Definition 2.2. *A systematic $(n, K, d)_q$ code has **information locality r** if for every $i \in [k]$, there is a size $\leq r$ subset $S \subseteq [n] \setminus \{i\}$ such that for any $\vec{x} \in \mathcal{C}$, $\vec{x}|_i$ is determined by $\vec{x}|_S$.*

Other symbols, other than the information symbols, can also have locality, and occasionally we will use this.

3 Locality Lower Bounds

In this section we establish lower bounds on the block-length of codes with small information locality. We will then prove structural results for codes meeting this lower bound. As we will often use it, we now prove the Singleton bound. Recall that MDS codes are those that exactly meet this bound.

Lemma 3.1 (Singleton Bound)**.** *Let $\mathcal{C}$ be an $(n, K, d)_q$ code, with $K > 1$. Then,*

$$n \geq \log_q K + d - 1$$

Proof. As $K > 1$, there are at least two distinct codewords. These codewords are at least d-apart. Thus $d \leq n$. Therefore we can delete the first $d-1$ coordinates from each codeword, resulting in a code $\mathcal{C}' \subseteq \Sigma^{n-d+1}$. The new code $\mathcal{C}'$ has distance ≥ 1, as each pair of original codewords have distance $\geq d$ and we only deleted $d-1$ coordinates. Thus, we have an injective map from $\mathcal{C}$ to Σ^{n-d+1}, and thus $\log_q K \leq n - d + 1$. □

The lower bounds that we derive for local codes will follow by analyzing Algorithm 1. In each step of the algorithm we will identify a new local constraint "$S_j \implies i_j$", so that the coordinates $S_j \subseteq [n]$ with $|S_j| \approx r$ determine an information coordinate $i_j \in [k] \setminus S_j$. We then examine the sub-code $\mathcal{C}_j \subseteq \mathcal{C} \subseteq \Sigma^n$ that results from taking all codewords $\vec{x} \in \mathcal{C}$ with

the restriction $\vec{x}|_S = \vec{\sigma}_j$ for some $\vec{\sigma}_j \in \Sigma^{|S_j|}$. By averaging, we show that $\vec{\sigma}_j$ can be chosen so $\mathcal{C}_j$ is not too small. In particular, the dimension of $\mathcal{C}_j$ will be $\approx |S_j|$ smaller than that of $\mathcal{C}$. However, because this sub-code $\mathcal{C}_j$ now fixes the coordinates in S, it also fixes the coordinate i_j, so the sub-code $\mathcal{C}_j$ has $\approx |S_j| + 1$ fixed coordinates, which we can discard. Thus the code $\mathcal{C}_j$ (after discarding fixed coordinates) has block length $\approx |S_j| + 1$ smaller than that of $\mathcal{C}$. Thus, each step will reduce the block length of the code by 1 more than it reduces the dimension of the sub-code. There will be $\approx k/r$ such steps, so that the gap between block-length and dimension will grow by $\approx k/r$. Further, we show that the distance remains d throughout this process, so then applying the Singleton bound to the final sub-code and incorporating the $\approx k/r$ gap yields (1.1) for the original code. We now make this process precise.

Algorithm 1 Finding sub-codes via locality

1: **procedure** SUB-CODE($\mathcal{C}, n, k, d, r, q$)
2: $\mathcal{C}_0 = \mathcal{C}$ ▷ "i_0" and "S_0" are not used
3: $j = 0$
4: **while** $|\mathcal{C}_j| > 1$ **do**
5: $j \leftarrow j + 1$.
6: Choose i_j such that $i_j \notin R_{j-1} := \bigcup_{j' \in [j-1]} (S_{j'} \cup \{i_{j'}\})$ and the size $\leq r$ subset of coordinates $S_j \subseteq [n] \setminus \{i_j\}$ determine the coordinate i_j, for all $\vec{x} \in \mathcal{C}$.
7: Let $\vec{\sigma}_j \in \Sigma^{|S_j|}$ be the most frequent element in the multi-set $\{\vec{x}|_{S_j} : \vec{x} \in \mathcal{C}_{j-1}\}$.
8: Define $\mathcal{C}_j := \{\vec{x} : \vec{x} \in \mathcal{C}_{j-1}, \vec{x}|_{S_j} = \vec{\sigma}_j\}$.
9: **end while**
10: **end procedure**

