The Optimal Rate of MDS Variable Generation

Yizhou Zhao, Hua Sun
EE Department, University of North Texas
Email: yizhouzhao@my.unt.edu, hua.sun@unt.edu

2023 IEEE International Symposium on Information Theory (ISIT) | 978-1-6654-7554-9/23/$31.00 ©2023 IEEE | DOI: 10.1109/ISIT54713.2023.10206914

***Abstract*—A collection of K random variables are called (K,n)-MDS if any n of the K variables are independent and determine all remaining variables. In the MDS variable generation problem, K users wish to generate variables that are (K,n)-MDS using a randomness variable owned by each user. We show that to generate 1 bit of (K,n)-MDS variables for each $n\in\{1,2,\cdots,K\}$, the minimum size of the randomness variable at each user is $1+1/2+\cdots+1/K$ bits.**

I. Introduction

Maximum distance separable (MDS) codes are one of the most fascinating classes of codes in coding theory (see Chapter 11 of [1]), with a wide array of applications ranging from storage systems [2]–[4], private information retrieval [5]–[9], coded computation [10]–[12] to secret sharing [13], [14] and secure multiparty computation [15]–[17]. In this work, we take a Shannon theoretic view to study how to efficiently generate random variables that have the MDS property.

A collection of K random variables $Z_1^n,\cdots,Z_K^n$ are said to be (K,n)-MDS if any n of them are independent and uniquely determine the remaining variables (see Table 1 for an example). Consider K users, where User $k\in\{1,2,\cdots,K\}$ holds a random variable Z_k. From Z_k, each user wishes to generate random variables $Z_k^1,\cdots,Z_k^K$ such that $Z_1^n,\cdots,Z_K^n$ are (K,n)-MDS. The question we explore is - to generate 1 bit of Z_k^n for each $n\in\{1,2,\cdots,K\}$, how many bits of the source Z_k are required?

From Figure 1, we see that K bits are sufficient for Z_k, when Z_k^n are independent for each n. Interestingly, we show that if the correlation among Z_k^n is optimally exploited, the size of Z_k (normalized by the size of Z_k^n) can be reduced to $1+1/2+\cdots+1/K$, i.e., the harmonic number, which is roughly $\ln K$. Furthermore, information theoretic converse is provided to prove that this is absolutely minimum.

II. Problem Statement and Main Result

Consider K users, where User $k\in\{1,2,\cdots,K\}\triangleq[K]$ holds a random variable Z_k of size L_Z bits. From Z_k, User $k\in[K]$ wishes to generate K random variables, $(Z_k^1,\cdots,Z_k^K)=(Z_k^n)_{n\in[K]}\triangleq Z_k^{\le K}$, where each Z_k^n has entropy L bits.

$$H(Z_k^1,\cdots,Z_k^K|Z_k)=0,\quad H(Z_k^n)=L,\forall n,k\in[K]. \tag{1}$$

Further, the variables $Z_1^n,Z_2^n,\cdots,Z_K^n$ are required to satisfy the following (K,n)-MDS property.

$$H\big((Z_k^n)_{k\in\mathcal{U}}\big)=\min(|\mathcal{U}|,n)\times L,\ \forall\mathcal{U}\subset[K]. \tag{2}$$

In words, among $Z_1^n,Z_2^n,\cdots,Z_K^n$, any n variables are independent and determine the remaining $K-n$ variables.

The generation efficiency is measured by the rate R_Z, defined as $R_Z\triangleq L_Z/L$, which characterizes the number of bits each user holds for each bit of the MDS variables generated. A rate value R_Z is said to be achievable if there exists an MDS variable generation scheme (i.e., a design of variables $(Z_k^n)_{n,k\in[K]}$), for which constraints (1), (2) are satisfied, and the rate is no greater than R_Z. The infimum of achievable R_Z values is called the optimal rate R_Z^*.

Theorem 1 states the main result.

Theorem 1. *For K-user MDS variable generation, the optimal rate is $R_Z^*=1+1/2+\cdots+1/K$.*

III. Proof of Theorem 1: Converse

Before proceeding to the general proof, we first consider the setting where $K=3$ to illustrate the key ideas.

A. Example: $K=3$ and $R_Z\ge 1+\frac{1}{2}+\frac{1}{3}$

The converse proof has a recursive nature, where we consider the generation of Z_k^1, $Z_k^{\le 2}$, and $Z_k^{\le 3}$ successively and later steps rely on results obtained in previous steps.

Step 1: Consider Z_k^1. From the definition of MDS variables (2), we have

$$H(Z_k^1)=L,\forall k\in\{1,2,3\}. \tag{3}$$

Step 2: Consider $Z_k^{\le 2}$.

$$\begin{aligned}
& H\Big(Z_1^{\le 2}\Big)+H\Big(Z_2^{\le 2}\Big) && (4)\\
= \ & H\Big(Z_1^{\le 2},Z_2^{\le 2}\Big)+I\Big(Z_1^{\le 2};Z_2^{\le 2}\Big) && (5)\\
\ge \ & H\big(Z_1^2,Z_2^2\big)+I\big(Z_1^1;Z_2^1\big) && (6)\\
\overset{(2)}{=} \ & H\big(Z_1^2,Z_2^2\big)+H\big(Z_1^1\big) && (7)\\
\overset{(2)(3)}{=} \ & 2L+L\ =\ 3L && (8)
\end{aligned}$$

where (7) follows from the definition of $(K,1)$-MDS variables, i.e., Z_2^1 is determined by Z_1^1, and in (8), the first term is due to definition of $(K,2)$-MDS variables and the second term follows from (3), i.e., the result from *Step 1* and we have reduced the problem from considering $Z_k^{\le 2}$ to Z_k^1.

