The syzygy distinguisher

Hugues Randriam
ANSSI, Laboratoire de cryptographie
& Télécom Paris

Preliminaries

The McEliece cryptosystem (1978)

Alice

secret structural data

t-error decoding algorithm

m

e.g. binary Goppa code

error-correcting code C
generator matrix **G**

Bob

message **m**
noise $\mathbf{e} \leftarrow_{\$} \{|\mathbf{e}| = t\}$
ciphertext $\mathbf{y} = \mathbf{mG} + \mathbf{e}$

y

Eve

???

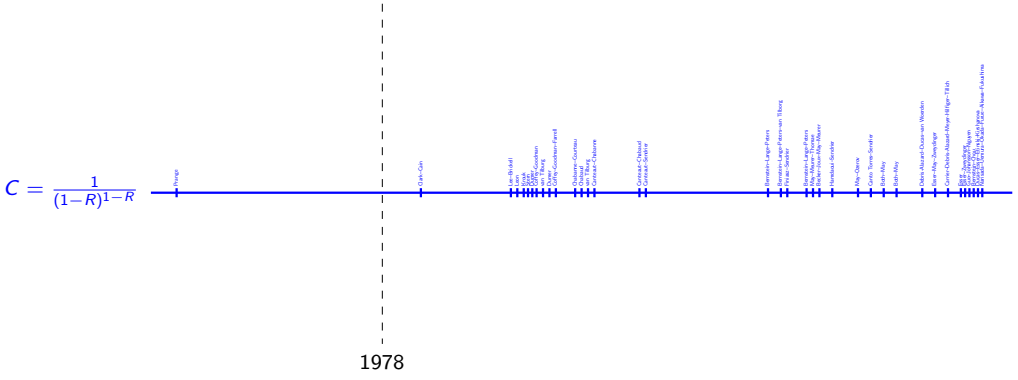
(1) **G** looks like a generic generator matrix
(2) decoding a generic linear code is hard

Note:
(1) ad hoc problem, trapdoor similar to those in today's multivariate cryptography
(2) well-studied problem, NP-hard, believed to be quantum-resistant

Stability of McEliece cryptanalysis

Asymptotic complexity for rate R, length $n \to \infty$ codes: $(C + o(1))^{\frac{n}{\log n}}$

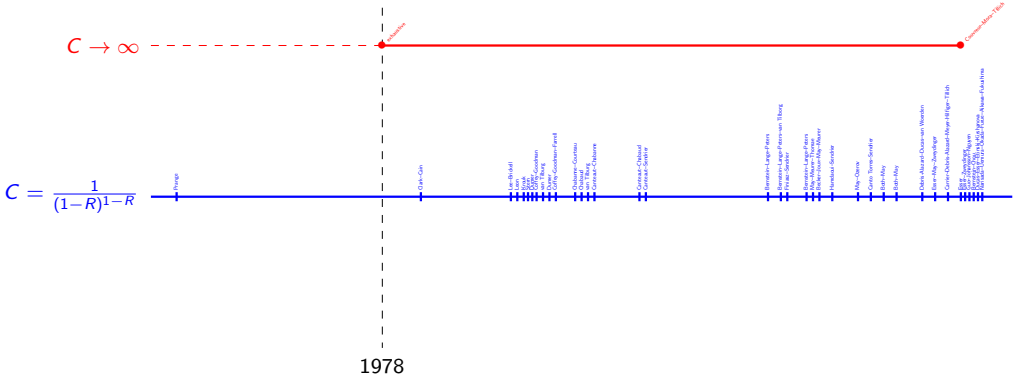
Blue: information set decoding — improving C would be a major result!



Stability of McEliece cryptanalysis

Asymptotic complexity for rate R, length $n \to \infty$ codes: $(C + o(1))^{\frac{n}{\log n}}$

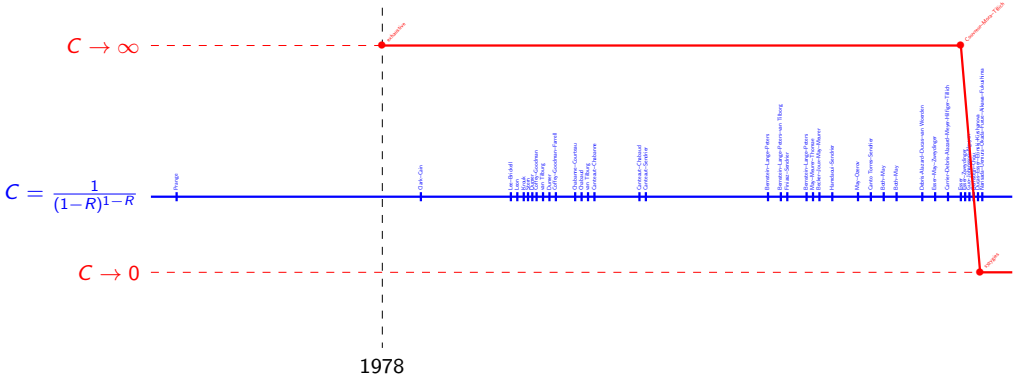
Blue: information set decoding — improving C would be a major result!
Red: Goppa structure distinguisher/recovery
(unmentioned results only work for extreme regimes or other types or codes, or need additional information)



Stability of McEliece cryptanalysis

Asymptotic complexity for rate R, length $n \to \infty$ codes: $(C + o(1))^{\frac{n}{\log n}}$

Blue: information set decoding — improving C would be a major result!
Red: Goppa structure distinguisher/recovery
(unmentioned results only work for extreme regimes or other types or codes, or need additional information)



continuous incremental improvements vs. sudden leaps, potentially devastating

Preliminaries — codes

- $\mathbb{F}$ field $\to$ $\mathbb{F}^n$ product algebra (componentwise multiplication)

$$\mathbf{x} = (x_1, \ldots, x_n),\ \mathbf{y} = (y_1, \ldots, y_n) \quad \to \quad \mathbf{xy} = \mathbf{x} * \mathbf{y} = (x_1 y_1, \ldots, x_n y_n)$$