Theorem 3.2. *Let $\mathcal{C}$ be a systematic $(n, q^k, d)_q$ code with information locality r. Then*

$$n \geq k + \left\lceil \frac{k}{r} \right\rceil + d - 2.$$

Proof. We first show that Algorithm 1 is well-defined. In particular, in Line 6, such an i_j exists. For, by hypothesis $|\mathcal{C}_{j-1}| > 1$, implying there are $\vec{x} \neq \vec{y} \in \mathcal{C}_{j-1} \subseteq \mathcal{C}$. As $\mathcal{C}$ is systematic, such codewords are determined by their first k coordinates, and thus must differ on at least one of those k coordinates. Further, they will not differ on coordinates in R_{j-1}, as those coordinates are fixed, as we fixed the coordinates in $S_{j'}$ for $j' < j$ by construction in Line 8 and this also fixes the $\{i_{j'}\}_{j'<j}$ by locality. Thus, any of the first k coordinate where $\vec{x}$ and $\vec{y}$ differ will suffice for i_j. In Line 6 the set S_j exists by our locality assumption on the code $\mathcal{C}$ and using that $i_j \in [k]$. Thus, the algorithm is well-defined.

We now analyze the algorithm. We first show that each new sub-code is not too small. Define $T_j := S_j \setminus R_{j-1}$ and define $t_j := |T_j|$, which is the number of coordinates we fix in constructing $\mathcal{C}_j$ that are not necessarily fixed by prior loops in the procedure. It follows that there are $\leq q^{t_j}$ many possibilities for the $\vec{\sigma}_j$ in Line 7, and thus by averaging

$$|\mathcal{C}_j| \geq |\mathcal{C}_{j-1}|/q^{t_j}. \tag{3.3}$$

We now use (3.3) to lower-bound the number of iterations of the algorithm. Let ℓ denote the largest value of j such that $|\mathcal{C}_j| > 1$ in the algorithm, so that $|\mathcal{C}_{\ell+1}| = 1$. Applying (3.3) repeatedly shows that

$$0 = \log_q |\mathcal{C}_{\ell+1}| \geq k - \sum_{j=1}^{\ell+1} t_j, \tag{3.4}$$

and so as $t_j \leq |S_j| \leq r$, we have that

$$k \leq (\ell + 1)r,$$

or equivalently,

$$\ell \geq \left\lceil \frac{k}{r} \right\rceil - 1. \tag{3.5}$$

We now reduce to the Singleton bound. We first note that R_ℓ is the disjoint union $R_\ell = \bigsqcup_{j=1}^{\ell} (T_j \sqcup \{i_j\})$ and thus

$$|R_\ell| = \ell + \sum_{j=1}^{\ell} t_j. \tag{3.6}$$

Note that this decomposition of R_ℓ is indeed disjoint. That is, first observe that the T_j are disjoint by construction, and the i_j are distinct by construction. Further, the i_j are disjoint from the $T_{j'}$, which can be seen by case analysis. For $j = j'$, we have that $i_j \notin S_j \supseteq T_j$ by the definition of the locality constraint on i_j. For $j' > j$, we see that $i_j \notin T_{j'}$ by definition of $T_{j'}$. Finally, for $j' < j$, we have that $i_j \notin R_{j'} \supseteq T_{j'}$ by definition of i_j.

