Westerbäck, Thomas; Freij-Hollanti, Ragnar; Hollanti, Camilla

Applications of polymatroid theory to distributed storage systems

Published in:
2015 53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015

DOI:
10.1109/ALLERTON.2015.7447009

Published: 04/04/2016

Document Version
Peer reviewed version

Please cite the original version:
Westerbäck, T., Freij-Hollanti, R., & Hollanti, C. (2016). Applications of polymatroid theory to distributed storage systems. In *2015 53rd Annual Allerton Conference on Communication, Control, and Computing, Allerton 2015* (pp. 231-237). [7447009] IEEE. https://doi.org/10.1109/ALLERTON.2015.7447009

Applications of Polymatroid Theory to Distributed Storage Systems

Thomas Westerbäck*
*Department of Mathematics and Systems Analysis
Aalto University, P.O. Box 11100
FI-00076 Aalto, Finland
(e-mails: firstname.lastname@aalto.fi)

Ragnar Freij-Hollanti† and Camilla Hollanti*
†Department of Communications and Networking
Aalto University, P.O. Box 13000
FI-00076 Aalto, Finland
(e-mail: firstname.lastname@aalto.fi)

***Abstract*—In this paper, a link between polymatroid theory and locally repairable codes (LRCs) is established. The codes considered here are completely general in that they are subsets of A^n, where A is an arbitrary finite set. Three classes of LRCs are considered, both with and without availability, and for both information-symbol and all-symbol locality. The parameters and classes of LRCs are generalized to polymatroids, and a generalized Singelton bound on the parameters for these three classes of polymatroids and LRCs is given. This result generalizes the earlier Singleton-type bounds given for LRCs. Codes achieving these bounds are coined *perfect*, as opposed to the more common term *optimal* used earlier, since they might not always exist. Finally, new constructions of perfect linear LRCs are derived from gammoids, which are a special class of matroids. Matroids, for their part, form a subclass of polymatroids and have proven useful in analyzing and constructing linear LRCs.**

I. Introduction

Within the past few years, *distributed storage systems* (DSSs) have revolutionized our traditional ways of storing, securing, and accessing data, and various big players like Facebook and Google nowadays provide their own cloud storage services. However, they do not come without regular failures, and hence have to be maintained by sophisticated repair processes. It has turned out that the number of nodes contacted for repair forms a bottle-neck in such vast data centers, calling for the notion of *locality*. In addition, clusters containing hot data, namely data that is frequently accessed simultaneously by many users, will benefit from multiple repair alternatives. This feature has further motivated the notion of *availability*.

A. Locally Repairable Codes

In this paper, we consider *locally repairable codes (LRCs)* with availability from the viewpoint of the interplay between its global parameters (n, k, d) and local parameters (r, δ, t). We will consider (n, k, d, r, δ)-LRCs, (n, k, d, r, δ, t)-LRCs, and $(n, k, d, r, \delta, t)'$-LRCs with 1-information-symbol locality, information-symbol locality, and/or all-symbol locality. These parameters and notions will be explained in detail in the sequel.

Let A be a finite set of size s and C a nonempty subset of A^n. Then we call C an (n, k)-code, where $k = \log_s(|C|)$. For $X = \{x_1, \ldots, x_l\} \subseteq [n] = \{1, \ldots, n\}$ and $\boldsymbol{z} \in A^n$, let $\boldsymbol{z}_X = (z_{x_1}, \ldots, z_{x_l})$. The *projection* of C into $A^{|X|}$ is defined as

$$C_X = \{\boldsymbol{c}_X = (c_{x_1}, \ldots, c_{x_l}) : \boldsymbol{c} \in C\,, |X| = l\}.$$

The *minimum (Hamming) distance* d of C can be defined as

$$d = \min\{|X| : X \subseteq [n] \text{ and } |C_{[n]\setminus X}| < |C|\}. \qquad (1)$$

In other words, for any $X \subseteq [n]$ with $|X| < d$, the symbols in X can be reconstructed by observing the symbols in $[n] \setminus X$ for every codeword in C, whereas for $|X| = d$ this is not necessarily true anymore.

We will consider two types of repair sets, (r, δ) and $(r, \delta)'$, where $r, \delta \in \mathbb{Z}$, $r \geq 1$, and $\delta \geq 2$. To this end, let $i \in [n]$ be a code symbol, equivalently a storage node, and $R \subseteq [n]$. The set $R \subseteq [n]$ is a *local repair set* with *repair locality* (r, δ) for the node i if

$$\begin{array}{ll} (i) & i \in R\,, \\ (ii) & |R| \leq r + \delta - 1\,, \\ (iii) & X \subseteq R \setminus \{i\}\,,\, |X| = |R| - (\delta - 1) \Rightarrow |C_X| = |C_R|\,. \end{array}$$

The set $R \subseteq [n]$ is a *local repair set* with *repair locality* $(r, \delta)'$ for the node i if the conditions (i) and (ii) above are satisfied, and in addition we have that

$$(iii)' \quad X \subseteq R\,,\, |X| = |R| - (\delta - 1) \Rightarrow |C_X| = |C_R|\,.$$

For (r, δ)-locality, the condition (iii) means that, for all codewords and subsets $X \subseteq R \setminus \{i\}$ such that $|X| \geq |R| - (\delta - 1)$, the symbols indexed by X are always sufficient to recover the symbol indexed by i. Also, the minimum distance of $C_{R\setminus\{i\}}$ is equal to or greater than $\delta - 1$.

For $(r, \delta)'$-locality, the condition $(iii)'$ means that, for all codewords and subsets $X \subseteq R \setminus \{i\}$ such that $|X| \geq |R| - (\delta - 1)$, the symbols indexed by X are always sufficient to recover the symbol indexed by i. Also, the minimum distance of C_R is equal to or greater than δ.