Remark 1. *In the above derivation, one naively looking step (6) deserves highlighting. To obtain the first entropy term, we drop Z_1^1,Z_2^1 from $Z_1^{\le 2},Z_2^{\le 2}$ and this turns out to be tight because when we generate $(K,2)$-MDS variables Z_k^2, all entropy in $(K,1)$-MDS variables Z_k^1 is fully used (thus information wholly absorbed, see the achievable scheme in Section IV-A). To obtain the second mutual information term,*

	User 1	User 2	User 3	User 4	User 5	$\cdots$
$(K,1)$-MDS	A_1	A_1	A_1	A_1	A_1	
$(K,2)$-MDS	A_2	B_2	A_2+B_2	A_2+2B_2	A_2+3B_2	
$(K,3)$-MDS	A_3	B_3	C_3	$A_3+B_3+C_3$	$A_3+2B_3+3C_3$	Z_5^3
$(K,4)$-MDS	A_4	B_4	C_4	D_4	$A_4+B_4+C_4+D_4$	
$\vdots$						Z_5
(K,K)-MDS	A_K	B_K	C_K	D_K	E_K	$\cdots$

Fig. 1. An example of MDS variables. $A_i, B_j, \cdots$ are uniform and from a prime field, e.g., $\mathbb{F}_5$.

we drop Z_1^2, Z_2^2 because the two $(K,2)$-MDS variables are independent, leaving us with only $(K,1)$-MDS variables so that we may use results from Step 1.

Symmetrically, we can prove that (8) holds for any 2 users,

$$H\left(Z_i^{\leq 2}\right)+H\left(Z_j^{\leq 2}\right)\geq 3L, \forall i,j\in\{1,2,3\}, i\neq j. \tag{9}$$

Step 3: Finally, consider $Z_k^{\leq 3}$. Denote the set of all permutations of $\{1,2,3\}$ as $\mathcal{S}_3 \triangleq \{\pi_i\}_{i\in 3!}$, where $\pi_i = (\pi_i(1),\pi_i(2),\pi_i(3))$ is a permutation of $\{1,2,3\}$.

$$\begin{aligned}
&3!\times 3L_Z\\
\overset{(1)}{\geq}&\sum_{\pi\in\mathcal{S}_3}\left[H\left(Z_{\pi(1)}^{\leq 3}\right)+H\left(Z_{\pi(2)}^{\leq 3}\right)+H\left(Z_{\pi(3)}^{\leq 3}\right)\right] &(10)\\
=&\sum_{\pi\in\mathcal{S}_3}\left[H\left(Z_{\pi(1)}^{\leq 3},Z_{\pi(2)}^{\leq 3},Z_{\pi(3)}^{\leq 3}\right)+I\left(Z_{\pi(1)}^{\leq 3};Z_{\pi(2)}^{\leq 3}\right)\right.\\
&\left.+I\left(Z_{\pi(3)}^{\leq 3};Z_{\pi(1)}^{\leq 3},Z_{\pi(2)}^{\leq 3}\right)\right] &(11)\\
\geq&\sum_{\pi\in\mathcal{S}_3}\left[H\left(Z_{\pi(1)}^{3},Z_{\pi(2)}^{3},Z_{\pi(3)}^{3}\right)+I\left(Z_{\pi(1)}^{1};Z_{\pi(2)}^{1}\right)\right.\\
&\left.+I\left(Z_{\pi(3)}^{\leq 2};Z_{\pi(1)}^{\leq 2},Z_{\pi(2)}^{\leq 2}\right)\right] &(12)\\
\overset{(2)}{=}&3!H\left(Z_1^3,Z_2^3,Z_3^3\right)+\sum_{\pi\in\mathcal{S}_3}H\left(Z_{\pi(1)}^1\right)+\sum_{\pi\in\mathcal{S}_3}H\left(Z_{\pi(3)}^{\leq 2}\right)\\
\overset{(2)(3)}{\geq}&3!\times 3L+3!\times L+\left[H\left(Z_1^{\leq 2}\right)+H\left(Z_2^{\leq 2}\right)\right]\\
&+\left[H\left(Z_1^{\leq 2}\right)+H\left(Z_3^{\leq 2}\right)\right]+\left[H\left(Z_2^{\leq 2}\right)+H\left(Z_3^{\leq 2}\right)\right]\\
\overset{(9)}{\geq}&3!\times 3L+3!\times L+3\times 3L &(13)\\
\Rightarrow&\ R_Z=L_Z/L\geq 1+1/2+1/3 &(14)
\end{aligned}$$

where in (11), the identity $H(X)+H(Y)=H(X,Y)+I(X;Y)$ is used twice.

Remark 2. *Similar to Remark 1, the key step is (12). For the first term, all entropy in $Z_k^{\leq 3}$ is preserved in Z_k^3; for the remaining two mutual information terms, we may drop the uncorrelated terms, after which they become the entropy terms in (13) due to the MDS property so that we may use results from Step 1 (i.e., Z_k^1) and Step 2 (i.e., $Z_k^{\leq 2}$).*

B. General Proof: $R_Z \geq 1+1/2+\cdots+1/K$

Let us start with two useful identities. The first identity, stated in the following lemma, transforms the sum of individual entropy terms to the sum of a joint entropy term and a number of mutual information terms.