Now consider the code $\mathcal{C}_\ell \subseteq \mathcal{C} \subseteq \Sigma^n$, whose codewords all agree on R_ℓ, by construction. As it is a sub-code of $\mathcal{C}$, $\mathcal{C}_\ell$ also has minimum distance $\geq d$. It follows that we can delete the coordinates $R_\ell \subseteq [n]$ from $\mathcal{C}_\ell$, and bijectively map $\mathcal{C}_\ell$ to the new code $\mathcal{C}' \subseteq \Sigma^{n-|R_\ell|}$, which still has distance $\geq d$. Applying (3.3) iteratively and using the decomposition (3.6), it follows that

$$\log_q |\mathcal{C}'| = \log_q |\mathcal{C}_\ell| \geq \log_q |\mathcal{C}| - \sum_{j=1}^{\ell} t_j = \log_q |\mathcal{C}| - |R_\ell| + \ell. \tag{3.7}$$

Applying the Singleton bound (Lemma 3.1) to $\mathcal{C}'$ and then using (3.7) yields

$$n - |R_\ell| \geq \log_q |\mathcal{C}'| + d - 1 \geq (\log_q |\mathcal{C}| - |R_\ell| + \ell) + d - 1 = k - |R_\ell| + \ell + d - 1,$$

which by adding the quantity $|R_\ell|$ to both sides, yields

$$n \geq k + \ell + d - 1, \tag{3.8}$$

from which using (3.5) yields

$$n \geq k + \left\lceil \frac{k}{r} \right\rceil + d - 2. \qquad \square$$

We now turn to structural results on optimal local codes, that is, codes with information locality r that meet Theorem 3.2 with equality. For simplicity, we will restrict ourselves to the case that $r|k$ so that k/r is integral. We will first show the local structure for $r = k$, in which case the code is of parameters $(k + d - 1, q^k, d)_q$ and is thus a MDS code.

Lemma 3.9. *Let $\mathcal{C}$ be a $(k+d-1, q^k, d)_q$ code, that is, a MDS code. Then for any subset of coordinates $S \subseteq [k+d-1]$ of size k, the multi-set $\{\vec{x}|_S : \vec{x} \in \mathcal{C}\}$ takes on all values in Σ^k exactly once, and for any $\vec{x} \in \mathcal{C}$, $\vec{x}|_S$ determines $\vec{x}$.*

Proof. Suppose $\vec{x}, \vec{y} \in \mathcal{C}$ (possibly equal) have $\vec{x}|_S = \vec{y}|_S$. Then as $|S| = k$, it follows that $\Delta(\vec{x}, \vec{y}) \leq (k+d-1) - k = d-1 < d$. As the minimum distance of any two distinct codewords in $\mathcal{C}$ is $\geq d$, it follows that $\vec{x} = \vec{y}$, so $\vec{x}|_S$ determines $\vec{x}$.

Thus, as there are q^k codewords, and the values of the coordinates $\vec{x}|_S$ (taking values in Σ^k, and $|\Sigma^k| = q^k$) determine the codeword, it follows that each value in Σ^k is taken by $\vec{x}|_S$ exactly once, ranging over $\vec{x} \in \mathcal{C}$. □

This lemma shows that for each symbol, MDS codes have locality k and no less. When $r|k$ but $r < k$ the situation becomes more complicated, and we derive the following result (Theorem 3.10). The heart of this result is item (2), which establishes that in non-linear optimal (r, d)-codes, all local constraints will constrain the involved symbols equally, and are thus reversible. This fact is trivial for linear codes, but more involved for non-linear codes. Once established, the arguments can follow the case for linear codes as done in [3].

Theorem 3.10. *Let $\mathcal{C}$ be a systematic $(n, q^k, d)_q$ code with information locality r, with $r|k$ and $r < k$. Suppose $n = k + \frac{k}{r} + d - 2$. Let (possibly equal) information coordinates $i, i' \in [k]$, have associated subsets $S \subseteq [n] \setminus \{i\}$ and $S' \subseteq [n] \setminus \{i'\}$ of size $\leq r$, such that $\vec{x}|_S$ determines $\vec{x}|_i$ and $\vec{x}|_{S'}$ determines $\vec{x}|_{i'}$, for all $\vec{x} \in \mathcal{C}$. Then*

1. *$|S| = r$.*
2. *For all $i'' \in S \cup \{i\}$, $\vec{x}|_{(S \cup \{i\}) \setminus \{i''\}}$ determines $\vec{x}|_{i''}$, for all $\vec{x} \in \mathcal{C}$.*
3. *$S \cup \{i\}$ and $S' \cup \{i'\}$ are either equal or disjoint.*
4. *Up to a permutation of coordinates, $\mathcal{C}$ is a code with k information symbols I, k/r parities L, each depending on a disjoint set of r information symbols, and $d-2$ other parties H, depending arbitrarily on the k information symbols.*

Proof. The proof will be by analyzing particular runs of Algorithm 1, using the analysis of Theorem 3.2. By showing that certain inequalities in that analysis must be tight, we will derive the desired results.

