ESTIMATING THE DIMENSION OF THE SUBFIELD SUBCODES OF HERMITIAN CODES

SABIRA EL KHALFAOUI AND GÁBOR P. NAGY

Abstract. In this paper, we study the behavior of the true dimension of the subfield subcodes of Hermitian codes. Our motivation is to use these classes of linear codes to improve the parameters of the McEliece cryptosystem, such that key size and security level. The McEliece scheme is one of the promising alternative cryptographic schemes to the current public key schemes since in the last four decades, they resisted all known quantum computing attacks. By analyzing computational data series of true dimension, we concluded that they can be estimated by the extreme value distribution function.

1. Introduction

Recently, there has been a big amount of research addressed to quantum computers that use quantum mechanical techniques to solve hard problems in mathematics [2]. The existence of these powerful machines threaten many of the public-key cryptosystem that are widely in use. Combined with Shor's algorithms [38], this would risk the confidentiality and integrity of today's digital communications. Post-quantum cryptography aims to construct and develop cryptosystem that must resist against quantum computing attacks.

McEliece [28] introduced the first code-based public-key cryptosystem in 1978, where he employed error correcting codes to generate the public and private key with security relying on two aspects: NP-completeness of decoding linear codes and the distinguishing of the chosen codes. The original McEliece scheme was constructed with binary Goppa codes which are the subfield subcodes of generalized Reed-Solomon codes. Even today, this proposal represents a good candidate for post-quantum cryptography [1]. There has been several attempts to find appropriate classes of codes and their parameters, which give rise to a secure and effective cryptosystem, for more details see [27, 31]. In this paper, we study the possibility of the application of subfield subcodes of Hermitian codes in the McEliece scheme. More precisely, we do the first step by investigating the true dimension of these codes for a broad spectrum of parameters, for partial results see [13, 34]. Our main observation is that the true dimension of subfield subcodes of Hermitian codes can be estimated by the extreme value distribution function.

In the literature, several attacks have been proposed against McEliece cryptosystem in general, and against McEliece systems based on AG codes, see [3, 6, 27]. Attacks can be divided into two classes: structural, or key recovery attacks, aimed at recovering the secret code, and decoding, or message recovery attacks, aimed at decrypting the transmitted ciphertext. The generic decoding attack against the McEliece scheme is the information set decoding (ISD) algorithm. The most recent

2010 *Mathematics Subject Classification.* 11T71, 14G50, 94B27.

Key words and phrases. AG code, Hermitian code, subfield subcode, extreme value distribution.

and most effective structural attack against AG code based McEliece systems is the Schur product distinguisher.

The structure of this paper is as follows. In section 2, we review the necessary backgrounds to define subfield subcodes, algebraic geometry codes and Hermitian codes. In section 3, we introduce some tools borrowed from statistics in order to handle our computed data on the true dimension of subfield subcodes of Hermitian codes, the latter being presented in section 4. Our main result is Proposition 5.1 in section 5 which shows the excellent fitting properties of the extreme value distribution to our measurements. In section 6, we applied this estimate to study the development of the key size of Hermitian subfield subcodes.

2. Backgrounds, formulas

In this section, we give an overview on subfield subcodes, AG codes and some of their properties, to find full details please refer to the monographs [17, 40, 41]. Our terminology on coding theory is standard, see [18, 40]. In particular, with an $\mathbb{F}_q$-linear code of length n, we mean a linear subspace of $\mathbb{F}_q^n$.

2.1. **Subfield subcodes.** Let h be an integer and r, q be prime powers with $q = r^h$. Then $\mathbb{F}_r$ is a subfield of $\mathbb{F}_q$ and the field extension $\mathbb{F}_q/\mathbb{F}_r$ has degree h. Let C be an $\mathbb{F}_q$-linear code of length n and dimension k. The $\mathbb{F}_q/\mathbb{F}_r$ subfield subcode of C is defined by

$$C|_{\mathbb{F}_r} = C \cap \mathbb{F}_r^n.$$

The trace polynomial $\mathrm{Tr}(x) = x + x^r + \cdots + x^{r^{h-1}}$ defines a map $\mathbb{F}_q \to \mathbb{F}_r$, which can be extended to a map $\mathbb{F}_q^n \to \mathbb{F}_r^n$ component wise. The trace code of the linear code C is

$$\mathrm{Tr}(C) = \{\mathrm{Tr}(c) \mid c \in C\}.$$

Clearly, both the subfield subcode and the trace code are $\mathbb{F}_r$-linear codes of length n. However, it is in general very hard to determine the true dimension of these new codes. The fascinating result given by Delsarte [8] in 1975 plays a key role for studying the class of the subfield subcodes of linear codes. It established a closed link between subfield subcodes and trace codes:

$$(C|_{\mathbb{F}_r})^\perp = \mathrm{Tr}(C^\perp).$$

Véron [44] used this equation to give the exact dimension formula

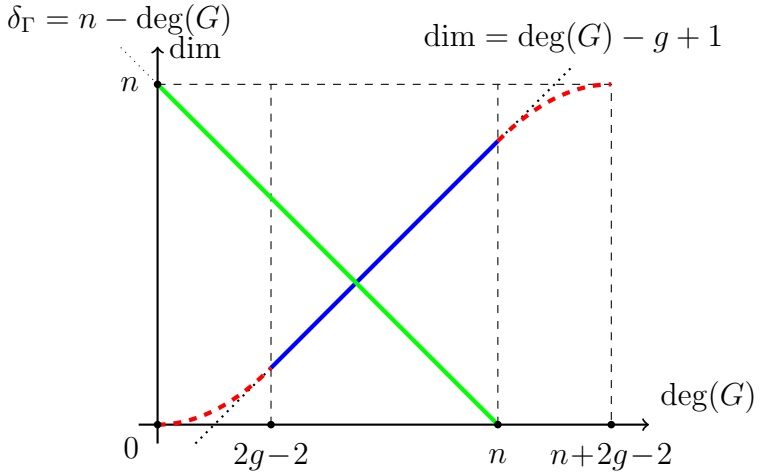
$$\dim_{\mathbb{F}_r}(C|_{\mathbb{F}_r}) = n - h(n-k) + \dim_{\mathbb{F}_r} \ker(\mathrm{Tr}). \tag{1}$$

In particular, we have the trace bound

$$\dim_{\mathbb{F}_r}(C|_{\mathbb{F}_r}) \geq n - h(n-k). \tag{2}$$

2.2. **Algebraic geometry codes.** In this section, we give an overview on the construction of algebraic geometry (AG) codes, which is a version of V.D. Goppa's original construction, since there are many ways to produce linear codes from algebraic curves. Also we give some details on the properties, parameters and duality of AG codes. AG codes are linear codes that use algebraic curves and finite fields for their construction. The construction can be done by evaluating functions (elements of the function field) or by computing residues of differentials. Our notation and terminology on algebraic plane curves over finite fields, their function fields, divisors and Riemann-Roch spaces are standard, see for instance [17, 29, 41].