Lemma 1. *For any random variables $Z_1,\cdots,Z_K$, we have*

$$\begin{aligned}
&H(Z_1)+H(Z_2)+\cdots+H(Z_K)\\
=\ &H(Z_1,Z_2,\cdots,Z_k)+I(Z_1;Z_2)+I(Z_3;Z_1,Z_2)\\
&+\cdots+I(Z_K;Z_1,Z_2,\cdots,Z_{K-1}). &(15)
\end{aligned}$$

Proof:

$$\begin{aligned}
&\left[H(Z_1)+H(Z_2)\right]+H(Z_3)+\cdots+H(Z_K)\\
=\ &H(Z_1,Z_2)+I(Z_1;Z_2)+H(Z_3)+\cdots+H(Z_K) &(16)\\
=\ &\left[H(Z_1,Z_2)+H(Z_3)\right]+I(Z_1;Z_2)+H(Z_4)\\
&+\cdots+H(Z_K) &(17)\\
=\ &H(Z_1,Z_2,Z_3)+I(Z_3;Z_1,Z_2)+I(Z_1;Z_2)+H(Z_4)\\
&+\cdots+H(Z_K)\ =\ \cdots &(18)\\
=\ &H(Z_1,Z_2,\cdots,Z_K)+I(Z_1;Z_2)+I(Z_3;Z_1,Z_2)\\
&+\cdots+I(Z_K;Z_1,Z_2,\cdots,Z_{K-1}). &(19)
\end{aligned}$$

■

The second identity, stated in the following lemma, transforms mutual information terms to joint entropy terms, for MDS variables.

Lemma 2. *For MDS variables $(Z_k^n)_{n,k\in[K]}$, we have*

$$I\left(Z_k^{\leq n};\left(Z_u^{\leq n}\right)_{u\in\mathcal{U}}\right)=H\left(Z_k^{\leq n}\right),\ \forall\mathcal{U}\subset[K]\backslash\{k\},|\mathcal{U}|=n. \tag{20}$$

Proof: The proof is immediate, by applying the definition of (K,n)-MDS variables in (2).

$$\begin{aligned}
&I\left(Z_k^{\leq n};\left(Z_u^{\leq n}\right)_{u\in\mathcal{U}}\right) &(21)\\
=&H\left(Z_k^{\leq n}\right)+H\left(\left(Z_u^{\leq n}\right)_{u\in\mathcal{U}}\right)-H\left(\left(Z_u^{\leq n}\right)_{u\in\mathcal{U}\cup\{k\}}\right) &(22)\\
\overset{(2)}{=}&H\left(Z_k^{\leq n}\right). &(23)
\end{aligned}$$

■

We are now ready to recursively bound the entropy of any n out of the K MDS variables $Z_k^{\leq n}$. This result is stated in the following lemma.

Lemma 3. *For MDS variables $(Z_k^n)_{n,k\in[K]}$, we have $\forall n\in[K]$, and $\forall\mathcal{U}\subset[K],|\mathcal{U}|=n$,*

$$\frac{1}{n}\sum_{k\in\mathcal{U}}H\left(Z_k^{\leq n}\right)\geq\left(1+\frac{1}{2}+\cdots+\frac{1}{n}\right)L. \tag{31}$$

$$\frac{(M+1)!}{M+1}\sum_{k\in\mathcal{U}} H\Big(Z_k^{\leq M+1}\Big)$$

$$= \frac{1}{(M+1)}\sum_{\pi\in\mathcal{S}_{M+1}}\Big[H\Big(Z_{k_{\pi(1)}}^{\leq M+1}\Big)+H\Big(Z_{k_{\pi(2)}}^{\leq M+1}\Big)+\cdots+H\Big(Z_{k_{\pi(M+1)}}^{\leq M+1}\Big)\Big] \tag{24}$$

$$\overset{(15)}{=} \frac{1}{(M+1)}\sum_{\pi\in\mathcal{S}_{M+1}}\Big[H\Big(Z_{k_{\pi(1)}}^{\leq M+1},Z_{k_{\pi(2)}}^{\leq M+1},\cdots,Z_{k_{\pi(M+1)}}^{\leq M+1}\Big)+I\Big(Z_{k_{\pi(1)}}^{\leq M+1};Z_{k_{\pi(2)}}^{\leq M+1}\Big)$$
$$+I\Big(Z_{k_{\pi(3)}}^{\leq M+1};Z_{k_{\pi(1)}}^{\leq M+1},Z_{k_{\pi(2)}}^{\leq M+1}\Big)+\cdots+I\Big(Z_{k_{\pi(M+1)}}^{\leq M+1};Z_{k_{\pi(1)}}^{\leq M+1},Z_{k_{\pi(2)}}^{\leq M+1},\cdots,Z_{k_{\pi(M)}}^{\leq M+1}\Big)\Big] \tag{25}$$

$$\geq \frac{1}{(M+1)}\sum_{\pi\in\mathcal{S}_{M+1}}\Big[H\Big(Z_{k_{\pi(1)}}^{M+1},Z_{k_{\pi(2)}}^{M+1},\cdots,Z_{k_{\pi(M+1)}}^{M+1}\Big)+I\Big(Z_{k_{\pi(1)}}^{1};Z_{k_{\pi(2)}}^{1}\Big)$$
$$+I\Big(Z_{k_{\pi(3)}}^{\leq 2};Z_{k_{\pi(1)}}^{\leq 2},Z_{k_{\pi(2)}}^{\leq 2}\Big)+\cdots+I\Big(Z_{k_{\pi(M+1)}}^{\leq M};Z_{k_{\pi(1)}}^{\leq M},Z_{k_{\pi(2)}}^{\leq M},\cdots,Z_{k_{\pi(M)}}^{\leq M}\Big)\Big] \tag{26}$$

$$\overset{(2)(20)}{=} \frac{1}{(M+1)}\sum_{\pi\in\mathcal{S}_{M+1}}\Big[(M+1)L+H\Big(Z_{k_{\pi(1)}}^{1}\Big)+H\Big(Z_{k_{\pi(3)}}^{\leq 2}\Big)+\cdots+H\Big(Z_{k_{\pi(M+1)}}^{\leq M}\Big)\Big] \tag{27}$$