We first establish further properties of the algorithm, in the case that $n = k + k/r + d - 2$, by extending the analysis given in Theorem 3.2. In particular, we highlight (3.8) and (3.4) respectively as

$$n \geq k + \ell + d - 1 \tag{3.11}$$

and

$$\sum_{j=1}^{\ell+1} t_j \geq k \tag{3.12}$$

where $t_j \leq |S_j| \leq r$. By the hypothesis that $n = k + k/r + d - 2$, we have that $\ell \leq k/r - 1 \in \mathbb{Z}$, from which it follows that

$$\sum_{j=1}^{\ell+1} t_j \leq (\ell+1)r \leq k \tag{3.13}$$

Combining (3.12) and (3.13) shows that these inequalities, and those inequalities used to derive them, must be met with equality. In particular, we have that $t_j = |S_j| = r$ for all j, and $\ell = k/r - 1$. From this, we can derive that the subsets $S_j \cup \{i_j\}$ in any run of Algorithm 1 must all have $r+1$ distinct elements, and further, that the family of subsets $\{S_j \cup \{i_j\}\}$ are disjoint.

Further, it follows that the inequalities used in the analysis of Theorem 3.2 to establish (3.11) must also be tight. In particular, $|\mathcal{C}_j| = |\mathcal{C}_{j-1}|/q^r$, for all j, implying that $|\mathcal{C}_j| = q^{k-rj}$, for all j. By the construction of $\mathcal{C}_j$ in Lines 7–8, it follows by an averaging argument that the multi-set $\{\vec{x}|_{S_j} : \vec{x} \in \mathcal{C}_{j-1}\}$ has q^r distinct elements (the maximum possible), each appearing equally often. In particular, this implies that any choice of $\vec{\sigma} \in \Sigma^r$ in Line 7 is valid, for any choice of the $\{\vec{\sigma}_{j'}\}_{j'<j}$ in the prior iteration of the loops. Further, once we have chosen these $k/r = \ell + 1$ values $\vec{\sigma}_j$, there is a unique codeword in $\vec{x} \in \mathcal{C}$ such that $\vec{x}|_{S_j} = \vec{\sigma}_j$ for all j, as $|\mathcal{C}_{k/r}| = 1$ by construction. It follows then that the values in the k coordinates $\sqcup_j S_j$ are completely independent, take on all q^k possible values, and uniquely determine a codeword in $\mathcal{C}$.

We will now use these facts applied to particular runs of the algorithm.

(1): Consider Algorithm 1 where we choose $i_1 \leftarrow i$ and $S_1 \leftarrow S$. The choice of i_1 is valid, for we are free to choose any $i_1 \in [n]$, as $R_1 = \emptyset$. The choice of S_1 is also valid, as S is a valid locality constraint on i_1. We then continue with Algorithm 1, to define the sets S_j and codes $\mathcal{C}_j$. From the analysis above, it follows that $|S_j| = r$ for all j, in particular $|S| = |S_1| = r$.

(2): For each $\vec{\tau} \in \Sigma^{r-1}$, consider the map $f_{\vec{\tau}} : \Sigma \to \Sigma$, such that $f_{\vec{\tau}}(\vec{x}|_{i''}) = \vec{x}|_i$ for all $\vec{x} \in \mathcal{C}$ such that $\vec{x}|_{S\setminus\{i''\}} = \vec{\tau}$. This is well-defined, as the coordinates $S = (S \setminus \{i''\}) \cup \{i''\}$ determine the coordinate i by locality, and the coordinates in S take on all q^r values by the analysis above, so for each $\vec{\tau}$, $f_{\vec{\tau}}$ must be defined for each input. The claim will be established by showing that, for each $\vec{\tau}$, this map is in fact a bijection, which will follow from showing that it is injective. To do this, we will analyze properties of sub-codes of $\mathcal{C}$, which will follow from analysis of Algorithm 1.