- a $[n, k]$-code (or $[n, k]_{\mathbb{F}}$-code) is a k-dimensional linear subspace $\mathsf{C} \subseteq \mathbb{F}^n$
- generalized Reed-Solomon code: for $\mathbf{x}, \mathbf{y} \in \mathbb{F}^n$, all x_i distinct, all y_i nonzero,

$$\mathrm{GRS}_k(\mathbf{x}, \mathbf{y}) = \langle \mathbf{y}, \mathbf{yx}, \ldots, \mathbf{yx}^{k-1} \rangle_{\mathbb{F}} = \{ \mathbf{y} f(\mathbf{x}) : \ f(X) \in \mathbb{F}[X]_{<k} \} \subseteq \mathbb{F}^n$$

with generator matrix the generalized Vandermonde matrix

$$\mathbf{G} = \mathbf{V}_k(\mathbf{x}, \mathbf{y}) = \begin{pmatrix} y_1 & \ldots & y_n \\ y_1 x_1 & \ldots & y_n x_n \\ \vdots & & \vdots \\ y_1 x_1^{k-1} & \ldots & y_n x_n^{k-1} \end{pmatrix} \in \mathbb{F}^{k \times n}$$

Preliminaries — more codes

- $\mathbb{F}_q \subseteq \mathbb{F}_{q^m}$ extension of finite fields, e.g. $q = 2$, $m = 12$
- alternant code: for $\mathbf{x}, \mathbf{y} \in (\mathbb{F}_{q^m})^n$, all x_i distinct, all y_i nonzero,

$$\begin{aligned}\mathrm{Alt}_t(\mathbf{x}, \mathbf{y}) &= \mathrm{GRS}_t(\mathbf{x}, \mathbf{y})^\perp \cap (\mathbb{F}_q)^n \\ &= \{\mathbf{c} \in (\mathbb{F}_q)^n : \ c_1 y_1 x_1^j + \cdots + c_n y_n x_n^j = 0 \quad (0 \le j < t)\}\end{aligned}$$

- this is a $[n, (\ge)n - mt]_q$-code
- Goppa code:

$$\mathrm{Gop}(\mathbf{x}, g) = \mathrm{Alt}_{\deg(g)}(\mathbf{x}, g(\mathbf{x})^{-1})$$

  for $g(X) \in \mathbb{F}_{q^m}[X]$ nonvanishing on $\mathbf{x}$, e.g. g irreducible
- we have efficient decoding algorithms for all these codes, provided we know the structural data $(\mathbf{x}, \mathbf{y})$ or $(\mathbf{x}, g)$ from which they were constructed.

Given a generator matrix $\mathbf{G}$, can we decide if it is that of an alternant/Goppa code? If so, can we recover (some) corresponding $(\mathbf{x}, \mathbf{y})$ or $(\mathbf{x}, g)$?

The square distinguisher (and slightly beyond)

The square distinguisher — products and powers of codes

- $\mathbb{F}^n$ endowed with componentwise multiplication $\rightarrow$ product of codes:

$$C, C' \subseteq \mathbb{F}^n \quad \rightarrow \quad CC' = C * C' = \langle \mathbf{c}\mathbf{c}' : \mathbf{c} \in C, \mathbf{c}' \in C' \rangle_{\mathbb{F}}$$

- powers $C^{\langle r \rangle} = C^{*r}$ defined inductively: $C^{\langle 0 \rangle} = \mathbb{F} \cdot \mathbf{1}$, $C^{\langle r+1 \rangle} = C^{\langle r \rangle} C$
- $\mathbf{c}_1, \ldots, \mathbf{c}_k$ basis of C (rows of generator matrix) $\rightarrow$ evaluation map

$$\begin{array}{rcl} \mathbb{F}[X_1, \ldots, X_k] & \rightarrow & \mathbb{F}^n \\ X_i & \mapsto & \mathbf{c}_i \end{array}$$

  maps subspace of homogeneous forms $\mathbb{F}[X_1, \ldots, X_k]_r$ onto $C^{\langle r \rangle}$
- in particular, the $\mathbf{c}_i \mathbf{c}_j$ $(1 \leq i \leq j \leq k)$ generate $C^{\langle 2 \rangle}$, and

$$\dim C^{\langle 2 \rangle} \leq \min\left(n, \frac{k(k+1)}{2}\right)$$

- Cascudo-Cramer-Mirandola-Zémor (2015): for random C, this inequality is an equality with high probability

The square distinguisher — FGOPT11, MP12, MT22

Theorem

For any q, m, n, t, there is an explicit positive constant $T = T_{\mathrm{Gop}}(q, m, n, t)$ such that the dual code C of any Goppa code with these parameters satisfies

$$\dim \mathsf{C}^{\langle 2 \rangle} \leq \min\left(n,\ \frac{k(k+1)}{2} - T\right).$$

Moreover, experimental evidence shows that for most parameter sets, this inequality is an equality with overwhelming probability.
(+ similar result for non-Goppa alternant codes, with another explicit T_{Alt})

Compare with the expected $\dim \mathsf{C}^{\langle 2 \rangle} = \min\left(n,\ \frac{k(k+1)}{2}\right)$ for (the dual of) a random code. Moreover, in all cases, $\dim \mathsf{C}^{\langle 2 \rangle}$ is efficiently computable.

$\rightarrow$ This provides a distinguisher in the regime $n > \frac{k(k+1)}{2} - T$, i.e. when the square of the dual of the alternant/Goppa code does not fill the whole space. In turn, this condition implies that this dual has small rate, or equivalently, that the primal code has high rate: typically 0.96 for n of cryptographic size.