$$= \frac{1}{(M+1)}\Bigg[\sum_{\pi\in\mathcal{S}_{M+1}}(M+1)L+\sum_{\pi\in\mathcal{S}_{M+1}}H\Big(Z_{k_{\pi(1)}}^{1}\Big)+\frac{1}{2}\sum_{\pi\in\mathcal{S}_{M+1}}\Big[H\Big(Z_{k_{\pi(1)}}^{\leq 2}\Big)+H\Big(Z_{k_{\pi(2)}}^{\leq 2}\Big)\Big]$$
$$+\cdots+\frac{1}{M}\sum_{\pi\in\mathcal{S}_{M+1}}\Big[H\Big(Z_{k_{\pi(1)}}^{\leq M}\Big)+H\Big(Z_{k_{\pi(2)}}^{\leq M}\Big)+\cdots+H\Big(Z_{k_{\pi(M)}}^{\leq M}\Big)\Big]\Bigg] \tag{28}$$

$$\geq \frac{(M+1)!}{(M+1)}\Bigg[(M+1)L+L+\Big(1+\frac{1}{2}\Big)L+\cdots+\Big(1+\frac{1}{2}+\cdots+\frac{1}{M}\Big)L\Bigg] \text{ (Induction)} \tag{29}$$

$$= (M+1)!\Big(1+\frac{1}{2}+\frac{1}{3}+\cdots+\frac{1}{M}+\frac{1}{M+1}\Big)L \tag{30}$$

Proof: The proof is based on mathematical induction on n.

Base case: When $n=1$, (31) becomes $H\big(Z_k^1\big)\geq L$, which follows directly from (2).

Induction step: Suppose (31) holds for $n\in[M]$, $1\leq M\leq K-1$, then we show that (31) also holds for $n=M+1$. Consider (31) when $n=M+1$ and suppose $\mathcal{U}=\{k_1,k_2,\cdots,k_{M+1}\}\subset[K]$. Denote the set of all permutations of $[M+1]$ as $\mathcal{S}_{M+1}=(\pi_i)_{i\in[(M+1)!]}$. The derivation is shown at the top of this page, where in (24), we include all permutations of users with indicies in $\mathcal{U}$ and (25) uses Lemma 1. In (26), we follow the insights in Remark 1 to drop terms, which cannot increase entropy or mutual information. In (27), we use Lemma 2 and the definition of MDS variables (2). In (28), we replace each term by averages, i.e., $\sum_{\pi\in\mathcal{S}_{M+1}}H\Big(Z_{k_{\pi(i)}}^{\leq n}\Big)=\sum_{\pi\in\mathcal{S}_{M+1}}H\Big(Z_{k_{\pi(j)}}^{\leq n}\Big),\forall i,j\in[M+1]$. In (29), we use the induction assumption that (31) holds for $n\in[M]$. ■

Equipped with Lemma 3, the final converse proof of R_Z follows immediately. Set $n=K$ in (31), i.e., $\mathcal{U}=[K]$, then

$$L_Z \overset{(1)}{\geq} \frac{1}{K}\sum_{k\in[K]}H\Big(Z_k^{\leq K}\Big) \tag{32}$$

$$\overset{(31)}{\geq} \Big(1+\frac{1}{2}+\cdots+\frac{1}{K}\Big)L \tag{33}$$

$$\Rightarrow\ R_Z=\frac{L_Z}{L} \geq 1+\frac{1}{2}+\cdots+\frac{1}{K}. \tag{34}$$

IV. Proof of Theorem 1: Achievability

The achievability proof is fairly straightforward. After setting up the dimensions following the insights from the converse proof, we only need random linear codes and transformations. Let us start with an example of $K=3$ and then proceed to the general proof.

A. Example: $K=3$ and $R_Z=1+1/2+1/3$

We show that when $K=3$, rate $R_Z=L_Z/L=1+1/2+1/3=11/6$ is achievable. To this end, suppose $L=6\log_2 q$, i.e., each MDS variable Z_k^n consists of 6 symbols from $\mathbb{F}_q$ and $L_Z=11\log_2 q$, i.e., each source variable Z_k consists of 11 symbols from $\mathbb{F}_q$. Suppose the prime power field size $q>72$.

Step 1: We describe the design of Z_k. We need 3 i.i.d. uniform 6×1 vectors over $\mathbb{F}_q$, denoted as S^1,S^2,S^3, then we set

$$Z_k=\big(\mathbf{H}_k^1S^1,\mathbf{H}_k^2S^2,\mathbf{H}_k^3S^3\big),\forall k\in\{1,2,3\} \tag{35}$$

where $\mathbf{H}_k^1\in\mathbb{F}_q^{6\times 6},\mathbf{H}_k^2\in\mathbb{F}_q^{3\times 6},\mathbf{H}_k^3\in\mathbb{F}_q^{2\times 6}$ need to satisfy some generic (full rank) properties (see Lemma 4 for details). For now, it suffices to think of them as random matrices over a large field, which will work with high probability. Note that Z_k has $6+3+2=11$ symbols, as desired.

Step 2: We describe the generation of MDS variables Z_k^n. For $(K,1)$-MDS variables Z_k^1, we set

$$Z_k^1 = \mathbf{H}_k^1 S^1 \tag{36}$$

which has 6 symbols.

For $(K,2)$-MDS variables Z_k^2, we set

$$Z_k^2 = \big(\mathbf{V}_k^{2\leftarrow 1}\mathbf{H}_k^1 S^1, \mathbf{H}_k^2 S^2\big) \tag{37}$$

where $\mathbf{V}_k^{2\leftarrow 1} \in \mathbb{F}_q^{3\times 6}$ transforms $(K,1)$-MDS variables to $(K,2)$-MDS variables (with maximum efficiency, see Remark 1 in the converse proof). Again, $\mathbf{V}_k^{2\leftarrow 1}$ need to satisfy some generic properties (stated later in Lemma 4), which hold with high probability over large fields. Note that Z_k^2 has 6 symbols.