FIGURE 1. Dimension and designed minimum distance of AG codes



Let q be a prime power and $\mathbb{F}_q$ be the finite field of order q. Let $\mathcal{X}$ be an algebraic curve i.e. an affine or projective variety of dimension one, which is absolutely irreducible and nonsingular and whose defining equations are (homogeneous) polynomials with coefficients in $\mathbb{F}_q$. Let g be the genus of $\mathcal{X}$ and denote by $\mathbb{F}_q(\mathcal{X})$ the function field of $\mathcal{X}$. For a divisor of D of $\mathbb{F}_q(\mathcal{X})$, the Riemann-Roch space is

$$\mathscr{L}(D) = \{f \in \mathbb{F}_q(\mathcal{X}) \mid (f) \succcurlyeq -D\} \cup \{0\},$$

where (f) is the principal divisor of f. The dimension $\ell(D)$ of $\mathscr{L}(D)$ is given by the Riemann-Roch Theorem [41, Theorem 1.1.15]:

$$\ell(D) = \ell(W - D) + \deg D - g + 1, \tag{3}$$

where W is a canonical divisor of $\mathbb{F}_q(\mathcal{X})$. Let G and D be two divisors of $\mathbb{F}_q(\mathcal{X})$ such that $D = P_1 + \cdots + P_n$ is the sum of n distinct rational places of $\mathbb{F}_q(\mathcal{X})$ and $P_i \notin \operatorname{supp}(G)$ for any i. With these data, two types of algebraic geometry codes can be constructed:

$$C_L(D, G) = \{(f(P_1), \cdots, f(P_n)) \mid f \in \mathscr{L}(D)\},$$
$$C_\Omega(D, G) = \{(res_{P_1}(\omega), \cdots, res_{P_n}(\omega)) \mid \omega \in \Omega(G - D)\}.$$

The codes $C_L(D, G)$ and $C_\Omega(D, G)$ are called the *functional* and the *differential codes,* respectively. These two codes are dual to each other. Moreover, the differential code $C_\Omega(D, G)$ is equivalent with the functional code $C_L(D, W + D - G)$. In particular, they have the same dimension and minimum distance, even though this equivalence does not preserve all important properties of the code. The formula

$$k = \ell(G) - \ell(G - D)$$

for the dimension k of $C_L(D, G)$ follows from the Riemann-Roch Theorem, which also provides a lower bound $\delta_\Gamma = n - \deg(G)$ for its minimum distance. The integer δ_Γ is called the *Goppa designed minimum distance* of the AG code.

We illustrate the behavior of the dimension k of $C_L(D, G)$ depending on the degree of the divisor G by Figure 1. In fact, (3) implies the exact value $k = \deg(G) - g + 1$ provided $2g - 2 < \deg(G) < n$. Furthermore, if $\deg(G) > n + 2g - 2$, then $k = n$. In the intervals $[0, 2g - 2]$, and $[n, n + 2g - 2]$, the dimension depends on the specific structure of the divisor G.

2.3. **On the decoding of AG codes.** Algebraic geometry codes are a generalization of Reed-Solomon codes, then it is not extraordinary that they benefit from similar decoding algorithms. The work on the decoding of AG codes seems to begin in 1986 when Driencourt gave a first decoding algorithm for codes on elliptic curves of characteristic 2 [9] correcting $\lfloor(\delta_\Gamma-1)/2\rfloor$ errors. By generalizing the work of Arimoto and Peterson [33] on employing a locator polynomial to decode Reed-Solomon codes, Justesen, Larsen, Jensen, Havemose and Høhold published [20] in 1989 a decoding algorithm for a larger class of AG codes, which can correct up to $\lfloor(\delta_\Gamma - g - 1)/2\rfloor$ errors, moreover in improved version [21] the error capability is increased to $\lfloor(\delta_\Gamma - g/2 - 1)/2\rfloor$. This method was generalized to arbitrary curves by Skorobogatov and Vladut [39], and independently by Krachkovskii [26], then extended by Duursma [10,11] to correct $\lfloor(\delta_\Gamma-1)/2\rfloor-\sigma$ errors, where σ is the Clifford defect of the curve [11, Definition 3.7] (is approximately $g/4$). In 1993, Feng and Rao [15] gave a majority voting scheme allowing a decoding up to $\lfloor(\delta_\Gamma - 1)/2\rfloor$ errors. Duursma generalized this result to all AG codes [12]. An efficient algorithm was described by Sakata, Justesen, Madelung, Jensen and Høhold in [35] using a multidimensional generalization of Massey-Berlekamp algorithm done by Sakata [36]. Kirfel and Pellikaan [22] noticed that one can decode beyond $\lfloor(\delta_\Gamma-1)/2\rfloor$ errors for 1–point AG codes by studying the Weierstrass semigroup. The reader can refer to [18, 19, 32] for more details on decoding methods.

2.4. **Hermitian codes.** The classes of AG codes we study in this paper are defined over the Hermitian curve. Let $\mathbb{F}_q$ be a finite field and define the Hermitian curve $\mathscr{H}_q$ by the affine equation $Y^q + Y = X^{q+1}$. Notice that $\mathscr{H}_q$ is defined over $\mathbb{F}_{q^2}$, that is, its rational points are points of the projective plane $PG(2, q^2)$, satisfying the homogeneous equation $Y^qZ + YZ^q = X^{q+1}$. With respect to the line $Z = 0$ at infinity, $\mathscr{H}_q$ has one infinite point $P_\infty = (0 : 1 : 0)$ and q^3 affine rational points $P_1, \dots, P_{q^3}$. As usual, we also look at the curve $\mathscr{H}_q$ as the smooth curve defined over the algebraic closure $\bar{\mathbb{F}}_{q^2}$. Then, there is a one-to-one correspondence between the points of $\mathscr{H}_q$ and the places of the function field $\bar{\mathbb{F}}_{q^2}(\mathscr{H}_q)$ of $\mathscr{H}_q$.