Further, a coordinate $i \in [n]$ has (r, δ, t)-*availability* (resp. $(r, \delta, t)'$-*availability*) if there are t local repair sets $R_1, \ldots, R_t$ for i with (r, δ)-locality (resp. $(r, \delta)'$-locality) such that

$$(iv) \quad j \neq l \Rightarrow R_j \cap R_l = \{i\}\,.$$

A subset $X \subseteq [n]$ has (r, δ, t)-availability (resp. $(r, \delta, t)'$-availability) if all elements $i \in X$ have (r, δ, t)-availability (resp. $(r, \delta, t)'$-availability).

A subset $K \subseteq [n]$ such that

$$|C_K| = |C| \text{ and } |C_{K\setminus\{i\}}| < |C|$$

for each element $i \in K$ is called an *information set*. This means that for any codeword the symbols indexed by K are enough to reconstruct all the other symbols of the codeword, but a strict subsets of these symbols is not.

Moreover, a *1-information set* K is an information set with the additional property that for every coordinate $i \in K$ and symbols $a, b \in C_{\{i\}}$,

$$|\{\boldsymbol{c} \in C : c_i = a\}| = |\{\boldsymbol{c} \in C : c_i = b\}|.$$

For example, a *systematic* (n, k)-code is a code for which k is an integer and there is an information set K of size k. This yields that K is a 1-information set where

$$|\{\boldsymbol{c} \in C : c_i = a\}| = |A|^{k-1},$$

for each $i \in K$ and symbol $a \in A$.

Let C be an (n, k, d)-code and $X \subseteq [n]$. Then C is an (n, k, d, r, δ)-LRC, (n, k, d, r, δ, t)-LRC, or $(n, k, d, r, \delta, t)'$-LRC over X if all elements in X have $(r, \delta, t = 1)$-locality, (r, δ, t)-availability, or $(r, \delta, t)'$-availability, respectively. If X is an information set, 1-information set, or $X = [n]$, then C has *information-symbol locality*, *1-information-symbol locality*, or *all-symbol locality*, respectively.

By a *linear* (n, k)-LRC we mean a subspace C of dimension k of $\mathbb{F}_q^n$, where $\mathbb{F}_q$ denotes the finite field of size q.

B. Related Work

There are several papers on different Singleton-type bounds for scalar, vector-linear, and nonlinear LRCs over finite fields, [1], [2], [3], [4], [5], [6] among others. Using entropy to analyze LRCs has, for example, been used in [4], [6], [7]. Further, combinatorial methods have been used for LRCs, *e.g.*, by the concept of regenerating sets [7] and matroids [8], [9]. For LRCs with availability, some constructions are proposed in [10], [6], [11].

C. Contributions and Organization

Every linear code has an associated matroid, however codes in general cannot be associated with a matroid. In this paper we extend the work in [9] on how matroids and linear LRCs are connected by associating any LRC over any finite set A with a polymatroid. Matroids are a subclass of polymatroids. Especially, we prove that the parameters associated with an LRC can be determined by its associated polymatroid. Moreover, we generalize the parameters associated with LRCs to polymatroids. Then, by using polymatroid theory, we get Singleton-type bounds for polymatroids which generalizes the bounds given in [1], [2], [3], [4], [5], [6] for LRCs. Moreover, these new bounds also give novel bounds on LRCs, since we simultaneously consider the parameters δ and t, alphabets that are not finite fields or finite vector spaces, and codes of size that is not a power of a prime.

A construction of linear LRCs with availablity is given, making use of an earlier result on a construction of linear LRCs from matroid theory in [9]. By this construction we are able to obtain a class of perfect linear LRCs with availability including all the parameters (r, δ, t). All the parameters for a class of perfect linear LRCs considering the availability $(r, \delta, t = 2)$ given in [10] are included in our construction.

In Section II, we give some fundamentals on polymatroid theory and entropy, and describe how codes $C \subseteq A^n$ can be associated to a polymatroids by the use of entropy. In Section III, the associated parameters of an LRC are generalized to polymatroids and bounds on these parameters for polymatroids and LRCs are given. In Section IV, a construction of linear LRCs is given which we then use to get a class of perfect LRCs with (r, δ, t)-availability.

II. Polymatroids and Codes

In this section we will show how (n, k)-codes can be associated to polymatroids via the notion of entropy. For more information on polymatroids, we refer the reader to [12].

A. Overview of Polymatroid Theory

For a finite set E, let 2^E denote the collection of all subsets of E. A pair $P = (\rho, E)$ is a (finite) *polymatroid* on E with a *set function* $\rho : 2^E \to \mathbb{R}$ if ρ satisfies the following three conditions for all subsets $X, Y \subseteq E$:

$$\begin{aligned} &(R1) \quad \rho(\emptyset) = 0\,, \\ &(R2) \quad X \subseteq Y \Rightarrow \rho(X) \leq \rho(Y)\,, \\ &(R3) \quad \rho(X) + \rho(Y) \geq \rho(X \cup Y) + \rho(X \cap Y)\,. \end{aligned}$$

A *matroid* is a polymatroid which additionally satisfies the following two conditions for all $X \subseteq E$:

$$\begin{aligned} &(R4) \quad \rho(X) \in \mathbb{Z}\,, \\ &(R5) \quad \rho(X) \leq |X|\,. \end{aligned}$$

For any polymatroid $P = (\rho, E)$ and $Y \subseteq E$ we obtain a new polymatroid $P_{|Y} = (\rho_{|Y}, Y)$, where

$$\rho_{|Y}(X) = \rho(X)$$

for any $X \subseteq Y$.