The square distinguisher — some ideas behind

- fundamental example: $\mathsf{C} = \mathsf{GRS}_k(\mathbf{x}, \mathbf{y})$, $\mathbf{c}_1 = \mathbf{y}$, $\mathbf{c}_2 = \mathbf{y}\mathbf{x}$, $\ldots$, $\mathbf{c}_k = \mathbf{y}\mathbf{x}^{k-1}$
  $\rightarrow$ $\mathbf{c}_i\mathbf{c}_j = \mathbf{y}^2\mathbf{x}^{i+j-2}$ $\rightarrow$ $\mathsf{C}^{\langle 2\rangle} = \mathsf{GRS}_{2k-1}(\mathbf{x}, \mathbf{y}^2)$
  $\rightarrow$ $\dim \mathsf{C}^{\langle 2\rangle} = \min(n, 2k-1)$ can be much smaller than $\min(n, \frac{k(k+1)}{2})$
- this comes from quadratic relations between codewords:

$$\mathbf{c}_i\mathbf{c}_j - \mathbf{c}_{i'}\mathbf{c}_{j'} = \mathbf{0} \qquad \text{whenever } i + j = i' + j'$$

- for any C, such relations live in $I_2(\mathsf{C}) = \ker(\mathbb{F}[X_1, \ldots, X_k]_2 \longrightarrow \mathsf{C}^{\langle 2\rangle})$, so

$$\dim \mathsf{C}^{\langle 2\rangle} = \frac{k(k+1)}{2} - \dim I_2(\mathsf{C})$$

- recall $\mathsf{Alt}_t(\mathbf{x}, \mathbf{y}) = \mathsf{GRS}_t(\mathbf{x}, \mathbf{y})^\perp \cap (\mathbb{F}_q)^n$; set $\mathsf{C} = \mathsf{Alt}_t(\mathbf{x}, \mathbf{y})^\perp$; then generically

$$\mathsf{C}_{\mathbb{F}_{q^m}} = \mathsf{GRS}_t(\mathbf{x}, \mathbf{y}) \oplus \mathsf{GRS}_t(\mathbf{x}^q, \mathbf{y}^q) \oplus \cdots \oplus \mathsf{GRS}_t(\mathbf{x}^{q^{m-1}}, \mathbf{y}^{q^{m-1}})$$

- thus $I_2(\mathsf{C}_{\mathbb{F}_{q^m}})$ contains the quadratic relations of all these GRS (and possibly some more), which contributes to make $\dim_{\mathbb{F}_{q^m}}(\mathsf{C}_{\mathbb{F}_{q^m}})^{\langle 2\rangle}$ small
- compatibility with extension of scalars: $\dim_{\mathbb{F}_q} \mathsf{C}^{\langle 2\rangle} = \dim_{\mathbb{F}_{q^m}}(\mathsf{C}_{\mathbb{F}_{q^m}})^{\langle 2\rangle}$

Remark: $I_2(\mathsf{C})$ and $\mathsf{C}^{\langle 2\rangle}$ are equivalent regarding dimension, so we can work with whichever is more convenient for computations, proofs, etc.

Improvements upon and around the square distinguisher

- Bardet-Mora-Tillich (2023) combine ideas from the square distinguisher, shortening/filtration arguments, and a careful Gröbner basis modeling, to get structural recovery attacks in some specific regimes.
- Other approach? In the alternant/Goppa case, even if $\mathsf{C}^{\langle 2 \rangle}$ fills the space, $I_2(\mathsf{C})$ is not a random space of quadratic relations: after extension of scalars, it contains uncommonly short relations such as $\mathbf{c}_i\mathbf{c}_j - \mathbf{c}_{i'}\mathbf{c}_{j'}$. Moreover, these short relations involve the structural basis: uncovering them could possibly lead to an attack. To exploit this, one needs:
  1. a good notion of length/weight/rank for quadratic relations
  2. then solve a nonlinear problem akin to `MinWeight`/`MinRank`.

  I thought a little bit, but failed at 1. and got discouraged by 2.
- But Couvreur-Mora-Tillich (2024) also had this idea, and they succeeded $\rightarrow$ extension of the distinguisher, with a trade-off between complexity and attainable rate. (Plus another attack in another specific regime.)
- This might be the most promising approach, and there is ongoing work further in this direction. Yet...

Theorem 2.8. *Let X be a set of 7 points in linearly general position in $\mathbb{P}^3$. There are just two distinct Betti diagrams possible for the homogeneous coordinate ring S_X:*

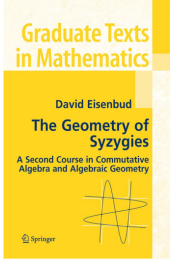
	0	1	2	3
0	1	–	–	–
1	–	3	–	–
2	–	1	6	3

and

	0	1	2	3
0	1	–	–	–
1	–	3	2	–
2	–	3	6	3

In the first case the points do not lie on any curve of degree 3. In the second case, the ideal J generated by the quadrics containing X is the ideal of the unique curve of degree 3 containing X, which is irreducible.

In nature, poisonous creatures will develop bright colors to warn others of their toxicity



Theorem 2.8. *Let X be a set of 7 points in linearly general position in $\mathbb{P}^3$. There are just two distinct Betti diagrams possible for the homogeneous coordinate ring S_X:*

	0	1	2	3
0	1	–	–	–
1	–	3	–	–
2	–	1	6	3

and

	0	1	2	3
0	1	–	–	–
1	–	3	2	–
2	–	3	6	3

In the first case the points do not lie on any curve of degree 3. In the second case, the ideal J generated by the quadrics containing X is the ideal of the unique curve of degree 3 containing X, which is irreducible.

Figure 1: a distinguisher for $[7, 4]$ GRS codes

Aim of this talk: make you understand this, and generalize.