With a Hermitian code we mean a functional AG code of the form $C_L(D, G)$, where the divisor D is defined as the sum $P_1 + \cdots + P_{q^3}$ affine rational points of $\mathscr{H}_q$. In our investigations, the divisor G can take two forms. In the *1-point case,* we set $G = sP_\infty$ with integer s. In the *degree 3 case,* we put $G = sP$, where P is a place of degree 3. Let P_1, P_2, P_3 be the extensions of P in the constant field extension of $\mathbb{F}_{q^2}(\mathscr{H}_q)$ of degree 3. Then P_1, P_2, P_3 are degree one places of $\mathbb{F}_{q^6}(\mathscr{H}_q)$ and, up to labeling the indices, $P_{j+1} = \mathrm{Frob}(P_j)$ where Frob is the q^2-th Frobenius map and the indices are taken modulo 3. Also, P may be identified with the $\mathbb{F}_{q^2}$-rational divisor $P_1 + P_2 + P_3$ of $\mathbb{F}_{q^6}(\mathscr{H}_q)$. Functional AG codes of the form $C_L(D, sP_\infty)$ and $C_L(D, sP)$ will be called 1-point Hermitian codes, and Hermitian codes over a degree 3 place, respectively. In the 1-point case, the basis of the Riemann-Roch space $\mathscr{L}(sP_\infty)$ can be given explicitly by [40]:

$$\mathcal{M}(s) := \left\{x^i y^j \mid 0 \le i \le q^2 - 1, 0 \le j \le q - 1, qi + (q + 1)j \le s\right\}.$$

In the degree 3 case, the basis of

$$\mathscr{L}(sP) = \left\{ \frac{f}{(\ell_1\ell_2\ell_3)^u} \mid f \in \mathbb{F}_{q^2}[X, Y], \deg f \le 3u, v_{P_i}(f) \ge v \right\} \cup \{0\}.$$

can be computed, see [24]. In this formula, $\ell_i = 0$ is the equation of the tangent line of $\mathscr{H}_q$ at P_i, and $s = u(q+1) - v$, $0 \le v \le q$.

The group $\mathrm{Aut}(\mathscr{H}_q)$ of all automorphisms of $\mathscr{H}_q$ is defined over $\mathbb{F}_{q^2}$. It is a group of projective linear transformations of $PG(2, q^2)$, isomorphic to the projective unitary group $PGU(3, q)$. Furthermore, $\mathrm{Aut}(\mathscr{H}_q)$ acts doubly transitively on the set $\{P_\infty, P_1, \ldots, P_{q^3}\}$ of $\mathbb{F}_{q^2}$-rational points. As it was pointed out in [24], the automorphism group of $\mathscr{H}_q$ acts transitively on the set of degree 3 places of $\mathbb{F}_{q^2}(\mathscr{H}_q)$, as well. Hence, the geometry of a degree 3 place is independent on the choice of P. However, the stabilizer G_P of P in $\mathrm{Aut}(\mathscr{H}_q)$ is not transitive on the set of $q^3 + 1$ rational points. In fact, G_P is a cyclic group of order $q^2 - q + 1$ and the number of G_P-orbits on the set of rational points is $q + 1$. (See [5, 24].)

3. Moments of the extended rate of subfield subcodes

In order to make our notation consistent, we make the following conventions. Let $\mathcal{X}$ be an algebraic curve over $\mathbb{F}_q$ and D, G effective divisors such that the AG code $C_L(D, G)$ is well defined. Assume that the objects δ and γ determine the curve $\mathcal{X}$ and the divisors D, G in a unique way. Let s be an integer and $\mathbb{F}_r$ be a subfield of $\mathbb{F}_q$. Then,

$$C^\gamma_{\delta,r}(s) = C_L(D, sG)|_{\mathbb{F}_r}$$

denotes the $\mathbb{F}_q/\mathbb{F}_r$ subfield subcode of the AG code $C_L(D, sG)$. The length of $C^\gamma_{\delta,r}(s)$ is $n = \deg(D)$.

For the integer s, let

$$R(s) = R^\gamma_{\delta,r}(s) = \frac{\dim_{\mathbb{F}_r} C^\gamma_{\delta,r}(s)}{n}$$

denote the rate of the subfield subcode $C^\gamma_{\delta,r}(s)$. We extend $R^\gamma_{\delta,r}$ to $\mathbb{R}$ in the usual way: $R^\gamma_{\delta,r}(x) = R^\gamma_{\delta,r}(\lfloor x \rfloor)$.

Lemma 3.1. *Let g be the genus of $\mathcal{X}$ and define*

$$\alpha = \left\lceil \frac{n + 2g - 2}{\deg(G)} \right\rceil.$$

Then $R(x)$ is a monotone increasing function, with

$$R(x) = \begin{cases} 0 & \text{for } x < 0, \\ 1 & \text{for } x \ge \alpha. \end{cases}$$

Proof. If $s \deg(G) > n + 2g - 2$, then $\deg(D + W - G) < 0$, and

$$C_\Omega(D, G) \cong C_L(D, D + W - G) = \{0\}.$$

Hence, if $s \ge \alpha$, then $C_L(D, sG) = \mathbb{F}_q^n$ and $C_L(D, sG)|_{\mathbb{F}_r} = \mathbb{F}_r^n$. □

The following observation has been made in [13, Theorem 5.1] for the special case of a one point divisor of a Hermitian curve.

Lemma 3.2. *For $0 \le x < n/(r \deg(G))$, we have $R(x) = 1/n$.*

Proof. As the divisor sG is positive for $s > 0$, the constant vectors are in $C_L(D, sG)|_{\mathbb{F}_r}$ and $R(s) \ge 1/n$ holds. Assume $R(s) > 1/n$, that is, the subfield subcode contains a non constant element $\boldsymbol{v} = (f(P_1), \ldots, f(P_n))$ with $f \in \mathscr{L}(sG)$. Since f cannot have more than $\deg(sG)$ zeros, $\boldsymbol{v}$ cannot have the same entry more than $s \deg(G)$ times. This implies $r \deg(sG) \ge n$. □

Lemma 3.1 implies that we can consider $R(x)$ as the distribution function of some random variable ξ, cf. [37, Definition 1, Section 2.3].