A polymatroid $P = (\rho, E)$ for which $\rho(\{x\}) \leq 1$ for all $x \in E$ is called a $1_{\leq}$-*polymatroid* throughout the paper. We say that two polymatroids $P = (\rho, E)$ and $P' = (\rho', E)$ on the same ground set E are equivalent if there is a constant $c \in \mathbb{R}$ such that $\rho(X) = c\rho'(X)$ for each $X \subseteq E$. Clearly, any polymatroid $P = (\rho, E)$ is equivalent to a $1_{\leq}$-polymatroid, wherefore we will only consider $1_{\leq}$-polymatroids for the rest of the paper. Note that if $P = (\rho, E)$ is a $1_{\leq}$-polymatroid, then $P_{|Y} = (\rho_{|Y}, Y)$ is also a $1_{\leq}$-polymatroid for every subset $Y \subseteq E$.

B. Some Basic Properties and Notions for $1_{\leq}$-polymatroids

The axioms (R1) and (R3) imply the following proposition.

Proposition 2.1: Let $P = (\rho, E)$ be a $1_{\leq}$-polymatroid. Then for any subset $X \subseteq E$ and element $x \in X$,

$$\begin{array}{ll} (i) & \rho(X) \leq |X|\,, \\ (ii) & 0 \leq \rho(X) - \rho(X \setminus \{x\}) \leq 1\,. \end{array}$$

Proof: The statement in (i) follows by induction on $|X|$: it is trivially true for $|X| = 0$; now let $y \in E \setminus X$. Then by the induction assumption and the axioms (R1) and (R3),

$$\rho(X \cup \{y\}) \leq \rho(X) + \rho(\{y\}) \leq |X| + 1.$$

For statement (ii), by axiom (R2), we immediately obtain that $0 \leq \rho(X) - \rho(X \setminus \{x\})$. Further, by the axioms (R1) and (R3),

$$\rho(X) - \rho(X \setminus \{x\}) \leq \rho(\{x\}) \leq 1.$$

■

By generalizing some notions from matroid theory we get the following corresponding notions for any $1_{\leq}$-polymatroid $P = (\rho, E)$ and $X \subseteq E$,

$$\begin{array}{ll} (i) & \eta(X) := |X| - \rho(X), \\ (ii) & \mathrm{cl}(X) := \{y \in E : \rho(X \cup \{y\}) = \rho(X)\}, \\ (iii) & X \text{ is a } \textit{flat} \text{ if } \mathrm{cl}(X) = X, \\ (iv) & X \text{ is } \textit{cyclic} \text{ if for all elements } x \in X, \\ & \rho(X) - \rho(X \setminus \{x\}) < 1. \end{array}$$

The collection of flats, cyclic sets, and cyclic flats of P are denoted by $\mathcal{F}$, $\mathcal{U}$ and $\mathcal{Z}$ respectively. Note that by definition $\emptyset \in \mathcal{U}$.

The following proposition will be needed for proving our generalized Singleton bound later on.

Proposition 2.2: Let $P = (\rho, E)$ be a $1_{\leq}$-polymatroid, then for any subsets $X, Y \subseteq E$,

$$\begin{array}{ll} (i) & \eta(X) \leq \eta(X \cup Y), \\ (ii) & \rho(\mathrm{cl}(X)) = \rho(X), \\ (iii) & X \subseteq Y \Rightarrow \mathrm{cl}(X) \subseteq \mathrm{cl}(Y), \\ (iv) & X \in \mathcal{F},\, x \in X,\, \rho(X \setminus \{x\}) < \rho(X) \Rightarrow \\ & (X \setminus \{x\}) \in \mathcal{F}, \\ (v) & X' \subseteq X \subseteq Y \Rightarrow \\ & \rho(X) - \rho(X \setminus X') \geq \rho(Y) - \rho(Y \setminus X'), \\ (vi) & X, Y \in \mathcal{U} \Rightarrow X \cup Y \in \mathcal{U}, \\ (vii) & X, Y \in \mathcal{U} \Rightarrow \mathrm{cl}(X \cup Y) \in \mathcal{Z}, \end{array}$$

Proof: For a proof of the results above we use some basic facts about polymatroids. A proof will appear in the journal version of this paper.

■

C. Codes and Entropy

We can associate any (n, k)-code with a random vector $\boldsymbol{Z} = (Z_1, \ldots, Z_n)$ with a joint probability distribution by

$$\Pr(\boldsymbol{Z} = \boldsymbol{z}) = \begin{cases} 1/|C| & \text{if} \quad \boldsymbol{z} \in C, \\ 0 & \text{if} \quad \boldsymbol{z} \notin C. \end{cases}$$

This gives, for the projections of the code, that

$$\Pr(\boldsymbol{Z}_X = \boldsymbol{z}_X) = |\{\boldsymbol{c} \in C : \boldsymbol{c}_X = \boldsymbol{z}_X\}|/|C|, \qquad (2)$$

where $X = \{x_1, \ldots, x_l\} \subseteq [n]$, $\boldsymbol{Z}_X = (Z_{x_1}, \ldots, Z_{x_l})$ and $\boldsymbol{z}_X \in A^{|X|}$. The *joint entropy* function of $\boldsymbol{Z}_X$ is then defined by using this probability as

$$H_C(\boldsymbol{Z}_X) = \sum_{\boldsymbol{z}_X \in A^{|X|}} \Pr(\boldsymbol{Z}_X = \boldsymbol{z}_X) \log_s \left(\frac{1}{\Pr(\boldsymbol{Z}_X = \boldsymbol{z}_X)} \right), \qquad (3)$$

where again $s = |A|$, and where we have the conventions that $0 \log_s 0 = 0$ and $H_C(\boldsymbol{Z}_\emptyset) = 0$.

D. Codes and Their Representations as Polymatroids

From [13] we have the following theorem.

