Statistical Decoding

Thomas Debris-Alazard and Jean-Pierre Tillich

Inria Saclay,
EPI GRACE

Code-based Cryptography and generic decoding problem

Code-based cryptography: McEliece (1978)...

$\rightarrow$ This is based on the difficulty of decoding for random linear codes

- Input: $\mathscr{C}$ binary code of length n, dimension k with parity-check matrix $\mathrm{H} \in \mathbb{F}_2^{n(1-R)\times n}, \mathrm{y} \in \mathbb{F}_2^n$, $t \in \mathbb{N}$
- Search: e where e has Hamming weight t such that $\mathrm{He}^T = \mathrm{Hy}^T$

$\rightarrow$ Decision problem NP-complete

The simplest information set decoding: Prange algorithm

We are looking for solving $He^T = s^T$:

$$\begin{cases} s_1 & = & h_{1,1}e_1 + h_{1,2}e_2 + \cdots + h_{1,n}e_n \\ & \vdots & \\ s_{n(1-R)} & = & h_{n(1-R),1}e_1 + h_{n(1-R),2}e_2 + \cdots + h_{n(1-R),n}e_n \end{cases}$$

$\rightarrow n(1-R)$ equations with n unknowns.

The simplest information set decoding: Prange algorithm

- If $e_i = 0$ on a set of nR positions i :

$$\begin{cases} s_1 & = h_{1,J_1} e_{J_1} + h_{1,J_2} e_{J_2} + \cdots + h_{1,J_{n(1-R)}} e_{J_{n(1-R)}} \\ & \vdots \\ s_{n(1-R)} & = h_{n(1-R),J_1} e_{J_1} + h_{n(1-R),J_2} e_{J_2} + \cdots + h_{n(1-R),J_{n(1-R)}} e_{J_{n(1-R)}} \end{cases}$$

$\rightarrow$ $n(1-R)$ equations with $n(1-R)$ unknowns .

Exponential complexity as exponentially small probability to pick a set with this property

Information set decoding

Most of the generic decoding algorithms come from the Prange algorithm (1962) :

Lee-Brickell (1988) - Leon (1988) - Stern (1988) - CC (1998) - - MMT (2011)- BLP (2011) - BJMM (2012) - MO (2015)

If $t = o(n)$, all these algorithms have the same asymptotic exponent (Canto-Torres&Sendrier 2016) :

$$\widetilde{\mathcal{O}}\left(2^{-\log_2(1-R)\cdot t}\right)$$

$\rightarrow$ Crucial when it comes to estimate key size of crypto-systems in code-based cryptography

Statistical decoding

It exists an algorithm which does not belong to this family:
Statistical decoding of Al. Jabri (2001)

Studied by R.Overbeck in 2006

No study of its asymptotic complexity!

Results

- Asymptotic exponent of statistical decoding given by a simple formula
- Statistical decoding has a worse complexity than the Prange algorithm for a certain range of error weights.

Statistical decoding: intuition

$$\mathrm{y} = \mathrm{c} + \mathrm{e} \text{ where } \mathrm{c} \in \mathscr{C}$$

$$\mathscr{C}^{\perp} = \{\mathrm{h} \in \mathbb{F}_2^n : \forall \mathrm{c} \in \mathscr{C},\ \langle \mathrm{h}, \mathrm{c} \rangle = 0\}$$

$$\mathrm{h} \in \mathscr{C}^{\perp} \Rightarrow \langle \mathrm{y}, \mathrm{h} \rangle = \langle \mathrm{e}, \mathrm{h} \rangle$$

- If $e_i = 1$ and $h_i = 1$,

$$\langle \mathrm{y}, \mathrm{h} \rangle = 1 \iff \#\left(Supp(\mathrm{e}) \cap Supp(\mathrm{h}) - \{i\}\right) \text{ even}$$

- If $e_i = 0$ and $h_i = 1$

$$\langle \mathrm{y}, \mathrm{h} \rangle = 1 \iff \#\left(Supp(\mathrm{e}) \cap Supp(\mathrm{h}) - \{i\}\right) \text{ odd}$$