[7,4] GRS codes

For any $[7,4]$-code C with $C^{\langle 2\rangle} = \mathbb{F}^7$, we have

$$\dim I_2(C) = \frac{k(k+1)}{2} - n = 3.$$

If C is a $[7,4]$ GRS code, with basis $\mathbf{c}_1, \mathbf{c}_2, \mathbf{c}_3, \mathbf{c}_4 = \mathbf{y}, \mathbf{yx}, \mathbf{yx}^2, \mathbf{yx}^3$, we have

$$\mathbf{c}_1\mathbf{c}_3 = \mathbf{c}_2^2, \quad \mathbf{c}_1\mathbf{c}_4 = \mathbf{c}_2\mathbf{c}_3, \quad \mathbf{c}_2\mathbf{c}_4 = \mathbf{c}_3^2$$

linearly independent quadratic relations, i.e.

$I_2(C) = \langle Q_1, Q_2, Q_3\rangle$ where $Q_1 = X_1X_3 - X_2^2,\ Q_2 = X_1X_4 - X_2X_3,\ Q_3 = X_2X_4 - X_3^2.$

These Q_1, Q_2, Q_3 are linearly independent over $\mathbb{F}$, i.e. they satisfy no scalar linear relation. But they satisfy linear relations with degree 1 coefficients, a.k.a. syzygies of total degree 3. There are two of them:

$$X_1Q_3 - X_2Q_2 + X_3Q_1 \ = \ X_2Q_3 - X_3Q_2 + X_4Q_1 \ = \ 0.$$

Proof: $\det \begin{pmatrix} X_1 & X_2 & X_3 \\ X_1 & X_2 & X_3 \\ X_2 & X_3 & X_4 \end{pmatrix} = \det \begin{pmatrix} X_2 & X_3 & X_4 \\ X_1 & X_2 & X_3 \\ X_2 & X_3 & X_4 \end{pmatrix} = 0.$

This characterizes $[7,4]$ GRS codes: generic triples of quadratic forms do not admit such syzygies.

The degree 3 syzygy distinguisher

- $S = \mathbb{F}[X_1, \ldots, X_k]$ graded by total degree, $I_r(\mathrm{C}) = \ker(S_r \longrightarrow \mathrm{C}^{\langle r \rangle})$
- syzygies as above lie in the kernel of the "Macaulay matrix"

$$\varphi_3 : I_2(\mathrm{C}) \otimes S_1 \longrightarrow I_3(\mathrm{C}) \subseteq S_3$$

- assume $\mathrm{C}^{\langle 2 \rangle} = \mathbb{F}^n$, so $\dim I_2(\mathrm{C}) = \binom{k+1}{2} - n$ and $\dim I_3(\mathrm{C}) = \binom{k+2}{3} - n$
- we have a certain number of contingent syzygies, forced by dimension:

$$\dim(\ker \varphi_3) \geq (k \dim I_2(\mathrm{C}) - \dim I_3(\mathrm{C}))^+ = (k-1)\left(\tfrac{k(k+1)}{3} - n\right)^+$$

- random code: we make the heuristic that this is an equality w.h.p.
- algebraic code: get a certain number $\widetilde{T} = \widetilde{T}_{\mathrm{Alt}}$ or $\widetilde{T}_{\mathrm{Gop}}$ of structural syzygies, function of q, m, n, t, proven or just guessed/inferred from experiments
- this gives a distinguisher when $\widetilde{T} > (k-1)\left(\tfrac{k(k+1)}{3} - n\right)^+$ i.e. when

$$n > \frac{k(k+1)}{3} - \frac{\widetilde{T}}{k-1}$$

Side-by-side

Square distinguisher:

$$\mathrm{ev} : S_2 \longrightarrow \mathsf{C}^{\langle 2 \rangle} \subseteq \mathbb{F}^n$$

- $\left(\frac{k(k+1)}{2} - n\right)^+$ contingent quadratic relations (from dimension)
- T structural quadratic relations (from alternant/Goppa structure)
- distinguishability threshold:

$$n > \frac{k(k+1)}{2} - T$$

Degree 3 syzygy distinguisher:

$$\varphi_3 : I_2(\mathsf{C}) \otimes S_1 \longrightarrow I_3(\mathsf{C}) \subseteq S_3$$

- $(k-1)\left(\frac{k(k+1)}{3} - n\right)^+$ contingent syzygies (from dimension)
- $\widetilde{T}$ structural syzygies (from alternant/Goppa structure)
- distinguishability threshold:

$$n > \frac{k(k+1)}{3} - \frac{\widetilde{T}}{k-1}$$

Numerical data on the degree 3 distinguisher

Benchmarking distinguishers:

- choose type of code, q, m; set $n = q^m \longrightarrow$ find the largest distinguishable t
- choose type of code, $q, m, t \longrightarrow$ find the shortest distinguishable n

$\text{type}_{m,t}$	$\text{Alt}_{8,3}$	$\text{Alt}_{9,4}$	$\text{Gop}_{6,3}$	$\text{Gop}_{7,3}$	$\text{Gop}_{8,4}$	$\text{Gop}_{9,6}$	$\text{Gop}_{10,7}$
$k = mt$	24	36	18	21	32	54	70
$\widetilde{T}$	16	261	886	1003	4000	26738	54084
$n_{\text{deg 3 dist.}}$	200	437	62	104	223	486	873
$n_{\text{square dist.}}$	–	–	63	106	225	487	876

Table 1: shortest deg 3 syzygy-distinguishable n, for $q = 2$

Remarks:

- for alternant codes, transition to nondistinguishability is abrupt
- for Goppa codes, transition more progressive, both for degree 3 distinguisher and for square distinguisher $\longrightarrow$ can catch slightly shorter n with good proba
- CMT24 does much better for $\text{Gop}_{6,3}$: $n_{\text{CMT24}} = 59$

