Partial MDS Codes with Regeneration

Lukas Holzbaur*, Sven Puchinger†, Eitan Yaakobi‡, and Antonia Wachter-Zeh*
*Technical University of Munich, {lukas.holzbaur, antonia.wachter-zeh}@tum.de
†Technical University of Denmark, svepu@dtu.dk
‡Technion — Israel Institute of Technology, yaakobi@cs.technion.ac.il

***Abstract*—Partial MDS (PMDS) and sector-disk (SD) codes are classes of erasure correcting codes that combine locality with strong erasure correction capabilities. We construct PMDS and SD codes where each local code is a bandwidth-optimal regenerating MDS code. In the event of a node failure, these codes reduce both, the number of servers that have to be contacted as well as the amount of network traffic required for the repair process. The constructions require significantly smaller field size than the only other construction known in literature. Further, we present a PMDS code construction that allows for efficient repair for patterns of node failures that exceed the local erasure correction capability of the code and thereby invoke repair across different local groups.**

I. Introduction

Distributed data storage is of ever increasing importance with the amount of data stored by cloud service providers and data centers in general reaching staggering heights. The data is commonly spread over a number of nodes (servers or hard drives) in a *distributed storage system* (DSS), with some additional redundancy to protect the system from data loss in the case of node failures (erasures). The resilience of a DSS against such events can be measured either by the minimal *number of nodes* that needs to fail for data loss to occur, i.e., the *distance* of the storage code, or by the expected time the system can be operated before a failure occurs that causes data loss, referred to as the *mean time to data loss*. For both measures the use of maximum distance separable (MDS) codes provides the optimal trade-off between storage overhead and resilience to data loss (note that replication is a trivial MDS code). However, MDS codes require a large number of nodes to be involved in the recovery of even a single node and straight-forward methods of recovery induce a large amount of network traffic. To address these issues, the concepts of *locally repairable codes* (LRCs) [3] and *regenerating codes* [4] have been introduced.

The latter allows for a larger number of nodes to be accessed in recovery, but only a function of the data stored on each node is retrieved, which can significantly decrease the repair traffic. The lower bounds on the required traffic for repair derived in [4] lead to two extremal code classes, namely *minimum bandwidth regenerating* (MBR) and *minimum storage regenerating* (MSR) *codes*. In this work we only consider MSR codes which repair a single failed node from all surviving nodes.

On the other hand, LRCs introduce additional redundancy to the system, such that a few/single node failures can be recovered from a small number of helper nodes, i.e., can be recovered *locally*. A particularly strong class of LRCs are *Partial MDS* (PMDS) *codes* [5], [6], [7], which guarantee to tolerate *all* failure patterns possible under these constraints. Specifically, an (r, s)-PMDS code of length μn can be partitioned into μ local groups of size n, such that any erasure pattern with r erasures in each local group plus any s erasures in arbitrary positions can be recovered.

However, the local recovery of nodes still induces a large amount of network traffic. To circumvent this bottleneck, several locally regenerating codes have been proposed, but only the construction of [8] has been shown to be PMDS [9] We propose several new constructions of locally regenerating PMDS codes with significantly smaller field size than the construction in [8]. Specifically, we construct a new PMDS code with two global parities ($s = 2$), where each local code is a d-MSR code. The construction is a non-trivial combination of the PMDS codes in [6] with the MSR codes in [10], and has field size in the order of $O(\mu r^2 n)$. Next, we present a new general construction of locally regenerating PMDS codes for any number of global parities. The construction combines an arbitrary family of universal PMDS codes (that is, the local codes can be chosen almost arbitrarily) and an MSR code whose rows are all MDS codes. This immediately leads to several new explicit locally regenerating PMDS codes using known universal PMDS families and the MSR codes in [10]: the PMDS codes in [8] result in a field size in the order of $O\big((rn)^{\mu(n-r)}\big)$, the ones in [11] give a field size in $O\big(\max\{rn, \mu + 1\}\big)^{n-r}\big)$, and the codes of [7] result in a locally regenerating PMDS code of field size $O(nr(2n\mu)^{s(r+1)-1})$. All new locally regenerating PMDS codes have the same subpacketization as the underlying MSR code from [10]. For the two-global-parities construction and the universal construction with the PMDS codes in [11] and [7], there are reasonable parameter ranges in which the respective construction has lowest field size among all known constructions. Moreover, for all parameters, at least one of the new constructions has a smaller field size than the known construction of [8].