As in (1), we first run Algorithm 1 with $i_1 \leftarrow i$ and $S_1 \leftarrow S$, and the algorithm yields the k/r coordinates $\{i_j\}_{j\in[k/r]}$ and respective k/r locality constraints $\{S_j\}_{j\in[k/r]}$, such that the family of $(r+1)$-sized subsets $\{S_j \cup \{i_j\}\}_j$ are disjoint.

Now observe that choosing the $\{S_j \cup \{i_j\}\}_j$ in reverse order is also a valid run of Algorithm 1, because the conditions in Line 6 hold regardless of the order of j in the sequence $(S_j \cup \{i_j\})_j$. Consider the code $\mathcal{C}'$ resulting from the algorithm run in reverse order, where we have fixed the coordinates $\{S_j \cup \{i_j\}\}_{j>1}$, and so $\mathcal{C}'$ is the last code in this reverse-run algorithm that has more than one codeword. As such, $|\mathcal{C}'| = q^r$, by the analysis above. It has $n - (k/r - 1)(r+1) = r + d - 1$ non-fixed coordinates, and distance $\geq d$. The non-fixed coordinates include i, S, and the remaining $d-2$ coordinates [1]. This implies that $\mathcal{C}'$ (when the fixed coordinates are dropped) is an MDS code. In particular, consider the two sets of coordinates, S and $S \cup \{i\} \setminus \{i''\}$, in the code $\mathcal{C}'$. By Lemma 3.9, both of these sets of coordinates take on all possible values in Σ^r, and determine the other $d-1$ coordinates in $\mathcal{C}'$.

[1] Note that $d \geq 2$ is implied here. That $d = 1$ is impossible can be seen most easily in the case when $r = k$. For then, the parameters imply that the code is simply the list of all q^k words, and so no information locality is possible, for each coordinate is fully independent of the rest. That $d = 1$ is impossible for $r < k$ can be seen by reducing to a sub-code where $r = k$ via Algorithm 1.

We now show $f_{\vec{\tau}}$ is injective, for any $\vec{\tau} \in \Sigma^{r-1}$. Suppose $f_{\vec{\tau}}(\rho) = f_{\vec{\tau}}(\omega) = \sigma$, for some (possibly equal) $\rho, \omega, \sigma \in \Sigma$. As the coordinates S take on all q^r values in $\mathcal{C}'$, there are two codewords, $\vec{x}, \vec{y} \in \mathcal{C}'$ such that $\vec{x}|_{S\setminus\{i''\}} = \vec{y}|_{S\setminus\{i''\}} = \vec{\tau}$, $\vec{x}|_{i''} = \rho$ and $\vec{y}|_{i''} = \omega$. By the locality in $\mathcal{C}$, this means that $\vec{x}|_i = \vec{y}|_i = \sigma$. By the locality in $\mathcal{C}'$ just established, $\vec{x}|_{i''}$ is determined by $\vec{x}|_{S\setminus\{i''\}} = \vec{y}|_{S\setminus\{i''\}} = \vec{\tau}$ and $\vec{x}|_i = \vec{y}|_i = \sigma$, and thus $\rho = \vec{x}|_{i''} = \vec{y}|_{i''} = \omega$. Thus, $f_{\vec{\tau}}$ must be injective, for every $\vec{\tau}$.

(3): Suppose $S \cup \{i\}$ and $S' \cup \{i'\}$ are not equal, and we seek to show they are disjoint. Let i'' be the index of a coordinate in one of the sets but not the other, and without loss of generality, $i'' \in (S \cup \{i\}) \setminus (S' \cup \{i'\})$. Define $S'' := S \cup \{i\} \setminus \{i''\}$, and observe that by (2) applied to $i'' \in S \cup \{i\}$ implies that the coordinates in S'' determine the coordinate i''.