Higher modules of syzygies

Geometric view on codes

- $\mathbf{G} \in \mathbb{F}^{k\times n}$ generator matrix of C $\leftrightarrow$ its rows $\mathbf{c}_1, \dots \mathbf{c}_k$ form a basis of C
- for any polynomial $f(\mathbf{X}) = f(X_1, \dots, X_k) \in S$ we have

$$\mathrm{ev}(f) = f(\mathbf{c}_1, \dots, \mathbf{c}_k) = (f(\mathbf{p}_1), \dots, f(\mathbf{p}_n))$$

  where $\mathbf{p}_1, \dots \mathbf{p}_n$ are the columns of $\mathbf{G}$
- $\mathsf{C} \sim \mathsf{C}'$ linearly isometric $\leftrightarrow$ $\mathsf{C}' = \mathbf{a}\mathsf{C}^\sigma$ for some $\mathbf{a} \in (\mathbb{F}^n)^\times$, $\sigma \in \mathfrak{S}_n$
- lin. isometry class of C $\leftrightarrow$ eq. class of multiset $\{\overline{\mathbf{p}}_1, \dots, \overline{\mathbf{p}}_n\} \subseteq \mathbb{P}^{k-1}(\mathbb{F})$ mod. projective automorphisms
- the homogeneous coordinate ring of C (or of $\{\overline{\mathbf{p}}_1, \dots, \overline{\mathbf{p}}_n\}$) is

$$\mathsf{C}^{\langle\cdot\rangle} = \bigoplus_{r\geq 0} \mathsf{C}^{\langle r\rangle}$$

- under ev, $\mathsf{C}^{\langle\cdot\rangle}$ is a graded S-module — actually it is a graded quotient of S:

$$0 \longrightarrow I(\mathsf{C}) \longrightarrow S \longrightarrow \mathsf{C}^{\langle\cdot\rangle} \longrightarrow 0$$

  where $I(\mathsf{C}) = \bigoplus_{r\geq 2} I_r(\mathsf{C})$ is the homogeneous ideal of C (or of $\{\overline{\mathbf{p}}_1, \dots, \overline{\mathbf{p}}_n\}$)

General definition of syzygies

- $S = \mathbb{F}[X_1, \ldots, X_k]$ graded by total degree
- M_0 finitely generated graded S-module
- $\mathcal{G} = \{g_1, \ldots, g_N\}$ a minimal system of homogeneous generators of M_0
- F_0 free S-module on $\mathcal{G}$: its elements are formal sums $\sum_i f_i[g_i]$, $f_i \in S$
- $M_1 = \ker(F_0 \twoheadrightarrow M_0)$ is the first module of syzygies of M_0

$$\sum\nolimits_i f_i[g_i] \in M_1 \quad \Longleftrightarrow \quad \sum\nolimits_i f_i g_i = 0 \text{ in } M_0.$$

F_0 and M_1 define M_0 by generators and relations. But M_1 not free in general...

- Iterate: F_i free S-module on a min. syst. of homog. gen. of M_i
- $M_{i+1} = \ker(F_i \twoheadrightarrow M_i)$ is the $(i+1)$-th module of syzygies of M_0.

We will apply this to $M_0 = \mathrm{C}^{\langle\cdot\rangle} \rightarrow M_1 = I(\mathrm{C})$, $M_2 =$"usual" syzygies, ...

Minimal resolution and Betti numbers

- The minimal resolution of M_0 is

$$\cdots \longrightarrow F_i \longrightarrow F_{i-1} \longrightarrow \cdots \longrightarrow F_1 \longrightarrow F_0$$

- Hilbert's syzygy theorem: this terminates, more precisely $F_i = 0$ for $i > k$
- Betti numbers: $\beta_{i,j}$ =number of degree j generators of F_i
  $\longrightarrow$ new invariants for codes
- the square distinguisher actually is a $\beta_{1,2}$-distinguisher ($\beta_{1,2} = \dim(I_2(C))$)
- the degree 3 syzygy distinguisher actually is a $\beta_{2,3}$-distinguisher
- Betti diagram:

	0	1	2	...	k
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$
j	$\beta_{0,j}$	$\beta_{1,j+1}$	$\beta_{2,j+2}$	$\cdots$	$\beta_{k,j+k}$
$j+1$	$\beta_{0,j+1}$	$\beta_{1,j+2}$	$\beta_{2,j+3}$	$\cdots$	$\beta_{k,j+k+1}$
$\vdots$	$\vdots$	$\vdots$	$\vdots$		$\vdots$

minimality: $\mathrm{im}(F_{i+1}) \subseteq (X_1, \ldots, X_k)F_i \longrightarrow$ if F_i generated in deg $\geq D$, then F_{i+1} generated in deg $\geq D+1$

Examples of Betti diagrams

	0	1	2	3
0	1	–	–	–
1	–	3	–	–
2	–	1	6	3

Figure 2: the $[7, 4]_2$ Hamming code

	0	1	2	3	4	5
0	1	–	–	–	–	–
1	–	10	16	–	–	–
2	–	1	5	26	20	5

Figure 3: the $[12, 6]_3$ Golay code

	0	1	2	3	4	5	6	7	8	9	10	11
0	1	–	–	–	–	–	–	–	–	–	–	–
1	–	55	320	891	1408	1210	320	55	–	–	–	–
2	–	1	11	55	220	650	1672	1870	1221	485	110	11

Figure 4: the $[23, 12]_2$ Golay code

Basic properties of Betti diagrams of codes

- Auslander-Buchsbaum: $M_0 = \mathrm{C}^{\langle \cdot \rangle}$ has depth 1 $\rightarrow$ columns range from 0 to $k-1$ (instead of k)
- $\mathrm{C}^{\langle \cdot \rangle}$ is a quotient of $S \rightarrow$ 0-th column $= (1, -, -, \dots)^\top$
- $I_{\leq 1}(\mathrm{C}) = 0 \rightarrow$ 0-th row $= (1, -, -, \dots)$

Definition (Mumford, 1966)

Castelnuovo-Mumford regularity of $\mathrm{C}^{\langle \cdot \rangle}$ is $\max\{j : \exists i,\ \beta_{i,i+j} \neq 0\}$.

Definition (Ran15)

Castelnuovo-Mumford regularity of projective code C is $\min\{j : \mathrm{C}^{\langle j \rangle} = \mathbb{F}^n\}$.

Theorem

These two definitions coincide.

