Information Sets of Multiplicity Codes

Daniel Augot, Françoise Levy-Dit-Vehel, Man Cuong Ngô

Information Sets of Multiplicity Codes

Daniel Augot
INRIA Saclay and LIX
Bâtiment Alan Turing
1 rue Honoré d'Estienne d'Orves
91120 Palaiseau
daniel.augot@inria.fr

Françoise Levy-dit-Vehel
ENSTA ParisTech
828 boulevard des Maréchaux
91762 Palaiseau
INRIA Saclay and LIX
levy@ensta.fr

Cuong M. Ngô
INRIA Saclay and LIX
Bâtiment Alan Turing
1 rue Honoré d'Estienne d'Orves
91120 Palaiseau
manh-cuong.ngo@inria.fr

***Abstract*—We here provide a method for systematic encoding of the Multiplicity codes introduced by Kopparty, Saraf and Yekhanin in 2011. The construction is built on an idea of Kopparty. We properly define information sets for these codes and give detailed proofs of the validity of Kopparty's construction, that use generating functions. We also give a complexity estimate of the associated encoding algorithm.**

***Index Terms*—Locally decodable codes, locally correctable codes, Reed-Muller codes, Multiplicity codes, information set.**

I. Introduction

Locally decodable codes (LDC) allow one to probabilistically retrieve one symbol of a *message* by looking at only a small fraction of its encoding. They were formally introduced by Katz and Trevisan in 2000 [1]. When the local decoding algorithm retrieves a symbol of the *codeword* instead of a message symbol, one speaks of locally correctable codes (LCC). For an extensive treatment of locally decodable and correctable codes, we refer the reader to [2].

For C to be an LCC code, it is only required to have C defined as $C \subset \mathbb{F}_q^n$, while the notion of an LDC code requires that C is provided with an encoding $\mathrm{Enc} : \mathbb{F}_q^k \to \mathbb{F}_q^n$. Considering codes which are $\mathbb{F}_q$-linear subspaces of $\mathbb{F}_q^n$, there is a reduction making an LDC code from an LCC code [2, Lemma 2.3]. This reduction heavily relies on the notion of *Information Set*.

A breakthrough of Kopparty, Saraf and Yekhanin [3] is a construction of high-rate LCCs with sublinear locality. These codes were termed *Multiplicity Codes*, and generalize the Reed-Muller codes, using derivatives.

A technical and practical issue remains, which is to make these codes LDCs. For these codes, the message space and the codeword space do not share the same alphabet, so the standard reduction from [2, Lemma 2.3] can not be applied. The problem was circumvented in [3] by using concatenation.

It is well known that LDCs can be used to build Private Information Retrieval (PIR) schemes, using a standard equivalence between LDCs and PIRs [1]. In [4], for the very particular case of Reed-Muller codes and Multiplicity codes, a better usage of these locally decodable codes in PIR schemes was introduced, using a partitioning of the m-dimensional affine space into few affine hyperplanes. The concatenation solution provided by [3] appears not helpful in this context, since it more or less breaks the underlying affine geometry.

In the Appendix of [5], Kopparty described an idea to make a systematic encoding for Multiplicity codes. We clarify the idea of [5], providing notation and proofs, and solve a unicity problem, necessary to have a valid systematic encoding.

Ngô is supported by INRIA and Alcatel-Lucent *Privacy* ADR

II. Problem statement

Let $q = p^t$ for some $t \in \mathbb{N}^*$ and p prime. We enumerate the field with q elements as $\mathbb{F}_q = \{\alpha_0, \alpha_1, \ldots, \alpha_{q-1}\}$. Considering, for $m \in \mathbb{N}^*$, m indeterminates $X_1, \ldots, X_m$ and m positive integers $i_1, \ldots, i_m$, we use the short-hand notation

$$\begin{aligned} \boldsymbol{X} &= (X_1, \ldots, X_m) & \boldsymbol{X}^{\boldsymbol{i}} &= X_1^{i_1} \cdots X_m^{i_m}, \\ \mathbb{F}_q[\boldsymbol{X}] &= \mathbb{F}_q[X_1, \ldots, X_m] & \boldsymbol{i} &= (i_1, \ldots, i_m) \in \mathbb{N}^m, \\ |\boldsymbol{i}| &= i_1 + \cdots + i_m & \boldsymbol{P} &= (p_1, \ldots, p_m) \in \mathbb{F}_q^m, \end{aligned}$$

i.e. we use bold symbols for vectors, points, etc, and standard symbols for uni-dimensionnal scalars, variables, etc. We denote by

$$\mathbb{F}_q[\boldsymbol{X}]_d = \{F \in \mathbb{F}_q[\boldsymbol{X}];\ \deg F \leq d\}\,.$$

We also let $V = \mathbb{F}_q^m = \{\boldsymbol{P}_1, \ldots, \boldsymbol{P}_n\}$, where $n = q^m$.