For $(K,3)$-MDS variables Z_k^3, we set

$$Z_k^3 = \big(\mathbf{V}_k^{3\leftarrow 1}\mathbf{H}_k^1 S^1, \mathbf{V}_k^{3\leftarrow 2}\mathbf{H}_k^2 S^2, \mathbf{H}_k^3 S^3\big) \tag{38}$$

where $\mathbf{V}_k^{3\leftarrow 1} \in \mathbb{F}_q^{2\times 6}$ and $\mathbf{V}_k^{3\leftarrow 2} \in \mathbb{F}_q^{2\times 3}$ transform $(K,1)$-MDS and $(K,2)$-MDS variables to $(K,3)$-MDS variables, respectively. The required generic conditions on $\mathbf{V}_k^{3\leftarrow 1}, \mathbf{V}_k^{3\leftarrow 2}$ will be stated in Lemma 4. Note that Z_k^3 has 6 symbols.

Step 3: We specify the conditions on the matrices used in the code construction, $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ such that MDS property (2) holds. For our proposed linear codes, it is straightforward to verify that we only need to guarantee (2) when[1] $|\mathcal{U}| = n$.

For $(K,1)$-MDS variables Z_k^1, we require

$$\mathbf{H}_k^1 \in \mathbb{F}_q^{6\times 6} \text{ has full rank} \tag{39}$$

so that

$$H(Z_k^1) \overset{(36)}{=} \text{rank}(\mathbf{H}_k^1)\log_2 q \overset{(39)}{=} 6\log_2 q = L \tag{40}$$

which follows from the uniformity of S^1 so that entropy of its linear transformation is specified by the rank of the transformation matrix $\mathbf{H}_k^1$.

For $(K,2)$-MDS variables Z_k^2, we require for any $\mathcal{U} = \{k_1,k_2\} \subset [K]$

$$\begin{bmatrix} \mathbf{V}_{k_1}^{2\leftarrow 1}\mathbf{H}_{k_1}^1 \\ \mathbf{V}_{k_2}^{2\leftarrow 1}\mathbf{H}_{k_2}^1 \end{bmatrix}_{6\times 6} \text{and} \begin{bmatrix} \mathbf{H}_{k_1}^2 \\ \mathbf{H}_{k_2}^2 \end{bmatrix}_{6\times 6} \text{have full rank} \tag{41}$$

so that

$$\begin{aligned}
&H(Z_{k_2}^2, Z_{k_3}^2) \\
&\overset{(37)}{=} H\big(\mathbf{V}_{k_1}^{2\leftarrow 1}\mathbf{H}_{k_1}^1 S^1, \mathbf{V}_{k_2}^{2\leftarrow 1}\mathbf{H}_{k_2}^1 S^1, \mathbf{H}_{k_1}^2 S^2, \mathbf{H}_{k_2}^2 S^2\big) && (42)\\
&= H\left(\begin{bmatrix} \mathbf{V}_{k_1}^{2\leftarrow 1}\mathbf{H}_{k_1}^1 \\ \mathbf{V}_{k_2}^{2\leftarrow 1}\mathbf{H}_{k_2}^1 \end{bmatrix} S^1\right) + H\left(\begin{bmatrix} \mathbf{H}_{k_1}^2 \\ \mathbf{H}_{k_2}^2 \end{bmatrix} S^2\right) && (43)\\
&= \left(\text{rank}\left(\begin{bmatrix} \mathbf{V}_{k_1}^{2\leftarrow 1}\mathbf{H}_{k_1}^1 \\ \mathbf{V}_{k_2}^{2\leftarrow 1}\mathbf{H}_{k_2}^1 \end{bmatrix}\right) + \text{rank}\left(\begin{bmatrix} \mathbf{H}_{k_1}^2 \\ \mathbf{H}_{k_2}^2 \end{bmatrix}\right)\right)\log_2 q && (44)\\
&\overset{(41)}{=} 12\log_2 q = 2L && (45)
\end{aligned}$$

[1]Our construction is based on linear transformations on uniform variables so that entropy terms boil down to rank terms. When $|\mathcal{U}| = n$, we will show that (2) is equivalent to requesting that certain square matrices $\mathbf{H}$ have full rank and $(Z_k^n)_{k\in\mathcal{U}}$ is invertible to $S^{\le n}$ (refer to (39) to (49)). As a result, when $|\mathcal{U}| < n$, (2) holds as it is associated with sub-matrices of $\mathbf{H}$, which must also have full rank; when $|\mathcal{U}| > n$, the additional Z_k^n terms are a function of $S^{\le n}$ thus contributing no more entropy.

where (43) follows from the independence of S^1 and S^2.