Now run Algorithm 1, choosing $i_1 \leftarrow i'$, $S_1 \leftarrow S'$, and $i_2 \leftarrow i''$, $S_2 \leftarrow S''$. The first round choice of coordinate/locality is clearly valid. That a second round will even by executed follows from the above analysis, as the code $\mathcal{C}_1$ has size $q^{k-r} > 1$, using that $r < k$. Further, the choices of coordinate/locality are valid because $i'' \notin R_1 = S' \cup \{i'\}$ and S'' determines i'' as established above. Thus, we have a valid initial run of the algorithm, which we can than run to completion to yield the family $\{S_j \cup \{i_j\}\}_j$. By the above analysis of the algorithm, it follows that the sets $\{S_j \cup \{i_j\}\}_j$ are disjoint. In particular, $S' \cup \{i'\}$ and $S'' \cup \{i''\} = S \cup \{i\}$ are disjoint, as desired.

(4): The above analysis shows that a run of Algorithm 1 yields the k/r disjoint sets of coordinates $\{S_j\}_j$ that take on all q^k values and uniquely determine the codeword. Treating these k coordinates as information symbols, and the corresponding coordinates $\{i_j\}_j$ as the k/r parities L, we get the desired reordering of the coordinates. □

The above analysis shows that the locality constraints are disjoint sets of size $r+1$, but does not indicate how many such constraints there are. In the case that $d < r+3$, we can show that there are exactly k/r such local constraints, and can characterize their structure. In fact, just as in the linear case in [3], we see that the locality structure of optimal (r,d)-codes resembles that of Pyramid codes [2].

From a practical standpoint the most important items below are items 3 and 4. They indicate that the lower bounds that have been known to hold for parity locality of optimal *linear* (r,d)-codes [3, Theorem 11] also apply to non-linear codes. These bounds are known to be tight [3, Theorem 15].

Theorem 3.14. *Let $\mathcal{C}$ be a systematic $(n, q^k, d)_q$ code with information locality r, with $r|k$ and $r < k$. Suppose $n = k + \frac{k}{r} + d - 2$ and $d < r + 3$. Then the $k/r + d - 2$ parity symbols can be partitioned into L and H, with $|L| = k/r$ and $|H| = d-2$, where*

1. *The parities in L, each depend on a disjoint subset of size r of the k information symbols.*
2. *The parities in H, each depend on all of the k information symbols.*
3. *The parities in L have locality exactly r.*
4. *The parities in H have locality $\geq k - (k/r - 1)(d-3) > r$.*

Proof. (1): We continue with the analysis given in the proof of Theorem 3.10. In particular, it implies that the n coordinates carry a family of subsets of size $r+1$, such that any r of the $r+1$ coordinates determine the other. By hypothesis, each information coordinate participates in such a subset, and the analysis in Theorem 3.10 shows that there are at least k/r such subsets. This leaves at most $n - k/r \cdot (r+1) = d-2$ coordinates that do not participate in these locality constraints, and as $d-2 < r+1$, there are in fact exactly k/r disjoint locality constraints and exactly $d-2$ coordinates not covered by locality constraints. Thus, there are $k/r \cdot (r+1) - k = k/r$ parity symbols participating in locality constraints, and let this set be L, which has the desired properties.

(2): We now show that the parities in $H := [n] \backslash ([k] \cup L)$ depend on each of the information symbols. Consider some information symbol $i \in [k]$, and consider $\sigma \neq \sigma' \in \Sigma$. Let $\vec{x}, \vec{x}' \in \Sigma^k$ be such that $\vec{x}|_i = \sigma$ and $\vec{x}'|_i = \sigma'$, and $\vec{x}|_j = \vec{x}'|_j$ for $i \neq j \in [k]$. It follows then that $\mathrm{Enc}(\vec{x}) \neq \mathrm{Enc}(\vec{x}')$, and these codewords agree in $(k-1) + k/r - 1$ places: they agree in $k-1$ information symbols by construction, and they agree in all but one of the k/r light parities (the light parity grouped with coordinate i being the exception). As the code has minimum distance d, and there are only d coordinates that these distinct codewords can differ on, it follows that $\mathrm{Enc}(\vec{x})$ and $\mathrm{Enc}(\vec{x}')$ differ on all these coordinates, in particular the $d-2$ heavy parities H. Thus, changing any information coordinate will change all heavy parities, showing that each coordinate in H depends on each of the k information coordinates.