Lemma 3.3. *Let $R(x)$ be the extended rate function of a class of subfield subcodes $C_L(D, sG)|_{\mathbb{F}_r}$. Define the integer α as in Lemma 3.1. Let ξ be a random variable with distribution function $R(x)$. Then*

$$\mathsf{E}(\xi) = \sum_{s=0}^{\alpha} 1 - R(s), \qquad \mathsf{E}(\xi^2) = \sum_{s=0}^{\alpha} (2s+1)(1-R(s)).$$

Proof. This follows from [37, Section 2.6, Corollary 2]. □

Remark. Considered as a distribution function, $R^{\gamma}_{\delta,r}(s)$ has an expectation $\mathsf{E}^{\gamma}_{\delta,r}$, a variance $\mathsf{Var}^{\gamma}_{\delta,r}$ and a standard deviation $\mathsf{D}^{\gamma}_{\delta,r}$. These constants can be computed from the true dimensions of the subfield subcodes using Lemma 3.3 and the well known formulas of random variables.

4. Computed true dimensions of Hermitian subfield subcodes

Let q be a prime power. We say that the object $\delta = q$ determines the Hermitian curve $\mathscr{H}_q$ over $\mathbb{F}_{q^2}$, together with the divisor D which is the sum of affine rational points of $\mathscr{H}_q$. The objects $\gamma =$ 1-pt or $\gamma =$ deg-3 determine the divisor G to be equal either to the rational infinite place P_∞, or the degree 3 Hermitian place P, respectively. That being said, for any integer s and subfield $\mathbb{F}_r$ of $\mathbb{F}_{q^2}$, the Hermitian subfield subcodes

$$C^{\text{1-pt}}_{q,r}(s) = C_L(D, sP_\infty)|_{\mathbb{F}_r}, \qquad C^{\text{deg-3}}_{q,r}(s) = C_L(D, sP)|_{\mathbb{F}_r}$$

are well defined and consistent with the notation of section 3. These codes are $\mathbb{F}_r$-linear codes of length $n = q^3$.

Let $R^{\text{1-pt}}_{q,r}(s)$ and $R^{\text{deg-3}}_{q,r}(s)$ be the true rates of the codes $C^{\text{1-pt}}_{q,r}(s)$ and $C^{\text{deg-3}}_{q,r}(s)$. Using the GAP [16] package `HERmitian` [30], we have been able to compute the true dimension values of the codes $C^{\text{1-pt}}_{q,q}(s)$, $C^{\text{deg-3}}_{q,q}(s)$ for

$$q \in \{2, 3, 4, 5, 7, 8, 9, 11, 13\}$$

and the binary codes $C^{\text{1-pt}}_{q,2}(s)$, $C^{\text{deg-3}}_{q,2}(s)$ for

$$q \in \{2, 4, 8, 16\}.$$

(Cf. [13] for preliminary results on explicit computation of subfield subcodes of Hermitian 1-point codes.)

As given in Lemma 3.3, we computed the expectations $\mathsf{E}^{\text{1-pt}}_{q,q}$, $\mathsf{E}^{\text{1-pt}}_{q,2}$, $\mathsf{E}^{\text{deg-3}}_{q,q}$, $\mathsf{E}^{\text{deg-3}}_{q,2}$, the variances $\mathsf{Var}^{\text{1-pt}}_{q,q}$, $\mathsf{Var}^{\text{1-pt}}_{q,2}$, $\mathsf{Var}^{\text{deg-3}}_{q,q}$, $\mathsf{Var}^{\text{deg-3}}_{q,2}$, and the standard deviations $\mathsf{D}^{\mathbf{1\text{-}pt}}_{q,r}$, $\mathsf{D}^{\text{1-pt}}_{q,2}$, $\mathsf{D}^{\text{deg-3}}_{q,q}$, $\mathsf{D}^{\text{deg-3}}_{q,2}$ for these true rates. The numerical results are shown in Table 1 for $q = 3, 4, 5, 7, 8, 9, 11, 13$ and $r = q$, and in Table 2 for $q = 2, 4, 8, 16$ and $r = 2$. In Figure 2, we present the ratios $\mathsf{E}^{\gamma}_{q,r} \deg(G)/n$ and $\mathsf{D}^{\gamma}_{q,r} \deg(G)/n$, where $\gamma \in \{\text{1-pt, deg-3}\}$. While our data sets are small, these figures motivate the following open problem.

Problem 4.1. *Are there constants $c_1, c_2 > 0$ such that*

$$\mathsf{E}^{\textit{1-pt}}_{q,q} \approx \mathsf{E}^{\textit{deg-3}}_{q,q} \approx c_1 q^3/\deg(G), \qquad \mathsf{D}^{\textit{1-pt}}_{q,q} \approx \mathsf{D}^{\textit{deg-3}}_{q,q} \approx c_2 q^3/\deg(G),$$

where $a \approx b$ means $a/b \to 1$ with $q \to \infty$.

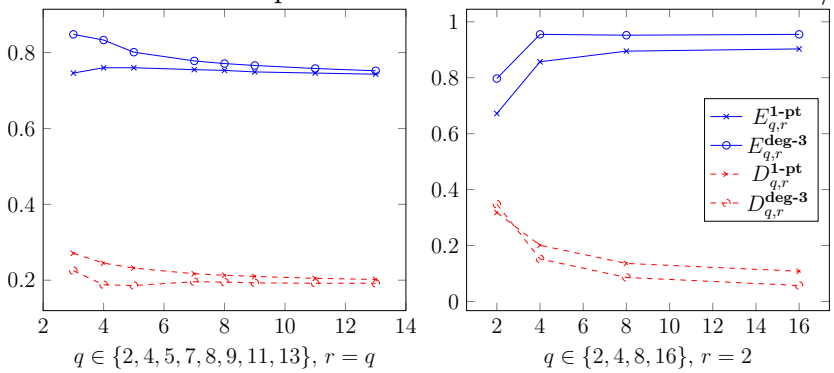
q	1-point codes		Codes over a degree 3 place	
	Expectation	Variance	Expectation	Variance
3	20.15	53.46	7.63	4.09
4	48.66	246.79	17.77	16.02
5	95.04	841.16	33.37	60.18
7	259.10	5 553.32	88.99	503.78
8	385.49	11 862.84	131.61	1 106.63
9	546.30	23 541.65	186.22	2 206.21
11	992.73	74 679.83	336.49	7 262.13
13	1 631.29	197 675.07	550.94	19 807.94