For $(K,3)$-MDS variables Z_k^3, we require

$$\begin{bmatrix} \mathbf{V}_1^{3\leftarrow 1}\mathbf{H}_1^1 \\ \mathbf{V}_2^{3\leftarrow 1}\mathbf{H}_2^1 \\ \mathbf{V}_3^{3\leftarrow 1}\mathbf{H}_3^1 \end{bmatrix}_{6\times 6}, \begin{bmatrix} \mathbf{V}_1^{3\leftarrow 2}\mathbf{H}_1^2 \\ \mathbf{V}_2^{3\leftarrow 2}\mathbf{H}_2^2 \\ \mathbf{V}_3^{3\leftarrow 2}\mathbf{H}_3^2 \end{bmatrix}_{6\times 6}, \text{and} \begin{bmatrix} \mathbf{H}_1^3 \\ \mathbf{H}_2^3 \\ \mathbf{H}_3^3 \end{bmatrix}_{6\times 6} \text{have full rank} \tag{46}$$

so that

$$\begin{aligned}
&H(Z_1^3, Z_2^3, Z_3^3) \\
&\overset{(38)}{=} H\big(\mathbf{V}_1^{3\leftarrow 1}\mathbf{H}_1^1 S^1, \mathbf{V}_2^{3\leftarrow 1}\mathbf{H}_2^1 S^1, \mathbf{V}_3^{3\leftarrow 1}\mathbf{H}_3^1 S^1\big) \\
&\quad + H\big(\mathbf{V}_1^{3\leftarrow 2}\mathbf{H}_1^2 S^2, \mathbf{V}_2^{3\leftarrow 2}\mathbf{H}_2^2 S^2, \mathbf{V}_3^{3\leftarrow 2}\mathbf{H}_3^2 S^2\big) \\
&\quad + H\big(\mathbf{H}_1^3 S^3, \mathbf{H}_2^3 S^3, \mathbf{H}_3^3 S^3\big) && (47)\\
&= \left(\text{rank}\left(\begin{bmatrix} \mathbf{V}_1^{3\leftarrow 1}\mathbf{H}_1^1 \\ \mathbf{V}_2^{3\leftarrow 1}\mathbf{H}_2^1 \\ \mathbf{V}_3^{3\leftarrow 1}\mathbf{H}_3^1 \end{bmatrix}\right) + \text{rank}\left(\begin{bmatrix} \mathbf{V}_1^{3\leftarrow 2}\mathbf{H}_1^2 \\ \mathbf{V}_2^{3\leftarrow 2}\mathbf{H}_2^2 \\ \mathbf{V}_3^{3\leftarrow 2}\mathbf{H}_3^2 \end{bmatrix}\right)\right. \\
&\quad \left. + \text{rank}\left(\begin{bmatrix} \mathbf{H}_1^3 \\ \mathbf{H}_2^3 \\ \mathbf{H}_3^3 \end{bmatrix}\right)\right)\log_2 q && (48)\\
&\overset{(46)}{=} 18\log_2 q = 3L. && (49)
\end{aligned}$$

Step 4: Finally, we show that there exist matrices $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ that satisfy the required full rank conditions. This result is stated in the following lemma.

Lemma 4. *If $q > 72$, we have $(\mathbf{V}_k^{n_2\leftarrow n_1})_{k,n_1,n_2\in\{1,2,3\},n_2>n_1}$, $(\mathbf{H}_k^n)_{k,n\in\{1,2,3\}}$ such that (39), (41), (46) are satisfied.*

Proof: The existence proof is based on probabilistic arguments. Draw each element of the matrices $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ independently and uniformly from $\mathbb{F}_q$. Denote the vector that contains all such elements as $\vec{v}$. View the determinant of each matrix in (39), (41), (46) as a polynomial in $\vec{v}$ and consider the product of all such polynomials, denoted by $f(\vec{v})$. $f(\vec{v})$ is product of $\binom{3}{1} + 2\binom{3}{2} + 3\binom{3}{3} = 12$ polynomials, each of which has degree at most 6, so the degree of $f(\vec{v})$ is at most $12\times 6 = 72$.

$f(\vec{v})$ is not the zero polynomial (proved later), so we can apply the Schwartz–Zippel lemma to obtain

$$\Pr(f(\vec{v}) = 0) \le 72/q < 1. \tag{50}$$

Therefore, there exists at least one assignment of $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ so that all matrices in (39), (41), (46) have full rank and thus the generated variables are indeed MDS.

Lastly, we are left to prove that $f(\vec{v})$ is not identically zero. To this end, it suffices to consider each matrix in (39), (41), (46) and show that for each such matrix, there exists one realization of $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ so that the matrix has full rank (and its determinant polynomial is not identically zero). This is proved next. A matrix that only involves $\mathbf{H}_k^n$ is trivial as we may set it as the identity matrix; a matrix that involves both $\mathbf{H}_k^n$ and $\mathbf{V}_k^{n_2\leftarrow n_1}$ can be set as the identity matrix as well because we can find a realization shown at the top of

$$(41): \quad \begin{bmatrix} \mathbf{V}_{k_1}^{2\leftarrow 1}\mathbf{H}_{k_1}^1 \\ \mathbf{V}_{k_2}^{2\leftarrow 1}\mathbf{H}_{k_2}^1 \end{bmatrix} = \mathbf{I}_6 \Leftarrow \mathbf{V}_{k_1}^{2\leftarrow 1} = \mathbf{V}_{k_2}^{2\leftarrow 1} = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0}_3 \end{bmatrix}, \mathbf{H}_{k_1}^1 = \begin{bmatrix} \mathbf{I}_3 & \mathbf{0}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 \end{bmatrix}, \mathbf{H}_{k_2}^1 = \begin{bmatrix} \mathbf{0}_3 & \mathbf{I}_3 \\ \mathbf{0}_3 & \mathbf{0}_3 \end{bmatrix} \tag{51}$$

$$(46): \quad \begin{bmatrix} \mathbf{V}_1^{3\leftarrow 1}\mathbf{H}_1^1 \\ \mathbf{V}_2^{3\leftarrow 1}\mathbf{H}_2^1 \\ \mathbf{V}_3^{3\leftarrow 1}\mathbf{H}_3^1 \end{bmatrix} = \mathbf{I}_6 \Leftarrow \mathbf{V}_1^{3\leftarrow 1} = \mathbf{V}_2^{3\leftarrow 1} = \mathbf{V}_3^{3\leftarrow 1} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_{2\times 4} \end{bmatrix},$$