(3): This follows from Theorem 3.10, as the light parities L can, under a permutation of coordinates, be regarded as information symbols, and thus cannot have locality $< r$.

(4): Let $[k] = \sqcup_{j \in [k/r]} I_j$ be the partition of the information symbols into size r subsets based on the grouping defined by the light parities. Pick arbitrary $\vec{\sigma}_j \in \Sigma^r$ for $j \in [k/r]$. Define the code $\mathcal{C}_j \subseteq \mathcal{C}$ to be $\mathcal{C}_j := \{\vec{x} : \vec{x} \in \mathcal{C}, \vec{x}|_{I'_j} = \vec{\sigma}_{j'}, j' \in [k/r] \setminus \{j\}\}$. It follows that in $\mathcal{C}_j$, all light parities are fixed except for the j-th. Thus, $\mathcal{C}_j$ has $(r+1) + (d-2) = r + d - 1$ unfixed symbols and has q^r codewords, as all but r information symbols are fixed. As $\mathcal{C}_j$ is a sub-code of $\mathcal{C}$, it follows that it has distance $\geq d$ and is thus is an MDS code.

Consider any heavy parity $h \in H$ determined by a set of coordinates $S \subseteq [n] \setminus \{h\}$ for all codewords in $\mathcal{C}$, and thus for all codewords in $\mathcal{C}_j$ for any j. By Lemma 3.9, we see that any r symbols in $\mathcal{C}_j$ are independent, so any locality constraint for h in $\mathcal{C}_j$ must involve at least r other symbols. As codewords in $\mathcal{C}_j$ are only unfixed on the coordinates I_j and H, it follows that $|S \cap (I_j \cup H \setminus \{h\})| \geq r$. Ranging this inequality over all $j \in [k/r]$, and accounting for double counting over $H \setminus \{h\}$, we see that $|S| \geq k - (k/r - 1)(d-3)$, as desired. □

References

[1] Minghua Chen, Cheng Huang, and Jin Li, "On maximally recoverable property for multi-protection group codes," *Proc. of IEEE International Symposium on Information Theory (ISIT)*, pp. 486-490, 2007.

[2] Minghua Chen, Cheng Huang, and Jin Li, "Pyramid codes: flexible schemes to trade space for access efficiency in reliable data storage systems," *Proc. of the Sixth IEEE International Symposium on Network Computing and Applications (NCA 2007)*, pp. 79-86, 2007.

[3] Parikshit Gopalan, Cheng Huang, Huseyin Simitci, and Sergey Yekhanin, "On the locality of codeword symbols," *IEEE Transactions on Information Theory,* vol. 58, pp. 6925-6934, 2012.

[4] Cheng Huang, Huseyin Simitci, Yikang Xu, Aaron Ogus, Brad Calder, Parikshit Gopalan, Jin Li, and Sergey Yekhanin, "Erasure coding in Windows Azure Storage," *Proc. of the 2012 USENIX conference on Annual Technical Conference*, pp. 2-2, 2012.

[5] Dimitris S. Papailiopoulos, and Alexandros G. Dimakis, "Locally repairable codes," *Proc. of IEEE International Symposium on Information Theory (ISIT)*, pp. 2771-2775, 2012.

[6] Maheswaran Sathiamoorthy, Megasthenis Asteris, Dimitris S. Papailiopoulos, Alexandros G. Dimakis, Ramkumar Vadali, Scott Chen, and Dhruba Borthakur, "XORing elephants: novel erasure codes for big data," *Arxiv*, abs/1301.3791, 2013.

[7] Chaoping Xing, "Nonlinear codes from algebraic curves improving the Tsfasman-Vladut-Zink bound," *IEEE Transactions on Information Theory,* vol. 49, pp. 1653-1657, 2003.