Table 1: Expectations and variances for Hermitian $\mathbb{F}_{q^2}/\mathbb{F}_q$ subfield subcodes

q	1-point codes		Codes over a degree 3 place	
	Expectation	Variance	Expectation	Variance
2	5.38	6.48	2.12	0.86
4	54.86	164.96	20.38	10.52
8	458.22	4 838.52	162.50	216.32
16	3 698.92	195 390.48	1 303.40	6 029.44

Table 2: Expectations and variances for Hermitian $\mathbb{F}_{q^2}/\mathbb{F}_2$ subfield subcodes

FIGURE 2. The ratios of expectations and standard deviations to $n/\deg(G)$



Remark. Our data suggests that for small q, the choice $c_1 = 0.75$ and $c_2 = 0.2$ is sound.

5. Distribution fitting

In general, no explicit formula is known for the true dimension of subfield subcodes of AG codes. Our goal is to use the method of distribution fitting in order to study the behavior of these true dimensions in the case of subfield subcodes of Hermitian codes.

As in the previous sections, we use the notation $\mathscr{H}_q$ for the Hermitian curve over $\mathbb{F}_{q^2}$, P_∞, P for the places of degree 1 and 3, D and $G \in \{P_\infty, P\}$ for the divisors,

and $C^{\gamma}_{q,r}(s)$, $\gamma \in \{\text{1-pt, deg-3}\}$, for the $\mathbb{F}_{q^2}/\mathbb{F}_r$ subfield subcodes $C_L(D, sG)|_{\mathbb{F}_r}$. Then, with fixed q, r and $\gamma \in \{\text{1-pt, deg-3}\}$ the dimensions of the subfield subcodes are given by the extended rate function $R^{\gamma}_{q,r}(x)$.

$$R^{\text{1-pt}}_{q,q}(x), \quad R^{\text{1-pt}}_{q,2}(x), \quad R^{\text{deg-3}}_{q,q}(x), \quad R^{\text{deg-3}}_{q,2}(x).$$

Our goal is to consider these functions as distribution functions and fit some well known probability distribution functions to our experimental rate function $R(x)$.

We obtain numerical results by using the distribution fitting methods offered by MATLAB's Statistics and Machine Learning Toolbox [43]. The technique MLE (Maximum Likelihood Estimation) is a method for estimating the parameters of a probability distribution from a data set. The method finds the parameter values maximizing the logarithm of the likelihood function [14]. In order to compare different distributions for a given data set, one can use the log-likelihood values for a ranking. This is implemented MATLAB's `fitmethis` function [7]. Notice that `fitmethis` also computes the AIC value for each distribution, which stands for Akaike Information Criterion, that measures the quality of a model (distribution) versus the other models. It has the formula

$$AIC = 2l - 2\log(\hat{L})$$

where l is the number of parameters and $\hat{L}$ is the maximum values of the likelihood function. In the case of AIC, smaller values correspond to better fitting distributions (see [23]).

In our comparisons, we restricted ourselves to parametric distributions having at most two parameters, that is, we used MATLAB's `fitmethis` to compare the log-likelihood values of the following distributions: normal, exponential, gamma, logistic, uniform, extreme value, Rayleigh, beta, Nakagami, Rician, inverse Gaussian, Birnbaum-Saunders, log-logistic, log-normal and Weibull. We can summarize the results as follows:

Proposition 5.1. (1) *The best fitting distribution was the extreme value distribution for $R^{\textit{1-pt}}_{q,q}(x)$, $q \in \{4, 5, 7, 8, 9, 11, 13\}$, for $R^{\textit{deg-3}}_{q,q}(x)$, $q \in \{7, 8, 9, 11, 13\}$, and for $R^{\textit{1-pt}}_{8,2}(x)$, $R^{\textit{1-pt}}_{16,2}(x)$, $R^{\textit{deg-3}}_{4,2}(x)$, $R^{\textit{deg-3}}_{8,2}(x)$, and $R^{\textit{deg-3}}_{16,2}(x)$.*

(2) *For the missing cases $R^{\textit{1-pt}}_{2,2}(x)$, $R^{\textit{1-pt}}_{3,3}(x)$, $R^{\textit{deg-3}}_{2,2}(x)$, $R^{\textit{deg-3}}_{3,3}(x)$, $R^{\textit{deg-3}}_{4,4}(x)$, and $R^{\textit{deg-3}}_{5,5}(x)$, the best fitting distribution was the gamma distribution.*

(3) *The second best fitting distribution was the extreme value distribution for $R^{\textit{1-pt}}_{3,3}(x)$, $R^{\textit{deg-3}}_{3,3}(x)$, $R^{\textit{deg-3}}_{4,4}(x)$, $R^{\textit{deg-3}}_{5,5}(x)$.*

Our results show that for $q \geq 3$, among the two-parameter distributions, the extreme value distribution function is a good estimation of the rate function of subfield subcodes of Hermitian codes.

The extreme value distribution is also referred to as Gumbel or type 1 Fisher-Tippet distribution. In probability theory, these are the limiting distributions of the minimum of a large number of unbounded identically distributed random variables. The extreme value distribution function is

$$F(x; \alpha, \beta) = 1 - \exp\left(-\exp\left(\frac{x-\alpha}{\beta}\right)\right),$$

with location parameter $\alpha \in \mathbb{R}$ and a scale parameter $\beta > 0$. The mean μ and the variance σ^2 are

$$\mu = \alpha + \beta\gamma, \qquad \sigma^2 = \frac{\pi^2}{6}\beta^2,$$

where

$$\gamma = \int_1^\infty \left(-\frac{1}{x} + \frac{1}{\lfloor x \rfloor}\right) dx \approx 0.57721566490153$$

is the Euler-Mascheroni constant, see [25, Section 1.4]. With given empirical mean and variance of the data series, the parameters can be computed by

$$\alpha = \mu - \frac{\sqrt{6}\gamma}{\pi}\sigma, \qquad \beta = \frac{\sqrt{6}}{\pi}\sigma.$$

In Figure 3 we visualized the fitting of the extreme value distribution function to our experimental results on the true dimension of subfield subcodes of Hermitian codes.