A. Reed-Muller codes over $\mathbb{F}_q$ and information sets

We define the following evaluation map

$$\begin{aligned} \mathrm{ev} :\ \mathbb{F}_q[\boldsymbol{X}] &\to \mathbb{F}_q^n \\ F &\mapsto (F(\boldsymbol{P}_1), \ldots, F(\boldsymbol{P}_n))\,. \end{aligned}$$

For an integer $d > 0$, we denote by $\mathbb{F}_q[\boldsymbol{X}]_d$ the set of polynomials of degree less than or equal to d, which has dimension $\binom{m+d}{m}$ over $\mathbb{F}_q$, see [5]. We can now recall the definition of Reed-Muller codes over $\mathbb{F}_q$, also called Generalized Reed-Muller codes [6]:

Definition 1 (Reed-Muller codes over $\mathbb{F}_q$): For $d \leq m(q-1)$, the $d^{th}-$ order *Reed-Muller code over* $\mathbb{F}_q$, RM_d is

$$\mathrm{RM}_d = \{\mathrm{ev}(F)\ \mid F \in \mathbb{F}_q[\boldsymbol{X}]_d\}\,.$$

From now on, we omit "over $\mathbb{F}_q$" and simply say Reed-Muller codes. The evaluation map ev maps $\binom{m+d}{m}$ symbols into n symbols. However, when $d \geq q$, the map ev is not injective, and the dimension k_d of RM_d is less than or equal to $\binom{m+d}{d}$.

A codeword $c \in \mathrm{RM}_d$ can be indexed by integers as $c = (c_1, \ldots, c_n)$ or by points as $c = (c_{\boldsymbol{P}_1}, \ldots, c_{\boldsymbol{P}_n})$, where $c_i = c_{\boldsymbol{P}_i} = F(\boldsymbol{P}_i)$.

Definition 2 (Information set): Let $\mathcal{C}$ be an $[n, k]$ linear code over $\mathbb{F}_q$. An *information set* of $\mathcal{C}$ is a subset $\mathcal{I} \subset \{1, \ldots, n\}$ such that the map:

$$\begin{array}{rccc} \varphi: & \mathcal{C} & \to & \mathbb{F}_q^k \\ & c & \mapsto & (c_i)_{i \in \mathcal{I}} \end{array} \tag{1}$$

is a bijection.

J.D. Key *et al.* [7] gave information sets for Reed-Muller codes, that we recall in the following Theorem.

Theorem 1 ([7]): An information set of RM_d is

$$\left\{ (\alpha_{i_1}, \ldots, \alpha_{i_m}) |\ 0 \le i_l \le q-1;\ 1 \le l \le m;\ \sum_{l=1}^{m} i_l \le d \right\}.$$

We denote this particular information set by $\mathcal{I}_d$, with $\mathcal{I}_d \subset V$. Denote by

$$\mathcal{K}_d = \{(i_1, \ldots, i_m) \mid 0 \le i_l \le q-1,\ 1 \le l \le m,\ \sum_{l=1}^{m} i_l = d\},$$

$$\mathcal{L}_d = \{(i_1, \ldots, i_m) \mid 0 \le i_l \le q-1,\ 1 \le l \le m,\ \sum_{l=1}^{m} i_l \le d\},$$

then we have $k_d = \dim(\mathrm{RM}_d) = |\mathcal{L}_d| = \sum_{l=0}^{d} |\mathcal{K}_l|$ (see [6]).

B. Multiplicity codes

First we recall the notion of Hasse derivative for multivariate polynomials. We write polynomials $F \in \mathbb{F}_q[\boldsymbol{X}] = \mathbb{F}_q[X_1, \ldots, X_m]$ without parentheses and without variables, and $F(\boldsymbol{X})$ (resp. $F(\boldsymbol{P})$) when the evaluation on indeterminates (resp. points) has to be specified. Given $\boldsymbol{i}, \boldsymbol{j} \in \mathbb{N}^m$, we denote:

- $\boldsymbol{i} \le \boldsymbol{j}$ if $i_l \le j_l$ for all $l = 1, \ldots, m$,
- $\boldsymbol{i} < \boldsymbol{j}$ if $\boldsymbol{i} \le \boldsymbol{j}$ and $i_l < j_l$ for some $1 \le l \le m$.

Given $\boldsymbol{i} \in \mathbb{N}^m$, and $F \in \mathbb{F}_q[\boldsymbol{X}]$, the $\boldsymbol{i}$-th *Hasse derivative* of F, denoted by $H(F, \boldsymbol{i})$, is the coefficient of $\boldsymbol{Z}^{\boldsymbol{i}}$ in the polynomial $F(\boldsymbol{X} + \boldsymbol{Z}) \in \mathbb{F}_q[\boldsymbol{X}, \boldsymbol{Z}]$, where $\boldsymbol{Z} = (Z_1, \ldots, Z_m)$. More specifically, let $F(\boldsymbol{X}) = \sum_{\boldsymbol{j} \ge \boldsymbol{0}} f_{\boldsymbol{j}} \boldsymbol{X}^{\boldsymbol{j}}$, then

$$F(\boldsymbol{X} + \boldsymbol{Z}) = \sum_{\boldsymbol{j}} f_{\boldsymbol{j}} (\boldsymbol{X} + \boldsymbol{Z})^{\boldsymbol{j}} = \sum_{\boldsymbol{i}} H(F, \boldsymbol{i})(\boldsymbol{X}) \boldsymbol{Z}^{\boldsymbol{i}},$$

where

$$H(F, \boldsymbol{i})(\boldsymbol{X}) = \sum_{\boldsymbol{j} \ge \boldsymbol{i}} f_{\boldsymbol{j}} \binom{\boldsymbol{j}}{\boldsymbol{i}} \boldsymbol{X}^{\boldsymbol{j}-\boldsymbol{i}},$$

with

$$\binom{\boldsymbol{j}}{\boldsymbol{i}} = \binom{j_1}{i_1} \cdots \binom{j_m}{i_m}.$$