Next, we consider PMDS codes with global regeneration properties that offer non-trivial repair schemes for the case where local recovery is not possible. Specifically, we give a PMDS code construction that, when punctured in any r positions in each local group, becomes an MSR code. The resulting code has a field size in $O(n^{\mu(n+s)})$ and subpacketization in $O((8n)^{\mu n(n+s)})$. The reduction in global repair bandwidth is of particular interest, as the bandwidth of connections between nodes of different local groups is often assumed to be smaller than of nodes within the same local groups. Accordingly, though being less likely to occur, the non-local repair of a larger number of erasures can take a substantial amount of time.

This work was partially presented at the IEEE International Symposium on Information Theory (ISIT) 2020 [1] and is available online [2].

The work of L. Holzbaur and A. Wachter-Zeh was supported by the German Research Foundation (Deutsche Forschungsgemeinschaft, DFG) under Grant No. WA3907/1-1. S. Puchinger has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement no. 713683. E. Yaakobi was supported in part by the Israel Science Foundation under Grant No. 1817/18, by the Technion Hiroshi Fujiwara Cyber Security Research Center, and by the Israel National Cyber Directorate. This work was also supported by the Technical University of Munich – Institute for Advanced Study, funded by the German Excellence Initiative and European Union 7th Framework Programme under Grant Agreement No. 291763.

II. Globally Regenerating PMDS Codes

We write $[a, b]$ for the set of integers $\{a, a+1, \dots, b\}$ and $[b]$ if $a = 1$. For a set of integers $R \subseteq [n]$ and a code $\mathcal{C}$ we write $\mathcal{C}|_R$ for the code obtained by restricting $\mathcal{C}$ to the positions indexed by R. For an $a \times b$ matrix $\boldsymbol{B}$ we denote by $\boldsymbol{B}_{i,j}$ the entry in the i-th row and j-th column. For the i-th row/column we write $\boldsymbol{B}_{i,:}$ and $\boldsymbol{B}_{:,i}$, respectively. For a set $\mathcal{S} \subset [b]$, we denote by $\boldsymbol{B}_{\mathcal{S}}$ the restriction of the matrix $\boldsymbol{B}$ to the columns indexed by $\mathcal{S}$.

We denote a linear code of length n, dimension k, and distance $d_{\min}$ over a field $\mathbb{F}_q$ by $[n, k, d_{\min}]_q$. If the field size or minimum distance is not relevant, we sometimes omit them. For a code over $\mathbb{F}_{q^\ell}$ that is linear over $\mathbb{F}_q$ we write $[n, k, d_{\min}; \ell]_q$, $[n, k, d_{\min}; \ell]$, or $[n, k; \ell]$, respectively. The parameter ℓ is referred to as the subpacketization of the code.

By definition, a PMDS code punctured in arbitrary r positions per group is an MDS code of distance $s + 1$. In the following we construct PMDS codes where each of these MDS codes is an MSR code. Our construction is based on Gabidulin codes a class of rank-metric codes that have been used repeatedly in the literature to construct LRCs and PMDS codes. For a formal definition see [2, Definition 8].

Definition 1 (Partial MDS array codes)**.** *Let $\mathcal{W} = \{W_1, W_2, \dots, W_\mu\}$ be a partition of $[\mu n]$ with $|W_i| = n \; \forall \; i \in [\mu]$. Let $\mathcal{C} \subset \mathbb{F}_q^{\ell \times \mu n}$ be a linear $[\mu n, (n-r)\mu - s; \ell]$ code. The code $\mathcal{C}$ is a* $\mathsf{PMDS}(\mu, n, r, s, \mathcal{W}; \ell)$ partial MDS array code *if*