The occurrence of the extreme value distribution in the context of subfield subcodes of AG codes may be somewhat surprising and we cannot give a plain mathematical explanation for this. However, the rank of random matrices over finite fields is known to be related to the class of Gumbel type distributions, see Cooper's result [4, Theorem 5] for the theoretical background. This theory has been applied to parameter estimates of random erasure codes by Studholme and Blake [42].

6. Application: Estimating the key size of McEliece Cryptosystem

The largest (but not the only) part of the public key of the McEliece cryptosystem is the matrix A which defines the underlying error correction code. A is either the $n \times k$ generator matrix, or the $n \times (n-k)$ parity check matrix. In either case, A may be assumed to be in standard form, which means that the public key is given by $k(n-k)$ elements of $\mathbb{F}_r$. Hence, the key size is

$$\log_2(r)k(n-k).$$

Hence, for a fixed field $\mathbb{F}_r$ and length n, the key size is propotional to $R(1-R)$, see [31]. The true values of $R^\gamma_{q,r}(s)(1-R^\gamma_{q,r}(s))$ can be estimated by $F(x)(1-F(x))$, where $F(x)$ is the extreme value distribution function, see Figure 4.

7. Conclusion and future work

The main goal of this study was to establish an approximating formula of the true dimension of the subfield subcodes of Hermitian codes. We conducted an experimental study to analyze the datasets of the true dimension of the $\mathbb{F}_r$-linear codes $C^{1-pt}_{q,r}(s)$, $C^{deg-3}_{q,r}(s)$ for $q \in \{2,3,4,5,7,8,9,11,13,16\}$, $r = 2$ or $r = q$, and s is an integer parameter running from 0 to $q^3 + (q+1)(q-2)$. This analysis helped us to derive new properties of their structure and led to an approach that might be useful for further research and applications.

Theoretically, the main contribution of this work is the set up of statistical formulas such as moment and expectation by mean of the extended rate function of the underlying classes of subfield subcodes of Hermitian codes.

From a statistical perspective, the main result is the comparison of the fitting of our datasets of true dimensions to well known distribution functions of MATLAB's Statistics and Machine Learning Toolbox, using the method of `fitmethis`.

FIGURE 3. Estimating the extended rate function by extreme value distribution for subfield subcodes Hermitian codes

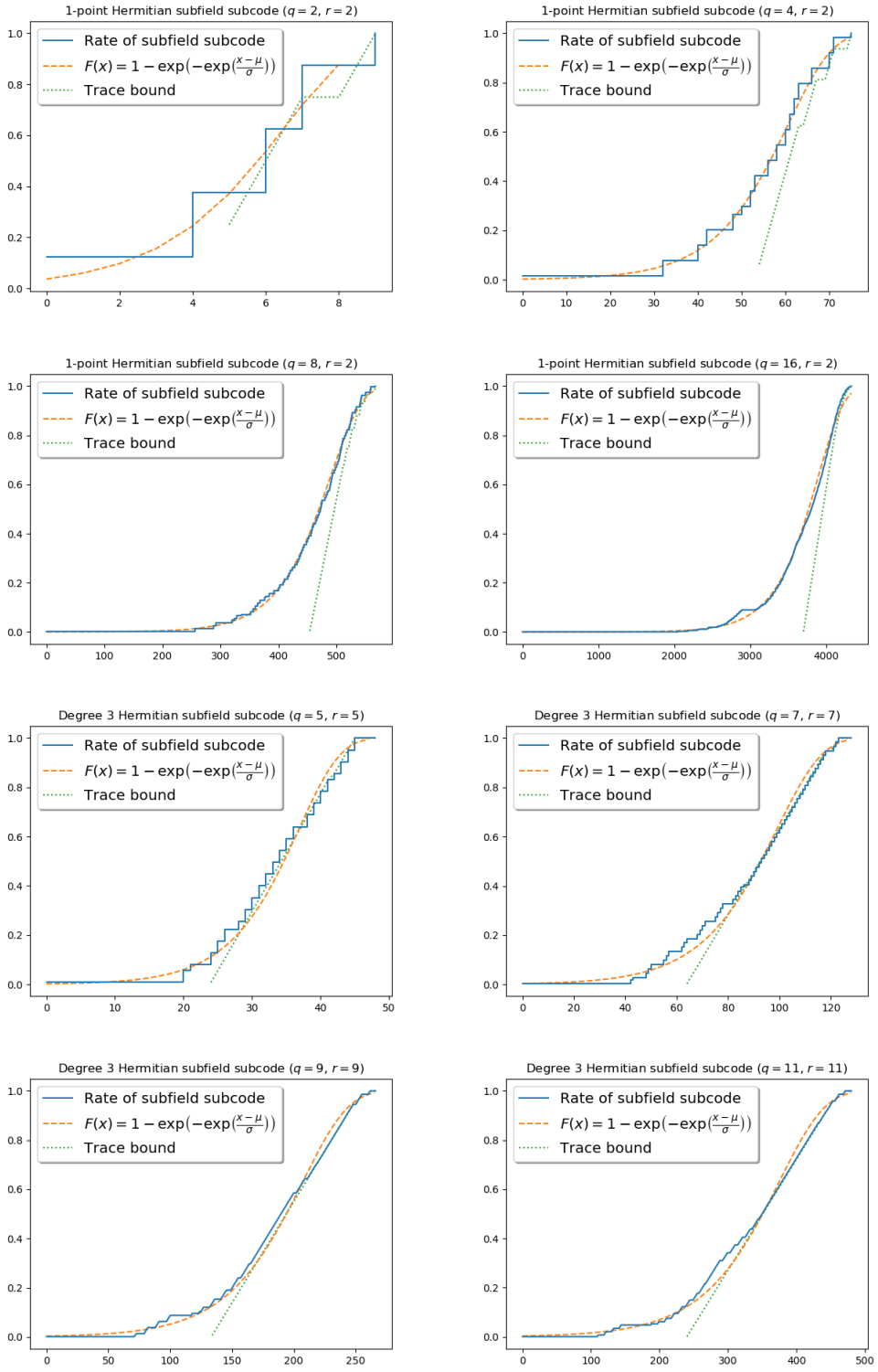
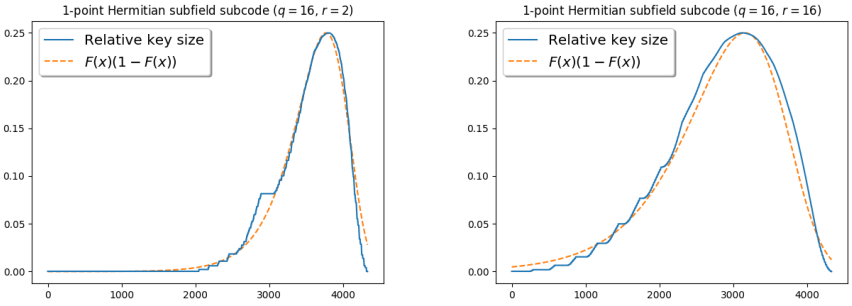