Theorem 2.3: The joint entropy function H of any random vector $(Z_1, \ldots, Z_n)$ over some underlying probability space defines a polymatroid $P = (\rho, [n])$, where for any subset $X \subseteq [n]$

$$\rho(X) = H(\boldsymbol{Z}_X).$$

Hence, by (2) and (3), every (n, k)-code C over A^n induces a polymatroid $P_C = (\rho_C, [n])$ where

$$\rho_C(X) = H_C(\boldsymbol{Z}_X)\,.$$

By the above formula and the log sum inequality we obtain the following proposition.

Proposition 2.4: Let C be an (n, k)-code over A with $|A| = s$. Then for the polymatroid $P_C = (\rho_C, [n])$ and for subsets $X, Y \subseteq [n]$,

$$\begin{array}{ll} (i) & P_C \text{ is a } 1_{\leq}\text{-polymatroid}, \\ (ii) & |C_{X \cup Y}| > |C_X| \iff \rho_C(X \cup Y) > \rho_C(X), \\ (iii) & |C| = s^{\rho_C([n])}, \\ (iv) & |C|/|A^n| = s^{\rho_C([n]) - n}. \end{array}$$

Example 2.1: From this definition of a polymatroid $P_C = (\rho_C, [n])$ we get the following characterization of certain classes of codes. Let again A be a finite set of size s.

- Linear codes over $\mathbb{F}_q$: $\rho_C(X) = \log_q(|C_X|) = \mathrm{rank}(\text{generator matrix over column-set } X) \in \mathbb{Z}$,
- Almost affine codes: $\rho_C(X) = \log_q(|C_X|) \in \mathbb{Z}$,
- Vector-linear codes, *i.e.*, C is a linear subspace of A^n, where $A = \mathbb{F}_q^\alpha$): $\rho_C(X) = \log_{q^\alpha}(|C_X|) \in \mathbb{R}$,
- Quasi-uniform codes over A: $\rho_C(X) = \log_s(|C_X|) \in \mathbb{R}$,
- A general code $C \subseteq A^n$: $\rho_C(X) = H_C(\boldsymbol{Z}_X) \in \mathbb{R}$.

III. LRCs and Polymatroid Theory

A. Code Parameters for Polymatroids

A parameter of a code is *polymatroid invariant* if it only depends on its associated polymatroid, *i.e.*, always has the same value on two codes with the same associated polymatroid.

We claim that the parameters (n, k, d) of a code $C \subseteq A^n$, (r, δ, t)-availability and $(r, \delta, t)'$-availability of a code symbol, as well as information-set locality and 1-information-set locality, are all polymatroid invariant properties of C. This follows from the definitions of these properties and the set function ρ_C by using projections and Proposition 2.4. Hence, we can naturally generalize the typical code parameters to $1_{\leq}$-polymatroids.

Definition 3.1: Let $P = (\rho, E)$ be a $1_{\leq}$-polymatroid. Then

(i) $n = |E|$,
(ii) $k = \rho(E)$,
(iii) $d = \min\{|X| : \rho(E \setminus X) < \rho(E)\}$,
(iv) if we let $x \in E$ and $r, \delta \in \mathbb{Z}$, where $r \geq 1$ and $\delta \geq 2$, then x has (r, δ, t)-availability if there are t subsets $R_1, \ldots, R_t \subseteq E$ such that for $i, j \in [t]$:
  (a) $x \in R_i$,
  (b) $|R_i| \leq r + \delta - 1$,
  (c) $Y \subseteq R_i \setminus \{x\}$, $|Y| = |R_i| - (\delta - 1) \Rightarrow \rho(Y) = \rho(R_i)$,
  (d) $i \neq j \Rightarrow R_i \cap R_j = \{x\}$,
  Similarly, x has $(r, \delta, t)'$-availability if there are t subsets $R_1, \ldots, R_t \subseteq E$ such that the conditions (a), (b) and (d) above are satisfied, and in addition
  (e) $Y \subseteq R_i$, $|Y| = |R_i| - (\delta - 1) \Rightarrow \rho(Y) = \rho(R_i)$,
(v) $K \subseteq E$ is an information set if $\rho(K) = k$ and $\rho(K \setminus \{x\} < k)$ for all $x \in K$,
(vi) $K \subseteq E$ is a 1-information set if K is an information set and $\rho(x) = 1$ for all $x \in K$.

Let now $x \in E$ and $R \subseteq E$. If the conditions (a)-(c) above are satisfied by x and R then, similarly as for codes, R is called a *local repair set* with *repair locality* (r, δ) for x. Further, if the conditions (a), (b), and (e) above are satisfied by x and R, then R is again called a *local repair set* with *repair locality* $(r, \delta)'$ for x.

We remark that the values of the parameters (n, k, d), (r, δ, t), and $(r, \delta, t)'$ for a code C and a node i are the same as for the associated polymatroid P_C and its element i. Further, a coordinate set K for a code C is an information set (resp. 1-information set) if and only if the corresponding set of elements K in P_C is an information set (resp. 1-information set).

B. Code Parameters in Terms of Cyclic Flats

Let $P = (\rho, E)$ be a $1_{\leq}$-polymatroid. First we remark that d is well-defined for any nontrivial P, that is, for any P whose set function ρ is not the zero function. Second, if there is an element $x \in E$ such that x is not in any cyclic flat, then $\rho(E \setminus \{x\}) = \rho(E) - 1$. This implies that $\rho(X - \{x\}) = \rho(X) - 1$ for all $X \subseteq E$ with $x \in X$. This, for its part, implies that there are no repair sets for x. Further, let R be a repair set of y with repair locality (r, δ) (resp. $(r, \delta)'$) and $x \in R$. Then $R \setminus \{x\}$ is a repair set of y with repair locality $(r - 1, \delta)$ (resp. $(r - 1, \delta)'$). Consequently, we are only interested in $1_{\leq}$-polymatroids $P = (\rho, E)$ for which $k \neq 0$ and the union of cyclic flats, denoted by $1_{\mathcal{Z}}$, is the whole set E.