- *$\mathcal{C}|_{W_i}$ is an $[n, n-r, r+1; \ell]$ MDS code for all $i \in [\mu]$*
- *$\mathcal{C}|_{[\mu n] \setminus \cup_{i=1}^{\mu} E_i}$ is an $[\mu n - r\mu, \mu n - r\mu - s, s+1; \ell]$ MDS code for any $E_i \subset W_i$ with $|E_i| = r \; \forall \; i \in [\mu]$.*

Definition 2 (Globally MSR PMDS Code)**.** *Let $\mathcal{C}$ be a* $\mathsf{PMDS}(\mu, n, r, s, \mathcal{W}; \ell)$ *code. We say that the code $\mathcal{C}$ is* globally MSR *if the restriction $\mathcal{C}|_{[\mu n] \setminus \cup_{i=1}^{\mu} E_i}$ is a $[\mu(n-r), \mu n - r\mu - s, s+1; \ell]$ MSR code for any $E_i \subset W_i$ with $|E_i| = r$ for all $i \in [\mu]$.*

The construction is based on two main observations: first, the principle used in the MSR codes of [10] can also be applied using Gabidulin codes instead of RS codes; second, performing linearly independent linear combinations of the symbols of a Gabidulin code yields another Gabidulin code with different code locators. Using these observations and carefully choosing the code locators for each row in an array of Gabidulin codewords, we assure the code obtained from puncturing r positions in each local group is MSR.

Construction 1 (Globally regenerating PMDS array codes)**.** *Let $\boldsymbol{B} \in \mathbb{F}_{q^M}^{\ell \times \mu(n-r)}$ and define the $[\mu n, \mu(n-r) - s; \ell]_{q^M}$ array code $\mathcal{C}(\mu, n, r, s, \boldsymbol{B}; \ell)_q$ as*

$$\{\boldsymbol{C} \in \mathbb{F}_q^{\ell \times \mu n} : \boldsymbol{C}_{a,:} = \boldsymbol{u}^{(a)} \cdot \boldsymbol{G}_{\boldsymbol{B}}^{(a)} \cdot \operatorname{diag}(\boldsymbol{G}_{\mathsf{MDS}}, \boldsymbol{G}_{\mathsf{MDS}}, \dots) \\ \forall \, \boldsymbol{u}^{(a)} \in \mathbb{F}_{q^m}, a = 0, \dots, \ell - 1\},$$

where $\boldsymbol{G}_{\boldsymbol{B}}^{(a)}$ is the generator matrix of the $[\mu(n-r), \mu(n-r) - s]$ Gabidulin code with code locators $\boldsymbol{B}_{a,:}$ and $\boldsymbol{G}_{\mathsf{MDS}}$ is a generator matrix of an $[n, n-r]_q$ MDS code.

It is easy to see that if the rows of the matrix $\boldsymbol{B}$ in Construction 1 contain linearly independent elements, then each row of the code is a PMDS code of the code family constructed in [8]. If the matrix $\boldsymbol{B}$ is chosen in a suitable way, then the MDS array codes obtained from erasing r positions in each local group are MSR codes as in [2, Definition 10], which can be seen as a Gabidulin-analog of Ye–Barg codes.

For the node repair algorithm of Ye–Barg codes [10], it is essential that the rows of a codeword can be partitioned into subsets for which there exist parity checks that differ exactly in position i , i.e., for which all entries are the same except for those at position i, which are all distinct. Replacing the RS code by a Gabidulin code entails a similar property of the code locators, which we formalize below along with a stronger requirement necessary for our construction.

