Estimates of extremal codeword weights of random linear codes over $\mathbf{F}_p$

A. M. Zubkov, V. I. Kruglov

Steklov Mathematical Institute of Russian Academy of Sciences

Saint-Petersburg, 2017

- Let p be any fixed prime number and $\mathbf{F}_p^N = \{X = (x_1, \ldots, x_N) \colon x_1, \ldots, x_N \in \mathbf{F}_p\}$.
- Any k-dimensional subspace $L \subset \mathbf{F}_p^N$ we understood as k-dimensional linear code.
- For any $X = (x_1, \ldots, x_N) \in \mathbf{F}_p^N$ we define its weight as $w(X) = \sum_{k=1}^N I\{x_k \neq 0\}$.
- For a linear code L let

$$v_s(L) = \left|\{X \in L \,:\, w(X) = s\}\right|,$$

the set $\{v_s(L), s = 0, \ldots, N\}$ is the weight spectrum of L.

Let $v_{\leqslant s}(L) = \sum_{u=1}^{s} v_s(L)$ and

$$\mu_*(L) = \min\{w(X)\colon X \in L\backslash\{0\}\}.$$

Theorem 1

If $L \subset \mathbf{F}_p^N$ is a random linear k-dimensional code in $\mathbf{F}_p^N$, then

$$\mathbf{E}v_{\leqslant s}(L) = \frac{p^k - 1}{p^N - 1} \sum_{u=1}^{s} C_N^u (p-1)^u,$$

$$\frac{1}{1 + \frac{p^N - p^k}{p^N - 1}(p-1)(\mathbf{E}v_{\leqslant s}(L))^{-1}} \leqslant$$

$$\leqslant \mathbf{P}\{\mu^*(L) \leqslant s\} \leqslant \min\left\{\mathbf{E}v_{\leqslant s}(L), 1\right\}.$$

It follows from Theorem 1 that

$$\mathbf{E}v_{\leqslant s+1}(L) \geqslant \frac{N-s}{s+1}(p-1)\mathbf{E}v_{\leqslant s}(L).$$

Let $v_{\geqslant s}(L)=\sum_{u=s}^{N} v_s(L)$ and

$$\mu^*(L)=\max\{w(X)\colon X\in L\}.$$

Theorem 2

If $L\subset \mathbf{F}_p^N$ is a random linear k-dimensional code in $\mathbf{F}_p^N$, then

$$\mathbf{E}v_{\geqslant s}(L)=\frac{p^k-1}{p^N-1}\sum_{u=s}^{N} C_N^u(p-1)^u$$

and

$$\frac{1}{1+\frac{p^N-p^k}{p^N-1}(p-1)(\mathbf{E}v_{\geqslant s}(L))^{-1}}\leqslant$$

$$\leqslant \mathbf{P}\{\mu^*(L)\leqslant s\}\leqslant \min\left\{\mathbf{E}v_{\geqslant s}(L),1\right\}.$$

Theorem 3 (Zubkov, Serov 2012)

Let $H(x,r) = x\ln\frac{x}{r} + (1-x)\ln\frac{1-x}{1-r}$*,* $\mathrm{sgn}(x) = \frac{x}{|x|}$ *for* $x \neq 0$ *and* $\mathrm{sgn}(0) = 0$*, let* $\{C_{N,r}(m)\}_{m=0}^{N}$ *be increasing sequences defined as follows:*

$$C_{N,r}(0) = (1-r)^N,\ C_{N,r}(N) = 1 - r^N,$$

$$C_{N,r}(m) = \Phi\left(\mathrm{sgn}\left(\tfrac{m}{N} - r\right)\sqrt{2NH\left(\tfrac{m}{N}, r\right)}\right), 1 \leqslant m < N.$$