FIGURE 4. Estimating the key size $n^2R(1-R)$



We found that the extreme value distribution is the best fitting one for $q > 5$ and the second best fitting distribution for smalles values of q. Also the gamma and the normal distributions have good fitting properties. Our proposal is to use the extreme value distribution function to estimate the true dimension of subfield subcodes of Hermitian codes. In the last section of this paper, we applied this formula to give an approximation for the key size of the McEliece scheme, depending on the parameter s.

In the future, we aim to replace Goppa codes in McEliece's original version with a family of codes that permit to reduce the public key size and to increase the code rate by maintaining a given level of security. Therefore, we intend to analyze McEliece cryptosystems based on subclasses of subfield subcodes of Hermitian codes. Our future work will include experiments, simulations, and security and cryptanalysis of the McEliece scheme in term of its public key size and other parameters. The measurements are based on attacks with supposed lowest complexity, e.g. information set decoding or the Schur product distinguisher.

ACKNOWLEDGMENT

The presented work was carried out within the project "Security Enhancing Technologies for the Internet of Things" 2018-1.2.1-NKP-2018-00004, supported by the National Research, Development and Innovation Fund of Hungary, financed under the 2018-1.2.1-NKP funding scheme. Partially supported by NKFIH-OTKA Grants 119687 and 115288.

The authors would like to thank Levente Buttyán (Budapest University of Technology, Hungary) for motivating discussions and Mátyás Barczy (University of Szeged, Hungary) for his help to deal successfully with the concepts from probability theory and statistics.

REFERENCES

[1] G. Alagic, J. Alperin-Sheriff, D. Apon, D. Cooper, Q. Dang, Y.-K. Liu, C. Miller, D. Moody, et al., *Status report on the first round of the NIST post-quantum cryptography standardization process*, US Department of Commerce, National Institute of Standards and Technology, 2019.

[2] F. Arute, K. Arya, R. Babbush, D. Bacon, J. C. Bardin, R. Barends, R. Biswas, S. Boixo, F. G. S. L. Brandao, D. A. Buell, B. Burkett, Y. Chen, Z. Chen, B. Chiaro, R. Collins, W. Courtney, A. Dunsworth, E. Farhi, B. Foxen, A. Fowler, C. Gidney, M. Giustina, R. Graff, K. Guerin, S. Habegger, M. P. Harrigan, M. J. Hartmann, A. Ho, M. Hoffmann, T. Huang, T. S. Humble, S. V. Isakov, E. Jeffrey, Z. Jiang, D. Kafri, K. Kechedzhi, J. Kelly, P. V. Klimov, S. Knysh, A. Korotkov, F. Kostritsa, D. Landhuis, M. Lindmark, E. Lucero, D. Lyakh, S. Mandrà, J. R. McClean, M. McEwen, A. Megrant, X. Mi, K. Michielsen, M. Mohseni, J. Mutus, O. Naaman, M. Neeley, C. Neill, M. Y. Niu, E. Ostby, A. Petukhov, J. C. Platt, C. Quintana, E. G. Rieffel, P. Roushan, N. C. Rubin, D. Sank, K. J. Satzinger, V. Smelyanskiy, K. J. Sung, M. D. Trevithick, A. Vainsencher, B. Villalonga, T. White, Z. J. Yao, P. Yeh, A. Zalcman, H. Neven, and J. M. Martinis, *Quantum supremacy using a programmable superconducting processor*, Nature **574** (2019), no. 7779, 505–510.

[3] M. Baldi, M. Bianchi, and F. Chiaraluce, *Security and complexity of the mceliece cryptosystem based on quasi-cyclic low-density parity-check codes*, IET Information Security **7** (2013), no. 3, 212–220.

[4] C. Cooper, *On the distribution of rank of a random matrix over a finite field*, Proceedings of the Ninth International Conference "Random Structures and Algorithms" (Poznan, 1999), 2000, pp. 197–212. MR1801132

[5] A. Cossidente, G. Korchmáros, and F. Torres, *On curves covered by the Hermitian curve*, J. Algebra **216** (1999), no. 1, 56–76. MR1694594

[6] A. Couvreur, I. Márquez-Corbella, and R. Pellikaan, *Cryptanalysis of McEliece cryptosystem based on algebraic geometry codes and their subcodes*, IEEE Trans. Inform. Theory **63** (2017), no. 8, 5404–5418. MR3683571

[7] F. de Castro, *fitmethis, Version 1.3.0.0*, 2020. MATLAB Central File Exchange.

[8] P. Delsarte, *On subfield subcodes of modified Reed-Solomon codes*, IEEE Trans. Information Theory **IT-21** (1975), no. 5, 575–576. MR0411819

[9] Y. Driencourt, *Some properties of elliptic codes over a field of characteristic 2*, International conference on applied algebra, algebraic algorithms, and error-correcting codes, 1985, pp. 185–193.

[10] I. M. Duursma, *Algebraic decoding using special divisors*, IEEE transactions on information theory **39** (1993), no. 2, 694–698.