$$\mathbf{H}_1^1 = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_{2\times 4} \\ \mathbf{0}_{4\times 2} & \mathbf{0}_{4\times 4} \end{bmatrix}, \mathbf{H}_2^1 = \begin{bmatrix} \mathbf{0}_2 & \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_{4\times 2} & \mathbf{0}_{4\times 2} & \mathbf{0}_{4\times 2} \end{bmatrix}, \mathbf{H}_3^1 = \begin{bmatrix} \mathbf{0}_{2\times 4} & \mathbf{I}_2 \\ \mathbf{0}_{4\times 4} & \mathbf{0}_{4\times 2} \end{bmatrix}; \tag{52}$$

$$\begin{bmatrix} \mathbf{V}_1^{3\leftarrow 2}\mathbf{H}_1^2 \\ \mathbf{V}_2^{3\leftarrow 2}\mathbf{H}_2^2 \\ \mathbf{V}_3^{3\leftarrow 2}\mathbf{H}_3^2 \end{bmatrix} = \mathbf{I}_6 \Leftarrow \mathbf{V}_1^{3\leftarrow 2} = \mathbf{V}_2^{3\leftarrow 2} = \mathbf{V}_3^{3\leftarrow 2} = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_{2\times 1} \end{bmatrix},$$

$$\mathbf{H}_1^2 = \begin{bmatrix} \mathbf{I}_2 & \mathbf{0}_{2\times 4} \\ \mathbf{0}_{1\times 2} & \mathbf{0}_{1\times 4} \end{bmatrix}, \mathbf{H}_2^2 = \begin{bmatrix} \mathbf{0}_2 & \mathbf{I}_2 & \mathbf{0}_2 \\ \mathbf{0}_2 & \mathbf{0}_2 & \mathbf{0}_2 \end{bmatrix}, \mathbf{H}_3^2 = \begin{bmatrix} \mathbf{0}_{2\times 4} & \mathbf{I}_2 \\ \mathbf{0}_{1\times 4} & \mathbf{0}_{1\times 2} \end{bmatrix} \tag{53}$$

this page, where $\mathbf{I}_i$ is the $i \times i$ identity matrix and $\mathbf{0}_i$ ($\mathbf{0}_{i\times j}$) is an $i \times i$ ($i \times j$) matrix wherein each element is zero. ■

B. General Proof: Any K

The general achievability proof of $R_Z = 1 + 1/2 + \cdots + 1/K$ is an immediate generalization of that of above example. Suppose $L = K! \log_2 q$ and the prime power field size[2] $q > K! \sum_{n\in[K]} n\binom{K}{n}$.

Step 1: Design Z_k. Set

$$Z_k = \left((\mathbf{H}_k^n S^n)_{n\in[K]} \right), \forall k \in [K] \tag{54}$$

where $S^n, n \in [K]$ are K i.i.d. uniform $K! \times 1$ vectors over $\mathbb{F}_q$ and $\mathbf{H}_k^n \in \mathbb{F}_q^{\frac{K!}{n}\times K!}$. Note that Z_k contains $L_Z/\log_2 q = \sum_{n\in[K]} K!/n$ symbols, so $R_Z = L_Z/L = \sum_{n\in[K]} 1/n$, as desired.

Step 2: Design Z_k^n. Set

$$Z_k^n = \left((\mathbf{V}_k^{n\leftarrow m}\mathbf{H}_k^m S^m)_{m\in[n-1]}, \mathbf{H}_k^n S^n \right), \forall m \in [K] \tag{55}$$

where $\mathbf{V}_k^{n\leftarrow m} \in \mathbb{F}_q^{\frac{K!}{n}\times\frac{K!}{m}}$.

Step 3: Conditions on $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ such that MDS property (2) holds. For (K, n)-MDS variables Z_k^n, we require for any $\mathcal{U} = \{k_1, k_2, \cdots, k_n\} \subset [K]$

$$\begin{bmatrix} \mathbf{V}_{k_1}^{n\leftarrow m}\mathbf{H}_{k_1}^m \\ \mathbf{V}_{k_2}^{n\leftarrow m}\mathbf{H}_{k_2}^m \\ \vdots \\ \mathbf{V}_{k_n}^{n\leftarrow m}\mathbf{H}_{k_n}^m \end{bmatrix}_{K!\times K!} \triangleq \mathbf{F}_{\mathcal{U}}^m, \forall m \in [n-1],$$

$$\text{and} \begin{bmatrix} \mathbf{H}_{k_1}^n \\ \mathbf{H}_{k_2}^n \\ \vdots \\ \mathbf{H}_{k_n}^n \end{bmatrix}_{K!\times K!} \triangleq \mathbf{H}_{\mathcal{U}}^n \text{ have full rank} \tag{56}$$

so that

$$H\left((Z_k^n)_{k\in\mathcal{U}}\right)$$

$$\overset{(55)}{=} H\left((\mathbf{H}_k^n S^n)_{k\in\mathcal{U}}\right) + \sum_{m\in[n-1]} H\left((\mathbf{V}_k^{n\leftarrow m}\mathbf{H}_k^m S^m)_{k\in\mathcal{U}}\right) \tag{57}$$

$$= \left(\text{rank}(\mathbf{H}_{\mathcal{U}}^n) + \sum_{m\in[n-1]} \text{rank}(\mathbf{F}_{\mathcal{U}}^m) \right) \log_2 q \tag{58}$$

$$\overset{(56)}{=} nK! \log_2 q = nL. \tag{59}$$

Thus (2) is guaranteed when $|\mathcal{U}| = n$. The cases where $\mathcal{U} \neq n$ follow in a straightforward manner (see the explanation in Footnote 1).