$\rightarrow$ Bias of the $\langle y, h \rangle$'s depending on $e_i = 1$ or 0

Notations

- $\mathscr{H}_w \subseteq \{\mathrm{h} \in \mathscr{C}^\perp : \ |\mathrm{h}| = w\}$ where $|\cdot|$ is Hamming weight
- $\mathscr{H}_{w,i} \subseteq \mathscr{H}_w \cap \{\mathrm{m} \in \mathbb{F}_2^n \ : \ m_i = 1\}$

We set a weight w, a noisy codeword $\mathrm{y} = \mathrm{c} + \mathrm{e}$ where $|\mathrm{e}| = t$, $\mathrm{c} \in \mathscr{C}$.

Two distributions

$$e_i = 1 \ : \ q_1(\mathsf{e}, w, i) \stackrel{\triangle}{=} \mathbb{P}_{\mathsf{h} \sim \mathscr{H}_{w,i}} \left(\langle \mathsf{y}, \mathsf{h} \rangle = \langle \mathsf{e}, \mathsf{h} \rangle = 1 \right)$$

$$e_i = 0 \ : \ q_0(\mathsf{e}, w, i) \stackrel{\triangle}{=} \mathbb{P}_{\mathsf{h} \sim \mathscr{H}_{w,i}} \left(\langle \mathsf{y}, \mathsf{h} \rangle = \langle \mathsf{e}, \mathsf{h} \rangle = 1 \right)$$

Two distributions

$$e_i = 1 \;:\; q_1(\mathrm{e}, w, i) \stackrel{\triangle}{=} \mathbb{P}_{\mathrm{h} \sim \mathscr{H}_{w,i}}\left(\langle \mathrm{y}, \mathrm{h} \rangle = \langle \mathrm{e}, \mathrm{h} \rangle = 1\right)$$

$$e_i = 0 \;:\; q_0(\mathrm{e}, w, i) \stackrel{\triangle}{=} \mathbb{P}_{\mathrm{h} \sim \mathscr{H}_{w,i}}\left(\langle \mathrm{y}, \mathrm{h} \rangle = \langle \mathrm{e}, \mathrm{h} \rangle = 1\right)$$

$$q_1(\mathrm{e}, w, i) = \frac{\sum_{j \text{ even}}^{w-1} \binom{t-1}{j}\binom{n-t}{w-1-j}}{\binom{n-1}{w-1}} = \frac{1}{2} + \varepsilon_1$$

$$q_0(\mathrm{e}, w, i) = \frac{\sum_{j \text{ odd}}^{w-1} \binom{t}{j}\binom{n-t-1}{w-1-j}}{\binom{n-1}{w-1}} = \frac{1}{2} + \varepsilon_0$$

Distinguish two distributions

Goal: distinguishing two distributions at distance $|\varepsilon_1 - \varepsilon_0|$

$\rightarrow$ Neymann-Pearson + Chernoff: sample of minimal size

$$P_w \triangleq \frac{\log_2(n)}{(\varepsilon_0 - \varepsilon_1)^2}$$

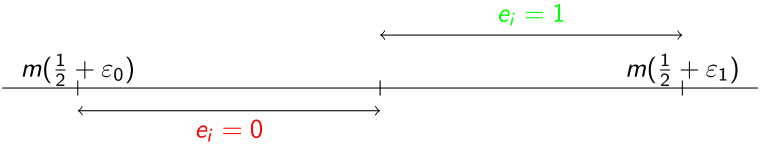
A distinguisher

$$V_m = \sum_{k=1}^{m} \mathrm{sgn}(\varepsilon_1 - \varepsilon_0)\langle y, h^k \rangle \in \mathbb{Z}$$

Proposition (Chernoff bound)

If $e_i = I$ we have:

$$\mathbb{P}\left(\left|V_m - m\,\mathrm{sgn}(\varepsilon_1 - \varepsilon_0)(1/2 + \varepsilon_I)\right| \geq m\frac{|\varepsilon_1 - \varepsilon_0|}{2}\right) \leq 2^{-2m\frac{(\varepsilon_1 - \varepsilon_0)^2}{2\ln(2)}}$$