Given $F, G \in \mathbb{F}_q[\boldsymbol{X}]$ and $\boldsymbol{i} \in \mathbb{N}^m$, we have (Leibniz rule [2]):

$$H(F \cdot G, \boldsymbol{i}) = \sum_{\boldsymbol{0} \le \boldsymbol{k} \le \boldsymbol{i}} H(F, \boldsymbol{k}) \cdot H(G, \boldsymbol{i} - \boldsymbol{k}) \tag{2}$$

Now, given a derivation order $s > 0$, we introduce an extended notion of evaluation. For a given $s > 0$, there are $\sigma = \binom{m+s-1}{m}$ Hasse derivatives of a polynomial F : $H(F, \boldsymbol{i})$, $\boldsymbol{i} \in \mathbb{N}^m$, $|\boldsymbol{i}| < s$. Denote by $S = \{\boldsymbol{j} \in \mathbb{N}^m;\ |\boldsymbol{j}| < s\}$, and let $\Sigma = \mathbb{F}_q^S$. An element $x \in \Sigma$ is written as

$$x = (x_{\boldsymbol{j}})_{\boldsymbol{j} \in S}, \quad x_{\boldsymbol{j}} \in \mathbb{F}_q.$$

We generalize the evaluation map at a point $\boldsymbol{P}$:

$$\begin{array}{rccc} \mathrm{ev}^s_{\boldsymbol{P}}: & \mathbb{F}_q[\boldsymbol{X}] & \to & \Sigma \\ & F & \mapsto & (H(F, \boldsymbol{v})(\boldsymbol{P}))_{\boldsymbol{v} \in S} \end{array}$$

and the total evaluation rule is

$$\begin{array}{rccc} \mathrm{ev}^s: & \mathbb{F}_q[\boldsymbol{X}] & \to & \Sigma^n \\ & F & \mapsto & \left(\mathrm{ev}^s_{\boldsymbol{P}_1}(F), \ldots, \mathrm{ev}^s_{\boldsymbol{P}_n}(F)\right). \end{array}$$

Definition 3 (Multiplicity Codes [3]): Given the above evaluation map and a degree $d < sq$, the corresponding *Multiplicity code* is

$$\mathrm{Mult}^s_d = \left\{ \mathrm{ev}^s(F)\ \mid F \in \mathbb{F}_q[\boldsymbol{X}]_d \right\}.$$

In the context of [3] the constraint $d < sq$ is required to ensure that ev^s restricted to $\mathbb{F}_q[\boldsymbol{X}]_d$ is injective.

C. Information sets of Multiplicity codes

The difficulty in defining an information set for a multiplicity code properly is that the $\mathbb{F}_q$-symbols of the message space are not the same as the $\mathbb{F}_q^S$-symbols of the codeword space. Recall that a codeword $c \in \Sigma^n$ can be indexed by points $\boldsymbol{P} \in V$:

$$c = (c_{\boldsymbol{P}})_{\boldsymbol{P} \in V}, \quad c_{\boldsymbol{P}} \in \Sigma.$$

Each $c_{\boldsymbol{P}}$ can be written $c_{\boldsymbol{P}} = ((c_{\boldsymbol{j}})_{\boldsymbol{P}})_{\boldsymbol{j} \in S}$, hence we can write

$$c = (c_{\boldsymbol{j}, \boldsymbol{P}})_{\boldsymbol{j} \in S, \boldsymbol{P} \in V}.$$

We can now define information sets of Multiplicity Codes:

Definition 4 (Information set of a Multiplicity Code): An *information set* of Mult^s_d is a set $\mathcal{I} \subset S \times \mathbb{F}_q^m$ such that the mapping

$$\begin{array}{rccc} \phi: & \mathrm{Mult}^s_d & \to & \mathbb{F}_q^{\mathcal{I}} \\ & c & \mapsto & (c_{\boldsymbol{j}, \boldsymbol{P}})_{(\boldsymbol{j}, \boldsymbol{P}) \in \mathcal{I}} \end{array}$$

is bijective.

In [5], an information set $\mathcal{I}$ of Mult^s_d based on information sets of Reed-Muller codes was suggested, namely, $\mathcal{I} = (\boldsymbol{j}, \mathcal{I}_{d_{\boldsymbol{j}}})_{\boldsymbol{j} \in S}$ where $\mathcal{I}_{d_{\boldsymbol{j}}}$ is the information set of the $d_{\boldsymbol{j}}$-th order Reed-Muller code as in (3), where the degree $d_{\boldsymbol{j}}$ is

$$d_{\boldsymbol{j}} = \min(m(q-1), d - \boldsymbol{j}q),\ \boldsymbol{j} \in S.$$

We prove that $\mathcal{I}$ is an information set in the next two Sections.

III. Systematic encoding algorithm

A. A polynomial decomposition

Given a multi-index $\boldsymbol{j} = (j_1, \dots, j_m)$, let $V_{\boldsymbol{j}}$ be the polynomial

$$V_{\boldsymbol{j}} = \prod_{i=1}^{m} (X_i^q - X_i)^{j_i}.$$

The following decomposition is given in [5] without proof.