Step 4: Finally, we are left to show that there exist matrices $(\mathbf{H}_k^n)_{k,n\in[K]}$, $(\mathbf{V}_k^{n_2\leftarrow n_1})_{k,n_1,n_2\in[K],n_2>n_1}$ that satisfy (56). Draw each element of the matrices $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ independently and uniformly from $\mathbb{F}_q$. Denote the vector that contains all such elements as $\vec{v}$. View the determinant of each matrix in (56) as a polynomial in $\vec{v}$ and consider the product of all such polynomials, denoted by $f(\vec{v})$. $f(\vec{v})$ is product of $\sum_{n\in[K]} n\binom{K}{n}$ polynomials, each of which has degree at most $K!$, so the degree of $f(\vec{v})$ is at most $K! \sum_{n\in[K]} n\binom{K}{n}$.

$f(\vec{v})$ is not the zero polynomial (whose proof is straightforward as we may find realizations of $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ such that each matrix in (56) is the identity matrix following the proof of Lemma 4), so we can apply the Schwartz–Zippel lemma to obtain $\Pr(f(\vec{v}) = 0) < 1$. Therefore, there exists at least one assignment of $\mathbf{H}_k^n, \mathbf{V}_k^{n_2\leftarrow n_1}$ so that all matrices in (56) have full rank and thus the generated variables are indeed MDS.

V. Conclusion

In this work, we characterize the optimal rate of MDS variable generation somewhat surprisingly and interestingly, as the harmonic number. The application of MDS variable generation to an intimately related problem - secure summation with user selection and related discussion can be found in the full version of this work [18].

Acknowledgement

This work is supported in part by NSF under Grant CCF-2007108 and Grant CCF-2045656.

[2]Similar to Shannon's original random coding proof to the achievability of channel capacity, our proof is existence based and no effort is devoted to minimizing the field size required.

REFERENCES

[1] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. Elsevier, 1977, vol. 16.

[2] M. Blaum, J. Bruck, and A. Vardy, "MDS Array Codes with Independent Parity Symbols," *IEEE Transactions on Information Theory*, vol. 42, no. 2, pp. 529–542, 1996.

[3] A. G. Dimakis, K. Ramchandran, Y. Wu, and C. Suh, "A Survey on Network Codes for Distributed Storage," *Proceedings of the IEEE*, vol. 99, pp. 476–489, 2011.

[4] V. Ramkumar, M. Vajha, S. B. Balaji, M. N. Krishnan, B. Sasidharan, and P. V. Kumar, "Codes for Distributed Storage," in *Concise Encyclopedia of Coding Theory*. Chapman and Hall/CRC, 2021, pp. 735–762.

[5] K. Banawan and S. Ulukus, "The Capacity of Private Information Retrieval from Coded Databases," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1945–1956, 2018.

[6] R. Freij-Hollanti, O. Gnilke, C. Hollanti, and D. Karpuk, "Private Information Retrieval from Coded Databases with Colluding Servers," *SIAM Journal on Applied Algebra and Geometry*, vol. 1, no. 1, pp. 647–664, 2017.

[7] H. Sun and S. A. Jafar, "Private Information Retrieval from MDS Coded Data with Colluding Servers: Settling a Conjecture by Freij-Hollanti et al." *IEEE Transactions on Information Theory*, vol. 64, no. 2, pp. 1000–1022, 2018.

[8] R. Zhou, C. Tian, H. Sun, and T. Liu, "Capacity-Achieving Private Information Retrieval Codes from MDS-Coded Databases with Minimum Message Size," *IEEE Transactions on Information Theory*, vol. 66, no. 8, pp. 4904–4916, 2020.

[9] H. Sun and C. Tian, "Breaking the MDS-PIR Capacity Barrier via Joint Storage Coding," *Information*, vol. 10, no. 9, p. 265, 2019.

[10] K. Lee, M. Lam, R. Pedarsani, D. Papailiopoulos, and K. Ramchandran, "Speeding Up Distributed Machine Learning Using Codes," *IEEE Transactions on Information Theory*, vol. 64, no. 3, pp. 1514–1529, 2017.

[11] S. Dutta, V. Cadambe, and P. Grover, "Short-Dot: Computing Large Linear Transforms Distributedly Using Coded Short Dot Products," *Advances In Neural Information Processing Systems*, vol. 29, 2016.

[12] S. Li and S. Avestimehr, "Coded Computing: Mitigating Fundamental Bottlenecks in Large-scale Distributed Computing and Machine Learning," *Foundations and TrendsÂ® in Communications and Information Theory*, vol. 17, no. 1, pp. 1–148, 2020. [Online]. Available: http://dx.doi.org/10.1561/0100000103

[13] R. J. McEliece and D. V. Sarwate, "On Sharing Secrets and Reed-Solomon Codes," *Communications of the ACM*, vol. 24, no. 9, pp. 583–584, 1981.

[14] A. Beimel, "Secret-Sharing Schemes: A Survey," in *International Conference on Coding and Cryptology*. Springer, 2011, pp. 11–46.

[15] M. Ben-Or, S. Goldwasser, and A. Wigderson, "Completeness Theorems for Non-Cryptographic Fault-Tolerant Distributed Computation," in *Proceedings of the twentieth annual ACM symposium on Theory of computing*. ACM, 1988, pp. 1–10.

[16] D. Chaum, C. Crépeau, and I. Damgard, "Multiparty Unconditionally Secure Protocols," in *Proceedings of the twentieth annual ACM symposium on Theory of computing*. ACM, 1988, pp. 11–19.

[17] R. Cramer, I. B. Damgard, and J. B. Nielsen, *Secure Multiparty Computation and Secret Sharing*. Cambridge University Press, 2015.

[18] Y. Zhao and H. Sun, "MDS Variable Generation and Secure Summation with User Selection," *arXiv preprint arXiv:2211.01220*, 2022.