$\rightarrow$ If $\mathrm{C}^{\langle 2 \rangle} = \mathbb{F}^n$, Betti diagram only has rows $0, 1, 2$ (and $I(\mathrm{C})$ gen. in deg $2, 3$).

Effective computation

Only interested in first row of Betti diagram (linear strand of resolution)

- $\beta_{r-1,r} = \dim(M_{r-1,r})$, where $M_{r-1,r}$ lowest degree part of M_{r-1}
- by definition $M_{r-1,r} = \ker(\varphi_r)$ where

$$\varphi_r : M_{r-2,r-1} \otimes S_1 \longrightarrow M_{r-2,r}$$

- but also $M_{r-1,r} = \ker(\psi_r)$ where

$$\psi_r : M_{r-2,r-1} \otimes S_1 \longrightarrow M_{r-3,r-2} \otimes S_2$$

given by $M_{r-2,r-1} \otimes S_1 \subseteq (M_{r-3,r-2} \otimes S_1) \otimes S_1 \to M_{r-3,r-2} \otimes S_2$

$\longrightarrow$ iteratively construct and take kernel of blockwise Macaulay matrix in

$$\mathbb{F}^{k\beta_{r-2,r-1} \times \binom{k+1}{2}\beta_{r-3,r-2}}$$

- looks "Gröbner-ish" but no use of Gröbner basis algorithm

Algebraic codes: structural syzygies

The Eagon-Northcott complex

Let R be a ring, and for $f \geq g$ let $\mathbf{\Phi} \in R^{f \times g}$ define a linear map

$$F = R^f \longrightarrow G = R^g.$$

The Eagon-Northcott complex of $\mathbf{\Phi}$ is

$$0 \to (\mathrm{Sym}^{f-g} G)^\vee \otimes \bigwedge^f F \to \cdots \to G^\vee \otimes \bigwedge^{g+1} F \to \bigwedge^g F \xrightarrow{\bigwedge^g \mathbf{\Phi}} \bigwedge^g G \simeq R.$$

It has length $f - g + 1$, and its r-th term is free of rank $\binom{g+r-2}{r-1}\binom{f}{g+r-1}$.

Under mild hypotheses (e.g. 1-genericity), it is exact: it defines a resolution of the quotient of R defined by the ideal $I_g(\mathbf{\Phi})$ of maximal minors of $\mathbf{\Phi}$.

The Eagon-Northcott complex for $g = 2$

When $g = 2$ and $\mathbf{\Phi} = \begin{pmatrix} x_1 & x_2 & \dots & x_f \\ x'_1 & x'_2 & \dots & x'_f \end{pmatrix}$ we can make everything explicit:

- $I_2(\mathbf{\Phi})$ is generated by the $q_{i,j} = x_i x'_j - x_j x'_i$ so there are $\binom{f}{2}$ of them
- relations between the q_{ij} are generated by the
  - $r_{ijk} = x_i q_{jk} - x_j q_{ik} + x_k q_{ij}$
  - $r'_{ijk} = x'_i q_{jk} - x'_j q_{ik} + x'_k q_{ij}$

  so there are $2\binom{f}{3}$ of them
- relations between the r_{ijk} and r'_{ijk} are generated by the
  - $s_{ijkl} = x_i r_{jkl} - x_j r_{ikl} + x_k r_{ijl} - x_l r_{ijk}$
  - $s'_{ijkl} = x_i r'_{jkl} - x_j r'_{ikl} + x_k r'_{ijl} - x_l r'_{ijk} + x'_i r_{jkl} - x'_j r_{ikl} + x'_k r_{ijl} - x'_l r_{ijk}$
  - $s''_{ijkl} = x'_i r'_{jkl} - x'_j r'_{ikl} + x_k r'_{ijl} - x_l r'_{ijk}$

  so there are $3\binom{f}{4}$ of them
- etc.

We observe that these are "short" relations.

The Eagon-Northcott complex for alternant codes

Let $\mathsf{C} = \mathsf{Alt}_t(\mathbf{x}, \mathbf{y})^\perp$, so

$$\mathsf{C}_{\mathbb{F}_{q^m}} = \mathsf{GRS}_t(\mathbf{x}, \mathbf{y}) \oplus \mathsf{GRS}_t(\mathbf{x}^q, \mathbf{y}^q) \oplus \cdots \oplus \mathsf{GRS}_t(\mathbf{x}^{q^{m-1}}, \mathbf{y}^{q^{m-1}}).$$

Let $e = \lfloor \log_q(t-1) \rfloor$.
Then $I_2(\mathsf{C}_{\mathbb{F}_{q^m}})$ contains $I_2(\mathbf{\Phi})$ where

$$\mathbf{\Phi} = \left(\begin{array}{cccc|cccc|c|cccc} X_1^{(0)} & X_2^{(0)} & \cdots & X_{t-q^e}^{(0)} & X_1^{(1)} & X_2^{(1)} & \cdots & X_{t-q^{e-1}}^{(1)} & \cdots & X_1^{(e)} & X_2^{(e)} & \cdots & X_{t-1}^{(e)} \\ X_{q^e+1}^{(0)} & X_{q^e+2}^{(0)} & \cdots & X_t^{(0)} & X_{q^{e-1}+1}^{(1)} & X_{q^{e-1}+2}^{(1)} & \cdots & X_t^{(1)} & \cdots & X_2^{(e)} & X_3^{(e)} & \cdots & X_t^{(e)} \end{array} \right).$$

Proof:

$$X_a^{(e-u)} X_{q^v+b}^{(e-v)} - X_{q^u+a}^{(e-u)} X_b^{(e-v)} \in I_2(\mathbf{\Phi})$$

evaluates to

$$(\mathbf{y}\mathbf{x}^a)^{q^{e-u}} (\mathbf{y}\mathbf{x}^{q^v+b})^{q^{e-v}} - (\mathbf{y}\mathbf{x}^{q^u+a})^{q^{e-u}} (\mathbf{y}\mathbf{x}^b)^{q^{e-v}} = \mathbf{0}$$

(or: the second row of $\mathbf{\Phi}$ evaluates to $\mathbf{x}^{q^e}$ times its first row).