Lemma 1: Let $F \in \mathbb{F}_q[\boldsymbol{X}]$ have total degree less than or equal to d, then F can be written as

$$F = \sum_{|\boldsymbol{j}| \leq d/q} F_{\boldsymbol{j}} \cdot V_{\boldsymbol{j}}, \tag{3}$$

for some polynomials $F_{\boldsymbol{j}} \in \mathbb{F}_q[\boldsymbol{X}]_{d_{\boldsymbol{j}}}$. There also exists a polynomial $F_{\boldsymbol{j}_0}$ where $|\boldsymbol{j}_0| = \lfloor d/q \rfloor$ and $\deg(F_{\boldsymbol{j}_0}) = d - \lfloor d/q \rfloor q$.

Proof: We consider a multivariate monomial $X_1^{u_1} \dots X_m^{u_m}$ and write $u_i = t_i q + r_i$ for all $i = 1, \dots, m$. First, we consider just $X_1^{u_1}$:

- if $t_1 = 0$, since $r_1 < q$, we do not need to prove anything;
- if $t_1 > 0$, we have:

$$\begin{aligned} X_1^{u_1} &= X_1^{r_1} \cdot ((X_1^q - X_1) + X_1)^{t_1} \\ &= \sum_{i=0}^{t_1} \binom{t_1}{i} X_1^{r_1 + (t_1 - i)} \cdot (X_1^q - X_1)^i \end{aligned}$$

Similarly, we recursively apply the above reduction with $X_1^{r_1+(t_1-i)}$ where $i = 0, \dots, t_1$, so we finally obtain:

$$\begin{aligned} X_1^{u_1} &= \sum_{i=0}^{t_1'} P_{1,i}(X_1) \cdot (X_1^q - X_1)^i \\ &= P_{1,t_1'}(X_1) \cdot (X_1^q - X_1)^{t_1'} + \sum_{i=0}^{t_1'-1} P_{1,i}(X_1) \cdot (X_1^q - X_1)^i, \end{aligned}$$

where $\deg(P_{1,i}) \leq q - 1$ for $i = 0, \dots, t_1'$, for some t_1'. We see that

$$\deg(P_{1,i}(X_1) \cdot (X_1^q - X_1)^i) < q(i+1) \leq q t_1' < u_1,$$

for all $i = 0, \dots, t_1' - 1$, so the term of degree $u_1 = \deg(X_1^{u_1})$ belongs to $P_{1,t_1'}(X_1) \cdot (X_1^q - X_1)^{t_1'}$, hence $\deg(P_{1,t_1'}) = u_1 - q t_1' = r_1 + q(t_1 - t_1')$.

Since $0 \leq r_1$, as $\deg(P_{1,t_1'}) \leq q-1$, it follows that $t_1 - t_1' = 0$, so we have $\deg(P_{1,t_1}) \leq \min(q - 1, u_1 - q t_1)$. Doing the same thing with the other variables $X_2, \dots, X_m$, we obtain:

$$\begin{aligned} X_1^{u_1} \cdots X_m^{u_m} &= \left(\sum_{i_1=0}^{t_1} P_{1,i_1}(X_1) \cdot (X_1^q - X_1)^{i_1} \right) \cdots \\ &\quad \cdots \left(\sum_{i_m=0}^{t_m} P_{m,i_m}(X_m) \cdot (X_m^q - X_m)^{i_m} \right) \\ &= \sum_{\boldsymbol{i} \leq (t_1, \dots, t_m)} B_{\boldsymbol{i}}(X_1, X_2, \dots, X_m) \cdot V_{\boldsymbol{i}}(X_1, X_2, \dots, X_m), \end{aligned}$$

where $B_{\boldsymbol{i}}(\boldsymbol{X}) = P_{1,i_1}(X_1) \cdots P_{m,i_m}(X_m)$ and $\deg(B_{\boldsymbol{i}}) = \sum_{j=1}^m \deg(P_{j,i_j}) \leq \min(m(q-1), \sum_{j=1}^m (u_j - q i_j)) = \min(m(q-1), (\sum_{j=1}^m u_j) - |\boldsymbol{i}|q)$. Since a multivariate polynomial is the sum of multivariate monomials, we obtain the result. We also note that if there would not exist an $F_{\boldsymbol{j}_0}$ such that $|\boldsymbol{j}_0| = \lfloor d/q \rfloor$ and $\deg(F_{\boldsymbol{j}_0}) = d - \lfloor d/q \rfloor q$, then the degree of the RHS of (3) would not be equal to $\deg(F)$. ∎

We prove the *uniqueness* of the $F_{\boldsymbol{j}}$'s in (3) in the next Section.