The following proposition gives a list of basic facts that will be needed later.

Proposition 3.1: Let $P = (\rho, E)$ be a $1_{\leq}$-polymatroid. Then for any element $x \in E$ and subsets $X, Y \in \mathcal{U}$,

(i) $if R$ is a repair set of x with (r, δ, t)-locality, then there is a repair set $Q \subseteq R$ of x with (r, δ, t)-availability,
(ii) $if R'$ is a repair set of x with $(r, \delta, t)'$-locality, then there is a repair set $Q' \subseteq R'$ of x with $(r, \delta, t)'$-availability,
(iii) $\mathrm{cl}(X) \in \mathcal{Z}$,
(iv) $\mathrm{cl}(X \cup Y) = \mathrm{cl}(\mathrm{cl}(X) \cup \mathrm{cl}(Y)) \in \mathcal{Z}$,
(v) $\rho(X) \leq |X| - (\delta - 1)$,
(vi) $\eta(X) \geq \delta - 1$,
(vii) $\rho(X \cup Y) \leq \rho(X) + \rho(Y) - \rho(X \cap Y)$,
$(viii)$ $\eta(X \cup Y) \geq \eta(X) + \eta(Y) - \eta(X \cap Y)$.

Proof: For a proof of the results above we use some basic facts about polymatroids. A proof will appear in the journal version of this paper. ∎

We are now ready to connect the parameters (n, k, d), (r, δ, t) and $(r, \delta, t)'$ of a polymatroid using cyclic flats.

Theorem 3.2: Let $P = (\rho, E)$ be a $1_{\leq}$-polymatroid with $k > 0$ and $1_{\mathcal{Z}} = E$. Then

(i) $n = |1_{\mathcal{Z}}|$,
(ii) $k = \rho(1_{\mathcal{Z}})$,
(iii) $d = \lfloor n - k + 1 - \max\{\eta(Y) : Y \in \mathcal{Z} \setminus \{1_{\mathcal{Z}}\}\} \rfloor$,
(iv) $x \in E$ has (r, δ, t)-availability if and only if there are t repair sets $R_1, \ldots, R_t \in \mathcal{U}$ with repair locality (r, δ), all of whose pairwise intersections equal $\{x\}$,
(v) $x \in E$ has $(r, \delta, t)'$-availability if and only if there are t repair sets $R'_1, \ldots, R'_t \in \mathcal{U}$ with repair locality $(r, \delta)'$, all of whose pairwise intersections equal $\{x\}$.

Proof: The statements (i) and (ii) follow directly from Definition 3.1. The statement (iv) follows from Proposition 3.1(i) and Definition 3.1. Similarly, statement (v) follows from Proposition 3.1(ii) and Definition 3.1.

For (iii), we first obtain that

$$\begin{aligned} d &= n - \max\{|Y| : Y \subseteq E,\, \rho(Y) < k\} \\ &= \max\{n - |Y| : Y \subseteq E,\, k - 1 \leq \rho(Y) < k\} \\ &= \max\{\lfloor n - |Y| - k + 1 + \rho(Y) \rfloor : Y \subseteq E,\, \rho(Y) < k\} \\ &= \lfloor n - k + 1 - \max\{\eta(Y) : Y \in \mathcal{Z} \setminus \{1_{\mathcal{Z}}\}\} \rfloor . \end{aligned} \tag{4}$$

In the equations above, the first equality is a consequence of (1), and the second of Axiom (R2) and Proposition 2.1(i). Further, for $X \subseteq E$,

$$\begin{aligned} &(i) \quad \eta(X) \leq \eta(\mathrm{cl}(X)), \\ &(ii) \quad x \in X,\, \rho(X \setminus \{x\}) = \rho(X) - 1 \Rightarrow \\ &\qquad \eta(X \setminus \{x\}) = \eta(X). \end{aligned} \tag{5}$$

Inequality (5)(i) is a consequence of Proposition 2.2. Now, by (4), (5) and by the fact that $\mathrm{cl}(\emptyset)$ is a cyclic flat and a subset of all flats, we obtain that

$$d = \lfloor n - k + 1 - \max\{\eta(Y) : Y \in \mathcal{Z} \setminus 1_{\mathcal{Z}}\} \rfloor .$$

∎

C. Generalized Singleton Bound for Polymatroids

We can define the above notions analogously for polymatroids. In this section, we will consider (n, k, d, r, δ)-polymatroids and $(n, k, d, r, \delta, t)'$-polymatroids with information-symbol locality, 1-information-symbol locality, and all-symbol locality, as well as (n, k, d, r, δ, t)-polymatroids with 1-information-symbol locality.

The approach in several parts of the proof of the following theorem is similar to the one used for the corresponding bound [10] for linear $(n, k, d, r, \delta = 2, t)$-LRCs with systematic information-symbol locality.