Shortening

Shortened subcode of C at $\mathcal{S}$: take all codewords that vanish over $\mathcal{S}$, then discard these coordinates.

Proposition

Assume

$$I_2(\mathsf{C}) \supseteq I_2(\mathbf{\Phi})$$

where $\mathbf{\Phi}$ *is a* $2 \times f$ *matrix of linear forms. Let* C_s *be a* s*-shortened subcode of* C*. Then*

$$I_2(\mathsf{C}_s) \supseteq I_2(\mathbf{\Phi}_s)$$

where $\mathbf{\Phi}_s$ *is a* $2 \times (f - s)$ *matrix whose columns are linear combinations of those of* $\mathbf{\Phi}$*, and* $\mathbf{\Phi}_s$ *is* 1*-generic if* $\mathbf{\Phi}$ *is.*

Lower bound on structural syzygies

Theorem

For $\mathsf{C} \in \mathsf{Alt}^{\perp}_{q,m,n,t}$ *set* $e = \lfloor \log_q(t-1) \rfloor$ *and* $f = (e+1)t - \frac{q^{e+1}-1}{q-1}$. *Then*

$$\beta_{r-1,r}(\mathsf{C}_s) \geq (r-1)\binom{f-s}{r} > 0$$

for $r \leq f - s$.

Proof: minimal resolution of C_s contains the Eagon-Northcott complex of $\mathbf{\Phi}_s$.

Remarks:

- f is close to k, so we will be able to shorten a lot
- can slightly improve for Goppa codes

Random codes: contingent syzygies

Linear algebra reminder

In a finite exact sequence of $\mathbb{F}$-vector spaces we have $\sum_i (-1)^i \dim(V_i) = 0$.

If $\varphi : U \longrightarrow V$ is a linear map between $\mathbb{F}$-vector spaces, we define:

- its index:

$$\begin{aligned} \operatorname{ind}(\varphi) &= \dim(U) - \dim(V) \\ &= \dim\ker(\varphi) - \dim\operatorname{coker}(\varphi) \end{aligned}$$

- its (rank) defect:

$$\begin{aligned} \operatorname{def}(\varphi) &= \min(\dim(U), \dim(V)) - \operatorname{rk}(\varphi) \\ &= \min(\dim\ker(\varphi), \dim\operatorname{coker}(\varphi)) \end{aligned}$$

thus

$$\dim\ker(\varphi) = \operatorname{ind}(\varphi)^+ + \operatorname{def}(\varphi), \qquad \dim\operatorname{coker}(\varphi) = \operatorname{ind}(\varphi)^- + \operatorname{def}(\varphi)$$

where for each real x we write $x^+ = \max(x, 0)$ and $x^- = (-x)^+$, so $x = x^+ - x^-$.

More on regularity 2 codes

Definition

The Hilbert series of C is $H_C(z) = \sum_{r \geq 0} z^r \dim C^{\langle r \rangle}$.

Theorem

Set $B_j = \sum_i (-1)^i \beta_{i,j}$ *and* $B(z) = \sum_j B_j z^j$. *Then* $B(z) = (1-z)^k H_C(z)$.

Recall $\varphi_r : M_{r-2,r-1} \otimes S_1 \longrightarrow M_{r-2,r}$

- $\beta_{r-1,r} = \dim \ker(\varphi_r)$
- $\beta_{r-2,r} = \dim \operatorname{coker}(\varphi_r)$.

If C has regularity 2 then

- $H_C(z) = 1 + kz + n(z^2 + z^3 + z^4 + \cdots)$
- $B_r = (-1)^r(\beta_{r-2,r} - \beta_{r-1,r}) = (-1)^{r-1} \operatorname{ind}(\varphi_r)$
- so

$$\operatorname{ind}(\varphi_r) = \left(\frac{k(k+1)}{r} - n \right) \binom{k-1}{r-2}.$$

The minimal resolution conjecture (caution!)

It follows

$$\beta_{r-1,r} \geq \operatorname{ind}(\varphi_r)^+ = \left(\frac{k(k+1)}{r} - n\right)^+ \binom{k-1}{r-2}.$$

Now:

- random linear maps tend to have small defect
- defect is 0 with probability exponentially close to 1 when index grows
- if C is random, we expect φ_r to behave like a random linear map
  $\rightarrow \beta_{r-1,r} = \left(\frac{k(k+1)}{r} - n\right)^+ \binom{k-1}{r-2}$ with high probability.

Backed by the Minimal resolution conjecture (Lorenzini 1993): claims it is so for generic codes

- Bad: counterexamples were found (see e.g. Eisenbud-Popescu)
- Good: only for very specific parameters $\rightarrow$ conjecture still "true enough" for our use

The minimal resolution conjecture (caution!)

It follows

$$\beta_{r-1,r} \geq \operatorname{ind}(\varphi_r)^+ = \left(\frac{k(k+1)}{r} - n\right)^+ \binom{k-1}{r-2}.$$

Now:

- random linear maps tend to have small defect
- defect is 0 with probability exponentially close to 1 when index grows
- if C is random, we expect φ_r to behave like a random linear map, do we??
  $\rightarrow \beta_{r-1,r} = \left(\frac{k(k+1)}{r} - n\right)^+ \binom{k-1}{r-2}$ with high probability.

Backed by the Minimal resolution conjecture (Lorenzini 1993): claims it is so for generic codes $\rightarrow$ over an infinite field

- Bad: counterexamples were found (see e.g. Eisenbud-Popescu)
- Good: only for very specific parameters $\rightarrow$ conjecture still "true enough" for our use

Syzygies and distance properties: the small defect heuristic

Experimental fact

For C *a* $[n, k, d, d^\perp]$*-code, we have*

1. $\beta_{r-1,r}(\mathsf{C}) > 0$, *hence* $\operatorname{def}(\varphi_r) > 0$, *for* $\frac{k(k+1)}{n} \leq r \leq k + 1 - d$
2. $\beta_{r-2,r}(\mathsf{C}) > 0$, *hence* $\operatorname{def}(\varphi_r) > 0$, *for* $d^\perp \leq r \leq \frac{k(k+1)}{n}$.