Statistical Decoding

$\rightarrow$ Difficulty: find enough vectors $\mathbf{h} \in \mathscr{H}_w$ with an algorithm $\texttt{ComputeParity}_w$

$\rightarrow$ We need: $O(P_w)$ where $P_w = \frac{\log_2(n)}{(\varepsilon_1 - \varepsilon_0)^2}$

Proposition

The complexity of statistical decoding is given up to a polynomial factor by:

- *If parity-check equations are already computed:* $O(P_w)$
- *Otherwise:* $O(P_w) + O(|\text{ComputeParity}_w|)$

$$|\text{ComputeParity}_w| \geq P_w$$

Asymptotic exponent

$$\pi(\omega,\tau) \stackrel{\triangle}{=} \underline{\lim}_{n\to+\infty} \frac{1}{n}\log_2 P_w$$

Let h be the binary entropy,

$$h(x) = -x\log_2(x) - (1-x)\log_2(1-x)$$

Theorem

We set $\omega \stackrel{\triangle}{=} \frac{w}{n}$, $\tau \stackrel{\triangle}{=} \frac{t}{n}$ *et* $\gamma \stackrel{\triangle}{=} \frac{1}{\omega}$,

- *If* $\tau \in \left(0, \frac{1}{2} - \sqrt{\omega - \omega^2}\right)$*:*
  $\pi(\omega,\tau) = 2\omega\log_2(r) - 2\tau\log_2(1-r) - 2(1-\tau)\log_2(1+r) + 2h(\omega)$
  where r is the smallest root of $(1-\omega)X^2 - (1-2\tau)X + \omega = 0$.
- *If* $\tau \in \left(\frac{1}{2} - \sqrt{\omega-\omega^2}, \frac{1}{2}\right)$*:* $\pi(\omega,\tau) = h(\omega) + h(\tau) - 1$.

Ingredient one: Bias and Krawtchouk polynomials

Polynomial of degree v, order m, p_v^m defined as:

$$p_v^m(X) = \frac{(-1)^v}{2^v} \sum_{j=0}^{v} (-1)^j \binom{X}{j} \binom{m-X}{v-j}$$

Ingredient one: Bias and Krawtchouk polynomials

Polynomial of degree v, order m, p_v^m defined as:

$$p_v^m(X) = \frac{(-1)^v}{2^v} \sum_{j=0}^{v} (-1)^j \binom{X}{j} \binom{m-X}{v-j}$$

$$\frac{(-2)^{w-2}}{\binom{n-1}{w-1}} p_{w-1}^{n-1}(t) = \varepsilon_0$$

$$-\frac{(-2)^{w-2}}{\binom{n-1}{w-1}} p_{w-1}^{n-1}(t-1) = \varepsilon_1$$

We used results of Mourad E.H Ismail & Plamen Simeonov (1998)