Theorem 3.3: Let $P = (\rho, E)$ be an (n, k, d, r, δ, t)-polymatroid with 1-information-symbol locality. Then

$$d \leq n - \lceil k \rceil + 1 - \left(\left\lceil \frac{(t(\lceil k \rceil - 1) + 1}{t(r-1)+1} \right\rceil - 1 \right) (\delta - 1).$$

Proof: Some parts of the proof are sketchy rather than rigorous. Let K be a 1-information set of P with (r, δ, t)-availability. By Theorem 3.2(iv), for each $x \in K$ there are t repair sets $R_1(x), \ldots, R_t(x) \in \mathcal{U}$ of x with (r, δ)-locality such that $i \neq j \Rightarrow R_i(x) \cap R_j(x) = \{x\}$. For $x \in K$ and $J \subseteq K$, let

$$R(x) = \bigcup_{i=1}^{t} R_i(x),\ Z(x) = \mathrm{cl}(R(x)) \text{ and } Z_J = \mathrm{cl}(\bigcup_{x \in J} Z(x).$$

By Proposition 3.1(iii) and (iv),

$$R(x),\ Z(x) \text{ and } Z_J \in \mathcal{Z}.$$

We claim for any $x \in K$ and $i \in [t]$ that

$$\begin{array}{ll} (i) & \rho(R_i(x)) \leq |R_i(x)| - (\delta - 1) \leq r, \\ (ii) & \eta(R_i(x)) \geq \delta - 1, \end{array}$$

For statement (i), by Proposition 2.1(i) and the definition of a repair set, we obtain that

$$\rho(R_i(x)) = \rho(Y) \leq |Y| = |R_i(x)| - (\delta - 1) \leq r$$

for any set $Y \subseteq R_i(x) \setminus \{x\}$ where $|Y| = |R_i(x)| - (\delta - 1)$. Statement (ii) follows directly from statement (i).

We claim for any $x \in K$, $i \in [t]$ and $I \subseteq [t] \setminus \{i\}$ that

$$\begin{array}{ll} (iii) & \rho(\bigcup_{l \in I} R_l(x)) \leq |I|(r-1) + 1, \\ (iv) & \eta(\bigcup_{l \in I} R_l(x)) \geq |I|(\delta - 1). \end{array}$$

Statement (iii) follows from induction on $|I|$. The statement follows from statement (i) for $|I| = 0$. Now, let $A = I \cup \{i\}$ and $y \in [t] \setminus A$, then by the induction assumption, and the axioms (R1) and (R3),

$$\begin{array}{rcl} \rho(\bigcup_{l \in (A \cup \{y\}} R_l(x)) & \leq & \rho(\bigcup_{l \in A} R_l(x) + \rho(R_y(x)) - \rho(x) \\ & \leq & |I|(r-1) + 1 + r - 1 \\ & = & (|I| + 1)(r-1) + 1. \end{array}$$

For statement (iv), by a similar argument as for statement (iii) above we have that

$$\rho(\bigcup_{l \in I} R_l(x)) \leq (\sum_{l \in I} \rho(R_l(x))) - 1.$$

Hence, (iv) follows from (ii) and (iii).

The property that $\rho(K) = k$ implies that $\rho(Z_K) = k$. Choose a subset $J = \{x_1, \ldots, x_j, x_{j+1}\} \subseteq K$ such that $\rho(Z_J) = k$ and $x_{i+1} \notin Z_{\{x_1, \ldots, x_i\}}$ for $1 \leq i \leq j$. For simplicity of notation, let $Z_{[i]}$ denote the cyclic flat $Z_{\{x_1, \ldots, x_i\}}$ for $1 \leq i \leq j+1$.

By Statements (iii) and (iv) and Proposition 2.2(i) and (ii), we immediately obtain that

$$\begin{array}{ll} (iv) & \rho(Z(x_i)) = \rho(R(x_i)) \leq t(r-1) + 1, \\ (v) & \eta(Z(x_i) \geq \eta(R(x_i)) \geq t(\delta - 1), \end{array}$$

for $i \in [j+1]$. Hence, for $1 \leq i \leq j$,

$$\begin{array}{ll} (vi) & \rho(Z_{[i+1]}) - \rho(Z_{[i]}) \leq t(r-1) + 1, \\ (vii) & \eta(Z_{[i+1]}) - \eta(Z_{[i]}) \geq t(\delta - 1). \end{array}$$

Statement (vi) is a consequence of (iv) and axiom (R3). Statement (vii) follows from the facts that $x_{i+1} \notin Z_{[i]}$. Consequently,

$$\begin{array}{ll} (viii) & \rho(Z_{[j]}) \leq j(t(r-1) + 1), \\ (ix) & \eta(Z_{[j]}) \geq jt(\delta - 1). \end{array}$$

For $0 \leq l \leq t$, let

$$Z_{j+1}^{l} = \mathrm{cl}(\bigcup_{i=1}^{l} R_i(x_{j+1})).$$

Now, let s be the integer in $[t]$ such that

$$\rho(\mathrm{cl}(Z_{[j]} \cup Z_{j+1}^{s-1})) < k \text{ and } \rho(\mathrm{cl}(Z_{[j]} \cup Z_{j+1}^{s})) = k.$$

The theorem now follows from similar arguments and enumerations as given in the proof of Theorem 1 i [7]. ∎

Theorem 3.4: Let $P = (\rho, E)$ be an (n, k, d, r, δ)-polymatroid with information-symbol, 1-information-symbol, or all-symbol locality. Then

$$d \leq n - \lceil k \rceil + 1 - \left(\left\lceil \frac{(t(\lceil k \rceil - 1) + 1}{t(r-1)+1} \right\rceil - 1 \right) (\delta - 1).$$

Proof: The proof of the results for information-symbol locality follow by similar argument as for Theorem 3.3. The theorem now follows from the facts that every (n, k, d, r, δ)-polymatroid with 1-information-symbol locality or all-symbol locality is an (n, k, d, r, δ)-polymatroid with information-symbol locality. ∎

Theorem 3.5: Let $P = (\rho, E)$ be an $(n, k, d, r, \delta, t)'$-polymatroid with information-symbol, 1-information-symbol or all-symbol locality. Then

$$d \leq n - \lceil k \rceil + 1 - \left(\left\lceil \frac{(t(\lceil k \rceil - 1) + 1}{t(r-1)+1} \right\rceil - 1 \right) (\delta - 1).$$

Proof: The proof of the results for information-symbol locality follow by similar argument as for Theorem 3.3. The theorem now follows from the facts that every $(n, k, d, r, \delta, t)'$-polymatroid with 1-information-symbol locality or all-symbol locality is an $(n, k, d, r, \delta, t)'$-polymatroid with information-symbol locality. ∎

D. Corollaries for LRCs

From Theorems 3.3, 3.4, and 3.5 we immediately get the following corollary for LRCs.