Conversely for random C*, and* r *out of these intervals,* $\operatorname{def}(\varphi_r) = 0$ *w.h.p.*

Thus:

1. if $d > k + 1 - \frac{k(k+1)}{n}$ we expect $\beta_{r-1,r} = 0$ for $r > \frac{k(k+1)}{n}$
2. if $d^\perp > \frac{k(k+1)}{n}$ we expect $\beta_{r-1,r} = \left(\frac{k(k+1)}{r} - n\right)\binom{k-1}{r-2}$ for $r < \frac{k(k+1)}{n}$.

Remark: these conditions hold for random codes of low enough rate.

More Betti diagrams

	0	1	2	3	4	5	6	7	8	9	10	11
0	1	–	–	–	–	–	–	–	–	–	–	–
1	–	55	319	880	1353	990	–	–	–	–	–	–
2	–	–	–	–	–	330	1617	1870	1221	485	110	11

Figure 5: an idealized $[23, 12]$-code according to the minimal resolution conjecture

	0	1	2	3	4	5	6	7	8	9	10	11
0	1	–	–	–	–	–	–	–	–	–	–	–
1	–	55	319	884	1397	1224	490	121	18	1	–	–
2	–	–	4	44	234	820	1738	1888	1222	485	110	11

Figure 6: a (pseudo)random $[23, 12]_2$-code ($d = 3$, $d^\perp = 4$)

Critical values for r: $d^\perp$, $\frac{k(k+1)}{n}$, $k + 1 - d$

The distinguisher

Principle

Fix an r:

- if distance condition in the small defect heuristic is met, random codes have $\left(\frac{k(k+1)}{r} - n\right)^+ \binom{k-1}{r-2}$ contingent syzygies
- dual alternant/Goppa codes have $\beta^*_{r-1,r}$ structural syzygies (e.g. from Eagon-Northcott complex)

$\rightarrow$ distinguishability threshold $n \geq \left\lceil \frac{k(k+1)}{r} - \frac{\beta^*_{r-1,r}-1}{\binom{k-1}{r-2}} \right\rceil$.

Also works for shortened subcodes $\rightarrow$ smaller n, k, R $\rightarrow$ helps with complexity and with distance condition.

We will restrict to $r > \frac{k(k+1)}{n}$ $\rightarrow$ compute $\beta_{r-1,r}$ and check whether $= 0$ or > 0.
Asymmetry in the heuristic:

- part 1 needed for the distinguisher to work
- part 2 used only for complexity estimate

Practical and non-practical results

- contains the square distinguisher as a special case = the $\beta_{1,2}$-distinguisher
- outperforms CMT24 in all experiments, both in terms of
  - distinguishable parameters
  - efficiency
- largest practically manageable parameters, with a naive, non-optimized Magma implementation: $q, m, n, t = 2, 10, 1024, 10 \rightarrow k = mt = 100$ then for 40-shortened subcodes we consistently find
  - $\beta_{3,4} = 30$ in the Goppa case
  - $\beta_{3,4} = 0$ in the random case
- `Classic McEliece 348864`: $q, m, n, t = 2, 12, 3488, 64 \rightarrow k = mt = 768$
  - 377-shortened dual Goppa codes have $\beta_{49,50} > 0$
  - expect $\beta_{49,50} = 0$ for shortened random codes (distance condition is ok)
  - but complexity estimate $\approx 2^{528}$

How far could we go with a more optimized implementation, sparse linear algebra, etc.?

Asymptotics

Fix a dual rate R, take $n \to \infty$, and $m = \lceil \log_q(n) \rceil$, $k \approx Rn$, $t = \frac{k}{m}$.
Let C be a dual alternant/Goppa code with these parameters.
Recall: for s-shortened subcodes of C,

$$\beta_{r-1,r}(\mathsf{C}_s) > 0$$

for all $r \leq f - s$, where $f = (e+1)t - \frac{q^{e+1}-1}{q-1}$, $e = \lfloor \log_q(t-1) \rfloor$.

Lemma

This f is very close to k, namely

$$k - f \sim R \frac{\log_q \log_q(n)}{\log_q(n)} n.$$

We can distinguish at $\beta_{r-1,r}$, after shortening $s := f - r$ times, as long as $r > \frac{(k-f+r)(k-f+r+1)}{n-f+r}$.

Asymptotics

Theorem

Such codes can be distinguished at $\beta_{r-1,r}$, after shortening $f-r$ times, where

$$r \sim \frac{R^2}{1-R}\left(\frac{\log_q \log_q(n)}{\log_q(n)}\right)^2 n.$$

Complexity is at most

$$q^{\left(\omega \frac{R^2}{1-R}+o(1)\right)\frac{(\log_q \log_q(n))^3}{(\log_q(n))^2} n}$$

which is subexponential in $\frac{n}{\log(n)}$.

Remarks:

- better than ISD algorithms, exponential in $\frac{n}{\log(n)}$
- asymptotic gain $\frac{(\log\log(n))^3}{\log(n)} \to 0$ but very slowly

Conclusion

- Is McEliece broken? — No.
- Will it be broken soon? — I don't know, and I wouldn't bet in any direction.
- Is our understanding of its security stable? — No: these last 3 years saw considerable progress from the algebraic approach (not limited to this work), and this is likely to continue.

TODO:

- Improve implementation, theoretically and practically.
- Provide missing proofs, especially regarding links between Betti numbers and distance properties.
- This is not a black-box distinguisher, it provides a lot of structural information $\rightarrow$ use it (joint with other techniques) for structural recovery?
- Betti numbers are new code invariants. Find other applications, e.g. to the monomial/linear equivalence problem?