B. Corresponding systematic encoding

Considering a point $\boldsymbol{P} \in V$, we have $V_{\boldsymbol{j}}(\boldsymbol{P} + \boldsymbol{Z}) = \sum_{\boldsymbol{i}} H(V_{\boldsymbol{j}}, \boldsymbol{i})(\boldsymbol{P}) \boldsymbol{Z}^{\boldsymbol{i}}$, and,

$$\begin{aligned} V_{\boldsymbol{j}}(\boldsymbol{P} + \boldsymbol{Z}) &= \prod_{i=1}^{m} ((P_i + Z_i)^q - (P_i + Z_i))^{j_i} \\ &= \prod_{i=1}^{m} (Z_i^q - Z_i)^{j_i} = \boldsymbol{Z}^{\boldsymbol{j}} \prod_{i=1}^{m} \left(Z_i^{q-1} - 1 \right)^{j_i}. \end{aligned}$$

So, we have proved the following [5]:

$$H(V_{\boldsymbol{j}}, \boldsymbol{i})(\boldsymbol{P}) = \begin{cases} 0 & \boldsymbol{i} \not\geq \boldsymbol{j} \\ (-1)^{|\boldsymbol{i}|} & \boldsymbol{i} = \boldsymbol{j}. \end{cases} \tag{4}$$

When we compute the Hasse Derivative of F, we find

$$\begin{aligned} H(F, \boldsymbol{i}) &= \sum_{|\boldsymbol{j}| \leq d/q} H(F_{\boldsymbol{j}} V_{\boldsymbol{j}}, \boldsymbol{i}) \\ &= \sum_{|\boldsymbol{j}| \leq d/q} \sum_{\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{j}} H(F_{\boldsymbol{j}}, \boldsymbol{u}) H(V_{\boldsymbol{j}}, \boldsymbol{v}) \\ H(F, \boldsymbol{i})(\boldsymbol{P}) &= \sum_{|\boldsymbol{j}| \leq d/q} \sum_{\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{j}, \boldsymbol{v} \leq \boldsymbol{j}} H(F_{\boldsymbol{j}}, \boldsymbol{u})(\boldsymbol{P}) H(V_{\boldsymbol{j}}, \boldsymbol{v})(\boldsymbol{P}). \end{aligned}$$

Thanks to (4), the summation reduces to

$$\begin{aligned} H(F, \boldsymbol{i})(\boldsymbol{P}) &= \sum_{\boldsymbol{j} \leq \boldsymbol{i}} \sum_{\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{i}, \boldsymbol{v} \leq \boldsymbol{j}} H(F_{\boldsymbol{j}}, \boldsymbol{u})(\boldsymbol{P}) H(V_{\boldsymbol{j}}, \boldsymbol{v})(\boldsymbol{P}) \\ &= (-1)^{|\boldsymbol{i}|} F_{\boldsymbol{i}}(\boldsymbol{P}) + \\ &\quad \sum_{\boldsymbol{j} < \boldsymbol{i}} \sum_{\boldsymbol{u} + \boldsymbol{v} = \boldsymbol{i}, \boldsymbol{v} \leq \boldsymbol{j}} H(F_{\boldsymbol{j}}, \boldsymbol{u})(\boldsymbol{P}) H(V_{\boldsymbol{j}}, \boldsymbol{v})(\boldsymbol{P}). \end{aligned} \tag{5}$$

Thus we can find the evaluation of $F_{\boldsymbol{i}}$ at $\boldsymbol{P} \in \mathbb{F}_q^m$ if we know:

- $H(F, \boldsymbol{i})(\boldsymbol{P})$;
- the polynomials $F_{\boldsymbol{j}}$ for every $\boldsymbol{j} < \boldsymbol{i}$.

Now, using the information set $\mathcal{I}_{d_j}$ of the Reed-Muller code RM_{d_j} given by Theorem 1, we can determine $F_{\boldsymbol{j}}$ given the values $F_{\boldsymbol{j}}(\boldsymbol{P})$, $\boldsymbol{P} \in \mathcal{I}_{d_j}$. So the set $\mathcal{I}$:

$$\mathcal{I} = (\boldsymbol{j}, \mathcal{I}_{d_j})_{\boldsymbol{j} \in S} \tag{6}$$

enables to find $F_{\boldsymbol{j}}$ from its values on $\mathcal{I}_{d_j}$. Under unicity of (3), we have the following :

Proposition 1: An information set of Mult_d^s is given by (6).

Given a message M of length $k = \binom{m+d}{m}$ over $\mathbb{F}_q$, we consider the polynomial $F \in \mathbb{F}_q[\boldsymbol{X}]_d$ whose list of coefficients is given by M. Then, the classical non-systematic encoding of M is $\mathrm{ev}^s(F) \in \Sigma^n$.

For the systematic encoding, we write the message as $M = (M_{\boldsymbol{j},\boldsymbol{P}})$, where $\boldsymbol{P} \in \mathcal{I}_{d_j}$ and $|\boldsymbol{j}| \leq d/q$, and we *define* F to be the *unique* polynomial such that $H(F,\boldsymbol{j})(\boldsymbol{P}) = M_{\boldsymbol{j},\boldsymbol{P}}$. We then *construct* F according to the above discussion : From the values $H(F,\boldsymbol{j})(\boldsymbol{P})$, we find $F_{\boldsymbol{j}}$ thanks to (5). Then we find F using (3) and finally we evaluate F on the remaining $(\boldsymbol{j},\boldsymbol{P}) \notin \mathcal{I}$. The systematic encoding of M over V is $\mathrm{ev}^s(F)$. We summarize this systematic encoding in Algorithm 1.