Corollary 3.6: Let C be an (n, k, d, r, δ, t)-LRC with 1-information-symbol locality. Then

$$d \leq n - \lceil k \rceil + 1 - \left(\left\lceil \frac{(t(\lceil k \rceil - 1) + 1}{t(r-1)+1} \right\rceil - 1 \right) (\delta - 1).$$

Corollary 3.7: Let C be an (n, k, d, r, δ)-LRC with 1-information-symbol, information-symbol or all-symbol locality. Then

$$d \leq n - \lceil k \rceil + 1 - \left(\left\lceil \frac{(t(\lceil k \rceil - 1) + 1}{t(r-1)+1} \right\rceil - 1 \right) (\delta - 1).$$

Corollary 3.8: Let C be an $(n, k, d, r, \delta, t)'$-LRC with 1-information-symbol, information-symbol or all-symbol locality. Then

$$d \leq n - \lceil k \rceil + 1 - \left(\left\lceil \frac{(t(\lceil k \rceil - 1) + 1}{t(r-1)+1} \right\rceil - 1 \right) (\delta - 1).$$

One remark on the bounds given above is that, if all the parameters (n, r, δ, t) are fixed as well as the alphabet size s, then the bound for d always increases when the number of codewords goes from s^k to $s^k + 1$. This, for example, implies that if there is a linear LRC that achieves some bound given above and there is a nonlinear LRC with the same parameters on (n, r, δ, t) but with a better rate then the nonlinear LRC will always have a smaller d than the linear LRC.

Further, there are many polymatroids which cannot be realised as a polymatroid P_C of any code $C \subseteq A^n$. For example the nonentropic polymatroids. Hence, the bounds given for polymatroids above are valid for many other types of polymatroids than just the P_C-polymatroids. The same is true for matroids, many of which are not representable by a linear code. In general, it is extremely hard to determine whether a given matroid is representable (over any field). It is conjectured, but to the best of the authors' knowledge not yet proven, that

$$\lim_{n \to \infty} \frac{|\{\text{Representable matroids on } n \text{ elements}\}|}{|\{\text{Matroids on } n \text{ elements}\}|} = 0.$$

Moreover, there are many non-code objects that can be associated to matroids or polymatroids, *e.g.*, graphs, hypergraphs, matchings, and designs. The bounds given above for polymatroids also give us results for all these additional objects.

IV. Constructions of Perfect linear (n, k, d, r, δ, t)-LRCs

Typically (Singleton-type) bound-achieving codes have been referred to as optimal. However, we rather choose to use the term perfect, since there might not always exist codes achieving the bound. However, in our interpretation, *optimal* should always refer to the best option one can possibly have. Hence it feels wrong to us to say that no optimal code exists, even though there would be a code that almost achieves the bound and is known to be the best possible code. To this end, we give the following definition.

Definition 4.1: We will call an (n, k, d, r, δ, t)-polymatroid or (n, k, d, r, δ, t)-LRC which achieves the bounds given above *perfect*.

In [9] a construction of linear LRCs is derived from matroid theory. This construction was used in [9] to obtain linear $(n, k, d, r, \delta, t = 1)$-LRCs with all-symbol locality that are perfect or near-perfect. We summarize the construction in the following.

A construction of matroids 4.1 ([9]): Let $F_1, \ldots, F_m$ be subsets of a finite set E, k a non-negative integer and $\rho : \{F_i\}_{i \in [m]} \to \mathbb{Z}$ a function such that

$$\begin{array}{ll} (i) & 0 < \rho(F_i) < |F_i|, \\ (ii) & k \leq |F_{[m]}| - \sum_{i=1}^{m} (\eta(F_i)), \\ (iii) & |F_{[m] \setminus \{i\}} \cap F_i| < \rho(F_i) \text{ for all } i \in [m], \end{array} \tag{6}$$

where for every element $i \in [m]$ and subset $I \subseteq [m]$

$$\begin{array}{ll} (a) & \eta(F_i) = |F_i| - \rho(F_i), \\ (b) & F_I = \bigcup_{i \in I} F_i. \end{array}$$

Further, for every subset $I \subseteq [m]$, define

$$\rho(F_I) = \min\{|F_I| - \sum_{i \in I} \eta(F_i), k\} \text{ and } \rho(E) = k.$$

Theorem 4.1 ([9]): Let $F_1, \ldots, F_m$ be subsets of a finite set E, k a non-negative integer and $\rho : \{F_i\}_{i \in [m]} \to \mathbb{Z}$ a function such that the conditions (i)-(iii) in (6) are satisfied. Then the set-construction defines a matroid $M_{\mathcal{Z}} = (\rho_{\mathcal{Z}}, E)$ where

$$\begin{array}{ll} (i) & \mathcal{Z} = \{F_I : I \subseteq [m],\, \rho(F_I) < k\} \cup E, \\ (ii) & \rho_{\mathcal{Z}}(X) = \min\{\rho(F) + |X \setminus F| : F \in \mathcal{Z}\}, \\ (iii) & n = |E|, \\ (iv) & k = \rho(E), \\ (v) & d = n - k + 1 - \max\{\eta(F) : F \in \mathcal{Z} \setminus \{E\}\}, \\ (iv) & F_i \text{ is a repair set with} \\ & (r = \rho(F_i), \delta = \eta(F_i) + 1)\text{-locality for} \\ & \text{every element in } F_i, \\ (iv) & \text{a subset } K \subseteq [n] \text{ is an information set of } M_{\mathcal{Z}} \iff \\ & |K| = k \text{ and } |K \cap F| \leq \rho(F) \text{ for all } F \in \mathcal{Z}. \end{array}$$

Theorem 4.2 ([9]): Every matroid $M_{\mathcal{Z}}$ given from Theorem 4.1 is in a class of matroids called gammoids.