[11] ———, *Decoding–codes from curves and cyclic codes* (1993).

[12] ———, *Majority coset decoding*, IEEE transactions on information theory **39** (1993), no. 3, 1067–1070.

[13] S. El Khalfaoui and G. P. Nagy, *On the dimension of the subfield subcodes of 1-point Hermitian codes*, Advances in Mathematics of Communications **0** (2019), no. 0, 0, available at `arxiv:1906.10444`.

[14] S. R Eliason, *Maximum likelihood estimation: Logic and practice*, Sage, 1993.

[15] G.-L. Feng and T. R. N. Rao, *Decoding algebraic-geometric codes up to the designed minimum distance*, IEEE Transactions on Information Theory **39** (1993), no. 1, 37–45.

[16] *GAP – Groups, Algorithms, and Programming, Version 4.10.2*, The GAP Group, 2019.

[17] J. W. P. Hirschfeld, G. Korchmáros, and F. Torres, *Algebraic curves over a finite field*, Princeton Series in Applied Mathematics, Princeton University Press, Princeton, NJ, 2008. MR2386879

[18] T. Hoholdt and R. Pellikaan, *On the decoding of algebraic-geometric codes*, IEEE Transactions on Information Theory **41** (1995), no. 6, 1589–1614.

[19] T. Høholdt, J. H. Van Lint, and R. Pellikaan, *Algebraic geometry codes*, Handbook of coding theory **1** (1998), no. Part 1, 871–961.

[20] J. Justesen, K. J. Larsen, H. E. Jensen, A. Havemose, and T. Hoholdt, *Construction and decoding of a class of algebraic geometry codes*, IEEE Transactions on Information Theory **35** (1989), no. 4, 811–821.

[21] J. Justesen, K. J. Larsen, H. E. Jensen, and T. Hoholdt, *Fast decoding of codes from algebraic plane curves*, IEEE Transactions on Information Theory **38** (1992), no. 1, 111–119.

[22] C. Kirfel and R. Pellikaan, *The minimum distance of codes in an array coming from telescopic semigroups*, IEEE Transactions on information theory **41** (1995), no. 6, 1720–1732.

[23] S. Konishi and G. Kitagawa, *Information criteria and statistical modeling*, Springer Science & Business Media, 2008.
[24] G. Korchmáros and G. P. Nagy, *Hermitian codes from higher degree places*, J. Pure Appl. Algebra **217** (2013), no. 12, 2371–2381. MR3057317
[25] S. Kotz and S. Nadarajah, *Extreme value distributions*, Imperial College Press, London, 2000. Theory and applications. MR1892574
[26] V. Yu. Krachkovskii, *Decoding of codes on algebraic curves*, Conference odessa, 1988.
[27] P. Loidreau, *Strengthening mceliece cryptosystem*, International conference on the theory and application of cryptology and information security, 2000, pp. 585–598.
[28] R. J McEliece, *A public-key cryptosystem based on algebraic* (1978).
[29] A. J. Menezes, I. F. Blake, X. Gao, R. C. Mullin, S. A. Vanstone, and T. Yaghoobian, *Applications of finite fields*, Vol. 199, Springer Science & Business Media, 2013.
[30] G. P. Nagy and S. El Khalfaoui, *HERmitian, Computing with divisors, Riemann-Roch spaces and AG-odes of Hermitian curves, Version 0.1*, 2019. GAP package.
[31] R. Niebuhr, M. Meziani, S. Bulygin, and J. Buchmann, *Selecting parameters for secure mceliece-based cryptosystems*, International Journal of Information Security **11** (2012Jun), no. 3, 137–147.
[32] R. Pellikaan, *On the efficient decoding of algebraic-geometric codes*, Eurocode92, 1993, pp. 231–253.
[33] W. Peterson, *Encoding and error-correction procedures for the bose-chaudhuri codes*, IRE Transactions on Information Theory **6** (1960), no. 4, 459–470.
[34] F. Piñero and H. Janwa, *On the subfield subcodes of Hermitian codes*, Designs, codes and cryptography **70** (2014), no. 1-2, 157–173.
[35] S. Sakata, J. Justesen, Y. Madelung, H. E. Jensen, and T. Hoholdt, *Fast decoding of algebraic-geometric codes up to the designed minimum distance*, IEEE Transactions on Information Theory **41** (1995), no. 6, 1672–1677.
[36] S. Sakata, *Extension of the berlekamp-massey algorithm to n dimensions*, Information and Computation **84** (1990), no. 2, 207–239.
[37] A. N. Shiryaev, *Probability. 1*, 3rd ed., Graduate Texts in Mathematics, vol. 95, Springer, New York, 2016. Translated from the fourth (2007) Russian edition by R. P. Boas and D. M. Chibisov. MR3467826
[38] P. W. Shor, *Polynomial-time algorithms for prime factorization and discrete logarithms on a quantum computer*, SIAM Review **41** (1999), no. 2, 303–332, available at `https://doi.org/10.1137/S0036144598347011`.
[39] A. N. Skorobogatov and S. G. Vladut, *On the decoding of algebraic-geometric codes*, IEEE Transactions on Information Theory **36** (1990), no. 5, 1051–1060.
[40] S. A. Stepanov, *Codes on algebraic curves*, Springer Science & Business Media, 2012.
[41] H. Stichtenoth, *Algebraic function fields and codes*, Vol. 254, Springer Science & Business Media, 2009.
[42] C. Studholme and I. F. Blake, *Random matrices and codes for the erasure channel*, Algorithmica **56** (2010), no. 4, 605–620. MR2581065
[43] Inc. The MathWorks, *Statistics and machine learning toolbox*, Natick, Massachusetts, United State, 2019.
[44] P. Véron, *Proof of conjectures on the true dimension of some binary Goppa codes*, Des. Codes Cryptogr. **36** (2005), no. 3, 317–325. MR2162582

Bolyai Institute, University of Szeged, Aradi vértanúk tere 1, H-6720 Szeged, Hungary
Email address: `sabira@math.u-szeged.hu`

Department of Algebra, Budapest University of Technology and Economics, Egry József utca 1, H-1111 Budapest, Hungary

Bolyai Institute, University of Szeged, Aradi vértanúk tere 1, H-6720 Szeged, Hungary
Email address: `nagyg@math.u-szeged.hu`