Algorithm 1 Systematic encoding algorithm for multiplicity codes

Input: The message $M = (M_{\boldsymbol{i},\boldsymbol{P}})_{(\boldsymbol{i},\boldsymbol{P})\in\mathcal{I}}$ of dimension k.
Output: The systematic encoding of M over V.

1: Determine recursively the polynomials $F_{\boldsymbol{j}} \in \mathbb{F}_q[\boldsymbol{X}]$ with $|\boldsymbol{j}| \leq d/q$, using (5) where $H(F,\boldsymbol{i})(\boldsymbol{P})$ is given by
$$H(F,\boldsymbol{i})(\boldsymbol{P}) = M_{\boldsymbol{i},\boldsymbol{P}}, \quad \boldsymbol{i} \in S.$$
2: Compute the polynomial $F \in \mathbb{F}_q[\boldsymbol{X}]$ as
$$F = \sum_{|\boldsymbol{j}|\leq d/q} F_{\boldsymbol{j}} \cdot V_{\boldsymbol{j}},$$
where $V_{\boldsymbol{j}} = \prod_{i=1}^{m} (X_i^q - X_i)^{j_i}$.
3: **return** $\mathrm{ev}^s(F)$, the systematic encoding of M over V.

IV. Unicity of the decomposition

To have unicity of F constructed from the message $(M_{\boldsymbol{j},\boldsymbol{P}})_{(\boldsymbol{j},\boldsymbol{P})\in\mathcal{I}}$, and full correctness of Algorithm 1, the following statement suffices.

Lemma 2: The decomposition (3) in Lemma 1 is unique.

Proof: We prove this lemma by showing that the size of $\mathcal{I}$ defined by (6) is exactly the dimension k of the code. Assume that $d = rq+t$, hence $r \leq s-1$ and $t < q$ (since $d < sq$). Recall that the dimension of Reed-Muller codes satisfy $k_d = |\mathcal{L}_d| = |\mathcal{I}_d|$. There are some particular cases:

- When $d \geq m(q-1)$, $k_d = q^m$
- When $0 \leq d \leq q-1$, $k_d = \binom{m+d}{m}$
- When $d < 0$, $k_d = 0$.

Since we do not know any closed formula for k_d, we use generating functions (see [8], [9]). First, we give a brief introduction. If $f(x) = \sum_{n\geq 0} a_n x^n$, then we call a_n the n-th coefficient of x^n, and denote it by $a_n = [x^n]f(x)$. Recall that:

$$\frac{1}{(1-x)^k} = \sum_{n\geq 0} \binom{n+k-1}{k-1} x^n. \tag{7}$$

Using

$$\mathcal{K}_d = \{(i_1,\ldots,i_m) \mid 0 \leq i_l \leq q-1;\ 1 \leq l \leq m;\ \sum_{l=1}^{m} i_l = d\},$$

we have a one-to-one mapping between elements $(i_1,\ldots,i_m) \in \mathcal{K}_d$ and monomials $x^{i_1}x^{i_2}\ldots x^{i_m}$ of total degree d and individual degree not greater than $q-1$. Hence, for a degree d, consider the generating function:

$$\begin{aligned} f_m(x) &\triangleq \left(\frac{1-x^q}{1-x}\right)^m \\ &= \underbrace{(1+x+\cdots+x^{q-1})\cdots(1+x+\cdots+x^{q-1})}_{m \text{ times}}, \end{aligned}$$

then the coefficient of x^d of $f_m(x)$ is exactly the cardinality of $\mathcal{K}_d$, with the convention that $\mathcal{K}_d = \emptyset$ when $d > m(q-1)$. From this, we use that $k_d = |\mathcal{L}_d| = |\mathcal{K}_0| + \cdots + |\mathcal{K}_d|$, with:

$|\mathcal{K}_d| = [x^d]f_m(x)$, $|\mathcal{K}_{d-1}| = [x^{d-1}]f_m(x) = [x^d](xf_m(x))$, $\ldots$, $|\mathcal{K}_0| = [1]f_m(x) = [x^d](x^d f_m(x))$.

Therefore:

$$\begin{aligned} k_d &= [x^d](f_m(x) + xf_m(x) + x^2 f_m(x) + \cdots + x^d f_m(x)) \\ &= [x^d]\left(\frac{1-x^{D+1}}{1-x} f_m(x)\right) = [x^d]\left(\frac{f_m(x)}{1-x}\right). \end{aligned}$$

Note that $k_d = |\mathcal{I}_d|$. Similarly as above, we have:

$k_d = [x^d]\frac{f_m(x)}{1-x}$, $k_{d-q} = [x^d]\frac{x^q f_m(x)}{1-x}$, $\ldots$,
$k_{d-rq} = [x^d]\frac{x^{rq} f_m(x)}{1-x}$.