Equations of weight $\frac{Rn}{2}$

We compute the parity-check matrix $\mathbf{H}$ of the code $\mathscr{C}$

Gaussian elimination on $\mathbf{H}$: $[\mathbf{I}_{n(1-R)}|\mathbf{H}']$

The rows have a weight $\frac{Rn}{2}(1+o(1))$

$\rightarrow$ Polynomial cost per solution

Equations of weight $\frac{Rn}{2}$

We compute the parity-check matrix H of the code $\mathscr{C}$

Gaussian elimination on H : $[\mathrm{I}_{n(1-R)}|\mathrm{H}']$

The rows have a weight $\frac{Rn}{2}(1+o(1))$

$\rightarrow$ Polynomial cost per solution

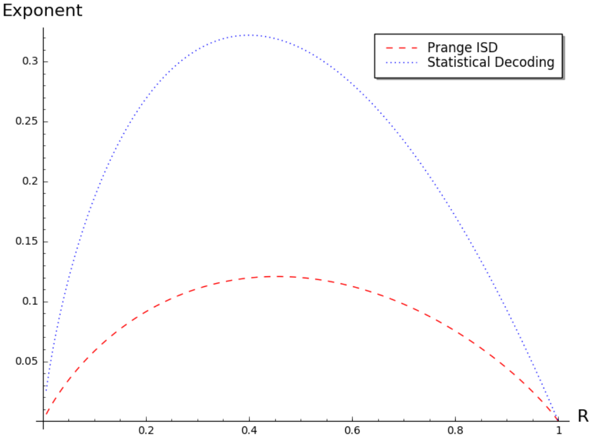
$$\pi^{complete}(\omega,\tau) \stackrel{\triangle}{=} \lim_{n\rightarrow+\infty} \frac{1}{n} \max\left(\log_2 P_w, \log_2 \left|\mathtt{ComputeParity}_w\right| \right)$$

Theorem

Let h be the binary entropy. With the previous algorithm of parity-check equations computation

- *If* $\tau = h^{-1}(1-R)$ *:* $\pi(R/2,\tau) = \pi^{complete}(R/2,\tau) = h(R/2) - R$;
- *If* $\tau = o(1)$ *:* $\pi(R/2,\tau) = \pi^{complete}(R/2,\tau) = -2\tau\log_2(1-R)$.

Comparison of exponents at $h^{-1}(1-R)$



Strategy

We are looking for a number P_w of vectors of $\mathscr{C}^{\perp}$ of weight w

$$P_w \searrow \text{ if } w \searrow$$

Finding parity-check equations of moderate (or small) weight w

Parity-check equations

In a random code there are $C_w \triangleq \frac{\binom{n}{w}}{2^{nR}}$ parity-check equations

$\rightarrow$ We are looking for the smallest w_0 such that:

$$P_{w_0} \leq C_{w_0}$$

The complexity of statistical decoding can not be $< P_{w_0}$.

Surprising fact

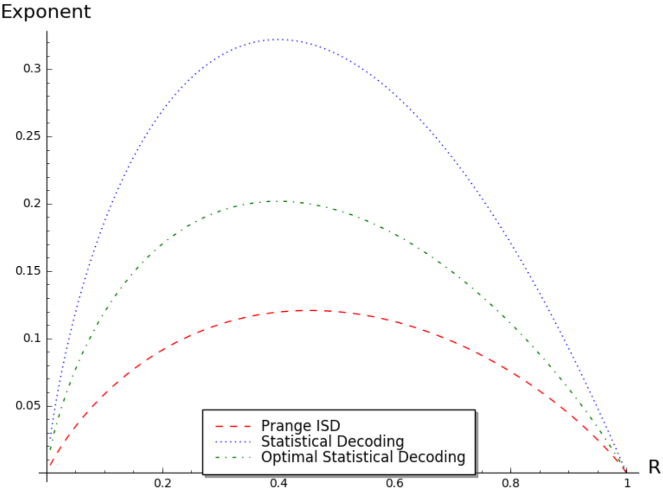
$t = nh^{-1}(1-R)$: number of errors which is the hardest to decode

$$\text{For } \tau = h^{-1}(1-R) : \forall w \geq w_0 : P_w = C_w$$

where

$$w_0 = n\left(\frac{1}{2} - \sqrt{\tau - \tau^2}\right)$$

Optimal exponent on Gilbert-Varshamov bound



Concluding remarks

- Iterative statistical decoding only improves a polynomial factor
- Consider a plenty of parity-check equation weights does not improve the asymptotic exponent
- Other kind of improvements
  $\rightarrow$ Consider a linear combination of information bits?

$$\langle \mathrm{h}, \mathrm{y} \rangle = h_1 \cdot y_1 + \sum_{j=2}^{n} h_j \cdot y_j \rightsquigarrow \langle \mathrm{h}, \mathrm{y} \rangle = \sum_{j \in J} h_j \cdot y_j + \sum_{j \in \overline{J}} h_j \cdot y_j$$

- Statistical decoding arises the issue of other kind of techniques to decode random linear codes.