We say that a matroid is *representable over a finite field* $\mathbb{F}_q$ if the matroid can be represented by a linear code over $\mathbb{F}_q$.

Theorem 4.3 ([14]): Every gammoid over a finite set E is representable over every finite field of size greater then or equal to $2^{|E|}$.

We remark that $2^{|E|}$ is just an upper bound on the smallest field size of a linear code that can be used to represent a gammoid. It is possible that a gammoid may be represented by a linear code over a field with much less size than $2^{|E|}$.

The following theorem was derived in [9], by use of the theorems above.

Theorem 4.4 ([9]): Every matroid $M_{\mathcal{Z}}$ given in Theorem 4.1 is isomorphic to $M_C = (\rho_C, [n])$, for some linear code C over a large enough field.

Using Construction 4.1 and Theorem 4.1 we are now able to construct (n, k, d, r, δ, t)-matroids M_C. To obtain the actual linear (n, k, d, r, δ, t)-LRC associated to M_C we can use [14] in which it is described how to derive a linear code associated to a gammoid.

Example 4.1: Construction of a perfect (n, k, d, r, δ, t)-matroid M_C for a linear code C.

Let $E = [36]$, $k = 4$,

$$\begin{array}{lllllll} F_1 & = & \{1, 5-8\}, & F_2 & = & \{1, 9-12\}, \\ F_3 & = & \{2, 13-16\}, & F_4 & = & \{2, 17-20\}, \\ F_5 & = & \{3, 21-24\}, & F_6 & = & \{3, 25-28\}, \\ F_7 & = & \{4, 29-32\}, & F_8 & = & \{4, 33-36\}, \end{array}$$

and $\rho(F_i) = 3$ for $i \in [8]$. Then, by Theorem 4.1,

(i) $\mathcal{Z} = \{\emptyset, F_1, \ldots, F_8, [36]\}$,
(ii) $K = \{1, 2, 3, 4\}$ is an information set,
(iii) for $i \in [8]$, F_i is a repair set with $(r = 3, \delta = 3)$-locality for every element $x \in F_i$,
(iv) $d = 36 - 4 + 1 - 2 = 31$,
(v) K has $(r = 3, \delta = 3, t = 2)$-locality.

By Theorem 4.4 and Corollary 3.6, the construction above defines a perfect linear $(36, 4, 31, 3, 3, 2)$-LRC with information-symbol locality since

$$36 - 4 + 1 - (\left\lceil \left(\frac{2(4-1)+1}{2(3-1)+1} \right\rceil - 1 \right) (3-1)) = 31 = d.$$

Theorem 4.5: If $n \geq k(t(r + \delta - 2) + 1))$, then there is a perfect linear (n, k, d, r, δ, t)-LRC with information-symbol locality.

Proof: For a proof of the results above we use the same kind of construction given in the example above. A proof will appear in the journal version of this paper. ■

ACKNOWLEDGMENTS

This work was partially supported by the Academy of Finland grants #276031, #282938, and #283262, and by a grant from Magnus Ehrnrooth Foundation, Finland. The support from the European Science Foundation under the ESF COST Action IC1104 is also gratefully acknowledged.

REFERENCES

[1] R. C. Singleton, "Maximum distance q-nary codes", *IEEE Trans. Inf. Theory*, 10, pp. 116–118, 1964.
[2] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the locality of codeword symbols," *IEEE Trans. Inf. Theory*, 58(11), pp. 6925–6934, September 2012.
[3] N. Prakash, G. M. Kamath, V. Lalitha, and P. V. Kumar, "Optimal linear codes with a local-error-correction property," *2012 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 2776 – 2780, 2012.
[4] D. S. Papailiopoulos, and A. G. Dimakis, "Locally repairable codes," *2012 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 2771–2775.
[5] N. Silberstein, A. S. Rawat, O. O. Koyluoglu, and S. Vishwanath, "Optimal locally repairable codes via rank-metric codes," *2013 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1819–1823.
[6] A. S. Rawat, D. S. Papailiopoulos, A. G. Dimakis, and S. Vishwanath "Locality and availability in distributed storage", *arXiv: 1402.2011v1*, 2014.
[7] A. Wang and Z. Zhang, "Repair locality from a combinatroial perspective" *2014 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1972–1976.
[8] I. Tamo, D. S. Papailiopoulos, A. G. Dimakis, "Optimal locally repairable codes and connections to matroid theory," *2013 IEEE Int. Symp. Inf. Theory (ISIT)*, pp. 1814–1818.
[9] T. Westerbäck, R. Freij, T. Ernvall, C. Hollanti "On the combinatorics of locally repairable codes via matroid theory", *arXiv:1501.00153*, 2015.
[10] A. Wang and Z. Zhang, "Repair locality with multiple erasure tolerance" *IEEE Trans. Inf. Theory*, 60(11), pp. 6979–6987, 2014.
[11] A. Wang and Z. Zhang, "Achieving arbitrary locality and availability in binary codes", arXiv:1501.04264v1, 2015.
[12] J. Oxley, "Matroid Theory" *2:ed, Oxford Graduate Texts in Mathematics, 21*. OxfordUniversity Press, 2011.
[13] S. Fujishige, "Polymatroidal dependence structure of a set of random variables," *Information and control*, 39(1), pp. 55–72, 1978.
[14] B. Lindström, "On the vector representations of induced matroids" *Bull. London Math. Soc.*, 5, pp. 85–90, 1973.