where $d = rq+t$ and $t < q$. For every $\boldsymbol{j}$ we have $\binom{m-1+|\boldsymbol{j}|}{m-1}$ such sets $(\boldsymbol{j},\mathcal{I}_{d_j})$. By (6), it follows that the size of $\mathcal{I}$ is thus

$$|\mathcal{I}| = \sum_{u=0}^{s-1} \sum_{|\boldsymbol{j}|=u} |\mathcal{I}_{d_j}| = \sum_{j=0}^{r} \binom{m-1+j}{m-1} k_{d-jq}, \tag{8}$$

which implies

$$\begin{aligned} |\mathcal{I}| &= [x^d]\frac{f_m(x)}{1-x} + \binom{m-1+1}{m-1} \cdot [x^d]\left(\frac{x^q}{1-x} f_m(x)\right) \\ &\quad + \cdots + \binom{m-1+r}{m-1} \cdot [x^d]\left(\frac{x^{rq}}{1-x} f_m(x)\right) \\ &= [x^d]\left(\frac{\sum_{i=0}^{r}\binom{m-1+i}{m-1}x^{iq}}{1-x} f_m(x)\right). \end{aligned}$$

Using (7), we have:

$$\sum_{i\geq 0} \binom{m-1+i}{m-1} x^{iq} = \frac{1}{(1-x^q)^m}, \quad \text{so}$$

$$\begin{aligned} |\mathcal{I}| &= [x^d]\left(\frac{\sum_{i=0}^{r}\binom{m-1+i}{m-1}x^{iq}}{1-x} f_m(x)\right) \\ &= [x^d]\left(\frac{\sum_{i\geq 0}\binom{m-1+i}{m-1}x^{iq}}{1-x} f_m(x)\right) \\ &= [x^d]\left(\frac{1}{(1-x^q)^m(1-x)} f_m(x)\right) \\ &= [x^d]\left(\frac{1}{(1-x^q)^m(1-x)}\left(\frac{1-x^q}{1-x}\right)^m\right) \\ &= [x^d]\left(\frac{1}{(1-x)^{m+1}}\right) \\ &= \binom{m+d}{m} = k, \end{aligned}$$

as we wanted to prove. To conclude the proof, we consider

$$\begin{array}{rccc} \psi: & \prod_{|\boldsymbol{j}|\leq d/q} \mathbb{F}_q[\boldsymbol{X}]_{d_{\boldsymbol{j}}} & \to & \mathbb{F}_q[\boldsymbol{X}]_d \\ & (F_{\boldsymbol{j}})_{|\boldsymbol{j}|\leq d/q} & \mapsto & \sum_{|\boldsymbol{j}|\leq d/q} F_{\boldsymbol{j}} \cdot V_{\boldsymbol{j}} = F \end{array}$$

Lemma 1 shows that ϕ is surjective. Since we have just proved

$$\dim(\mathbb{F}_q[\boldsymbol{X}]_d) = \binom{m+d}{m} = \sum_{|\boldsymbol{j}|\leq d/q} \dim(\mathbb{F}_q[\boldsymbol{X}]_{d_{\boldsymbol{j}}}),$$

the equality of dimensions of the range and of the domain implies that ψ is bijective, in particular one-to-one. ∎

Note that from Equation (8), we can compute easily the value of k_d recursively from k_{d-iq}'s where $0 \leq i \leq d/q$.

V. Systematic encoding for Derivative Codes

In this Section, we apply the previous results to the particular case of $m = 1$. This boils down to codes generalizing Reed-Solomon codes, using derivatives. These codes have been used in [10], where they were given the name of *Derivative Codes*. Let be given s and d as in Definition 3. In this case, the information sets $\mathcal{I}_{d_j}$ are

$$\mathcal{I}_{d_j} = \{i \mid 0 \leq i \leq d_j\}, \quad j = 0, \ldots, s-1.$$

The systematic encoding is described in Algorithm 2.

Algorithm 2 Systematic encoding algorithm for Derivative codes

Input: The message $M = (M_{i,P})$ of dimension k, where $P \in \mathcal{I}_{d_i}$ and $i < s$.

Output: The systematic encoding of M over $\mathbb{F}_q$.

1: Find the polynomials $F_i \in \mathbb{F}_q[X]$ where $i < s$, such that:

$$M_{i,P} = (-1)^i F_i(P) + \\ + \sum_{j=0}^{i-1} \sum_{v=j}^{i} H(F_j, i-v)(P) H(V_j, v)(P)$$

2: Define the polynomial $F \in \mathbb{F}_q[X]$ as

$$F = \sum_{j<s} F_j \cdot V_j,$$

where $V_j = (X^q - X)^j$.

3: **return** $\mathrm{ev}(F)$, the systematic encoding of M over $\mathbb{F}_q$.

VI. Conclusion

We have defined the notion of information set for Multiplicity codes as $\mathbb{F}_q$-linear codes. We filled in details of the work of Kopparty [5], who introduced a systematic encoding for such codes. Our work also allowed us to propose a new recursive formula for the size of Reed-Muller codes over $\mathbb{F}_q$, that makes use of a combinatorial proof of generating functions. Designing efficient algorithms for fast systematic encoding will be the topic of future work.

VII. Acknowledgment

The third author would like to thank Doron Zeilberger and Louis Joseph Billera for the suggestion of using generating functions in Section IV.

References

[1] J. Katz and L. Trevisan, "On the Efficiency of Local Decoding Procedures for Error-correcting Codes," in *Proceedings of the Thirty-second Annual ACM Symposium on Theory of Computing, STOC '00*, F. Yao and E. Luks, Eds. ACM, 2000, pp. 80–86.

[2] S. Yekhanin, *Locally Decodable Codes*, ser. Foundations and Trends in Theoretical Computer Science. NOW publisher, 2012, vol. 6.