Definition 3 (YB-Grouping Property)**.** *We say a matrix $\boldsymbol{B} \in \mathbb{F}_{q^m}^{\ell \times \mu(n-r)}$ has the **YB-grouping property w.r.t.** s if for each position i the rows of the matrix can be partitioned into $\frac{\ell}{s}$ subsets of s rows for which the elements in the i-th position are linearly independent and the elements in all other positions are the same for all s rows. Further, we say $\boldsymbol{B} \in \mathbb{F}_{q^M}^{\ell \times (\mu(n-r))}$ has the **scrambled YB grouping property** if $\boldsymbol{B} \cdot \operatorname{diag}(\boldsymbol{G}_1, \dots, \boldsymbol{G}_\mu)$ has the YB grouping property for all invertible matrices $\boldsymbol{G}_i \in \mathbb{F}_q^{(n-r) \times (n-r)}$.*

The code in Construction 1 can be obtained from a skew Ye–Barg code by multiplying it from the right by the $\mu(n-r) \times \mu n$ matrix $\operatorname{diag}(\boldsymbol{G}_{\mathsf{MDS}}, \boldsymbol{G}_{\mathsf{MDS}}, \dots)$. When puncturing arbitrary r positions in each local group, we do not obtain the original skew Ye–Barg code. However, we do get the original code multiplied from the right by an invertible matrix over $\mathbb{F}_q$. It is well-known that the rows of such a code are again Gabidulin codes. The *scrambled YB grouping property* captures the property required for these Gabidulin codes to again give a skew Ye–Barg code.

Theorem 1. *If $\boldsymbol{B}$ has the* scrambled YB grouping property*, then the code of construction is a globally-MSR PMDS code.*

A construction of a matrix $\boldsymbol{B}$ with this property is given in [2, Theorem 7] and leads to the following statement.

Theorem 2. *There is a globally-MSR PMDS code with field size*

$$(n-1)^{\mu(n-r+s-1)} \leq q^M < [2(n-1)]^{\mu(n-r+s-1)}$$

and subpacketization $\ell \leq 4^\mu q^{\mu(n-r)(n-r+s-1)}$.

References

[1] L. Holzbaur, S. Puchinger, E. Yaakobi, and A. Wachter-Zeh, "Partial MDS codes with local regeneration," in *2020 IEEE International Symposium on Information Theory (ISIT)*, 2020, pp. 628–633.

[2] L. Holzbaur, S. Puchinger, E. Yaakobi, and A. Wachter-Zeh, "Partial MDS codes with regeneration," *arXiv preprint arXiv:2009.07643*, 2020.

[3] P. Gopalan, C. Huang, H. Simitci, and S. Yekhanin, "On the locality of codeword symbols," *IEEE Trans. Inf. Theory*, vol. 58, no. 11, pp. 6925–6934, 2012.

[4] A. G. Dimakis, P. B. Godfrey, Y. Wu, M. J. Wainwright, and K. Ramchandran, "Network coding for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 56, no. 9, pp. 4539–4551, 2010.

[5] M. Blaum, J. L. Hafner, and S. Hetzler, "Partial-MDS codes and their application to raid type of architectures," *IEEE Trans. Inf. Theory*, vol. 59, no. 7, pp. 4510–4519, 2013.

[6] M. Blaum, J. S. Plank, M. Schwartz, and E. Yaakobi, "Construction of partial MDS and sector-disk codes with two global parity symbols," *IEEE Trans. Inf. Theory*, vol. 62, no. 5, pp. 2673–2681, 2016.

[7] R. Gabrys, E. Yaakobi, M. Blaum, and P. H. Siegel, "Constructions of partial MDS codes over small fields," *IEEE Trans. Inf. Theory*, vol. 65, no. 6, pp. 3692–3701, 2018.

[8] A. S. Rawat, O. O. Koyluoglu, N. Silberstein, and S. Vishwanath, "Optimal locally repairable and secure codes for distributed storage systems," *IEEE Trans. Inf. Theory*, vol. 60, no. 1, pp. 212–236, jan 2014.

[9] G. Calis and O. O. Koyluoglu, "A general construction for PMDS codes," *IEEE Commun. Lett.*, vol. 21, no. 3, pp. 452–455, 2016.

[10] M. Ye and A. Barg, "Explicit constructions of high-rate MDS array codes with optimal repair bandwidth," *IEEE Trans. Inf. Theory*, vol. 63, no. 4, pp. 2001–2014, 2017.

[11] U. Martínez-Peñas and F. R. Kschischang, "Universal and dynamic locally repairable codes with maximal recoverability via sum-rank codes," *IEEE Trans. Inf. Theory*, 2019.