[3] S. Kopparty, S. Saraf, and S. Yekhanin, "High-rate Codes with Sublinear-time decoding," in *Proceedings of the Forty-third Annual ACM Symposium on Theory of Computing, STOC'11*, S. Vadhan, Ed. New York, USA: ACM, 2011, pp. 167–176.

[4] D. Augot, F. Levy-dit-Vehel, and A. Shikfa, "A storage-efficient and robust private information retrieval scheme allowing few servers," in *Cryptology and Network Security - 13th International Conference, CANS 2014, Heraklion, Crete*, ser. Lecture Notes in Computer Science. Springer, 2014, pp. 222–239.

[5] S. Kopparty, "List-decoding multiplicity codes," *Electronic Colloquium on Computational Complexity (ECCC)*, vol. TR12-044, 2012.

[6] T. Kasami, S. Lin, and W. Peterson, "New generalizations of the Reed-Muller codes. I. Primitive codes," *IEEE Trans. Information Theory*, vol. 14, no. 2, pp. 189–199, 1968.

[7] J. Key, T. McDonough, and V. Mavron, "Information sets and partial permutation decoding for codes from finite geometries," *Finite Fields and Their Applications*, vol. 12, no. 2, pp. 232–247, Apr. 2006.

[8] R. P. Stanley, *Enumerative Combinatorics*, ser. Cambridge Studies in Advanced Mathematics. Cambridge University Press, 2011, vol. 1.

[9] H. S. Wilf, *Generatingfunctionology*. A. K. Peters, Ltd. Natick, 2006.

[10] V. Guruswami and C. Wang, "Linear-algebraic list decoding for variants of Reed–Solomon codes," *Information Theory, IEEE Transactions on*, vol. 59, no. 6, pp. 3257–3268, Jun. 2013.

Appendix
Complexity estimates

We give a rough and conservative estimate on the number of arithmetic operations in $\mathbb{F}_q$ needed for systematic encoding. Algorithm 1 finds a unique polynomial $F \in \mathbb{F}_q[\boldsymbol{X}]$ from the $F_{\boldsymbol{i}}$'s, those $F_{\boldsymbol{i}}$'s being found from the $\mathbb{F}_q$-symbols $M_{\boldsymbol{j},\boldsymbol{P}}$ at the $(\boldsymbol{j}, \boldsymbol{P}) \in \mathcal{I} = \left(\boldsymbol{j}, \mathcal{I}_{d_j}\right)_{\boldsymbol{j}\in S}$; then it evaluates back this polynomial F for $(\boldsymbol{j}, \boldsymbol{P}) \notin \mathcal{I}$. But (3) requires expensive multiplications of multivariate polynomials. Yet (5) also enables to bypass the computation of F, working only with $F_{\boldsymbol{j}}$'s, as follows. At step $\boldsymbol{i}$, a first pass consists in going through the points $\boldsymbol{P} \in \mathcal{I}_{d_i}$ to compute $F_{\boldsymbol{i}}(\boldsymbol{P})$. Then $F_{\boldsymbol{i}} \in \mathbb{F}_q[\boldsymbol{X}]_{d_i}$ is uniquely determined by its values on the information set $\mathcal{I}_{d_i}$. Note that $F_{\boldsymbol{i}}$ can be computed by applying the (precomputed) inverse of φ defined in (1), i.e. a matrix-vector product of cost $O(k_{d_i}^2)$. Once $F_{\boldsymbol{i}}$ is computed, using (5) again, the values $H(F, \boldsymbol{i})(\boldsymbol{P})$, for $\boldsymbol{P} \notin \mathcal{I}_{d_i}$ are computed. With $\sigma = |S|$, we have, for each $\boldsymbol{i} \in S$:

1) for each $\boldsymbol{P} \in \mathcal{I}_{d_i}$, $O(\sigma^2)$ for computing $F_{\boldsymbol{i}}(\boldsymbol{P})$ using (5); thus a total of $k_{d_i}\sigma^2$ for all $\boldsymbol{P} \in \mathcal{I}_{d_i}$;
2) $O(k_{d_i}^2)$ for recovering $F_{\boldsymbol{i}}$, using a matrix-vector product;
3) $O(\sigma k_{d_i})$ for computing the σ Hasse derivatives of $F_{\boldsymbol{i}}$, (termwise on $F_{\boldsymbol{i}}$, step-by-step through S);
4) $\tilde{O}(n)$ for at once evaluating $F_{\boldsymbol{i}}$ on all $P \notin \mathcal{I}_{d_i}$, neglecting logarithmic factors (multidimensionnal FFT)
5) for each $\boldsymbol{P} \notin \mathcal{I}_{d_i}$, $O(\sigma^2)$ for computing each $H(F, \boldsymbol{i})(\boldsymbol{P})$ using (5) again, for a total of $(n - k_{d_i})\sigma^2$.

Summing over the $\boldsymbol{i} \in S$, we get a "soft-O" estimate of $\tilde{O}\left(\sum_{\boldsymbol{i}\in S} n\sigma^2 + k_{d_i}^2\right) = \tilde{O}\left(n\sigma^3 + k^2\right)$, with a memory footprint of $O(\sigma n)$ for storing all the $F_{\boldsymbol{i}}$'s and their Hasse derivatives. Note that σn is the size of the output codeword.