Then for $m = 0, 1, \ldots, N-1$ *and for* $r \in (0,1)$

$$C_{N,r}(m) \leqslant \sum\nolimits_{u=0}^{m} C_N^u r^u (1-r)^{N-u} \leqslant C_{n,r}(m+1)$$

and equalities take place only for $C_{N,r}(0)$ *and* $C_{N,r}(N)$*.*

It follows from the Theorem 3 that typical values of minimal non-zero codeword weight $\mu_*(L)$ of random uniformly distributed k-dimensional linear code L in $\mathbf{F}_p^N$ are concentrated near the minimal root $s < \frac{N(p-1)}{p}$ of the equation

$$H\left(\frac{s}{N}, \frac{p-1}{p}\right) \approx \frac{1}{N}\left(k \ln p - \ln(4\pi k \ln p)\right).$$

In particular, if dimension k of the random code L and dimension N of the space $\mathbf{F}_p^N$ are growing proportionally, then the typical value s of the minimal non-zero codeword weight also is growing proportionally to N.

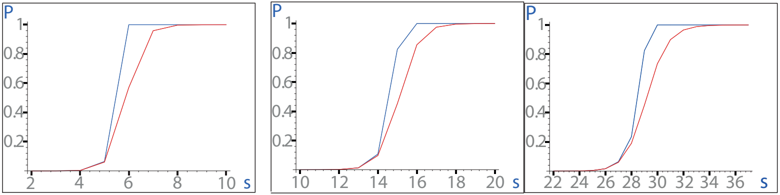
Analogously, typical values of maximal codeword weight $\mu^*(L)$ of random uniformly distributed k-dimensional linear code L in $\mathbf{F}_p^N$ are concentrated near the maximal root $s > \frac{N(p-1)}{p}$ of the same equation

$$H\left(\frac{s}{N}, \frac{p-1}{p}\right) \approx \frac{1}{N}\,\left(k \ln p - \ln(4\pi k \ln p)\right).$$

Remind that $H(x,r) = x \ln \frac{x}{r} + (1-x) \ln \frac{1-x}{1-r}$, so

$$H(x,r) = H(1-x, 1-r).$$

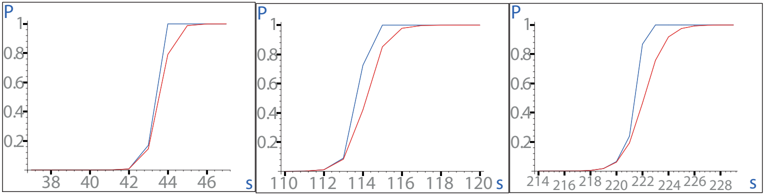
Bounds for $\mathbf{P}\{\mu_*(L) \leqslant s\}$ for $L \subset \mathbf{F}_2^N$, $N = 128$, and some $k = \dim L$.



From left to right: $k = 3N/4 = 96$, $k = N/2 = 64$, $k = N/4 = 32$.

Solutions of the equation $H\left(\frac{s}{N}, \frac{p-1}{p}\right) \approx \frac{1}{N}\ (k \ln p - \ln(4\pi k \ln p))$ are correspondingly 7.628, 17.293, 32.176

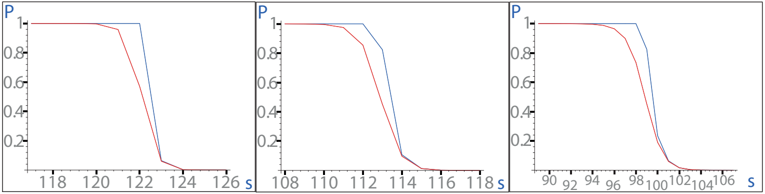
Bounds for $\mathbf{P}\{\mu_*(L) \leqslant s\}$ for $L \subset \mathbf{F}_2^N$, $N = 1024$, and some $k = \dim L$.



From left to right: $k = \frac{3N}{4} = 768$, $k = \frac{N}{2} = 512$, $k = \frac{N}{4} = 256$.

Solutions of the equation $H\left(\frac{s}{N}, \frac{p-1}{p}\right) \approx \frac{1}{N}\ (k \ln p - \ln(4\pi k \ln p))$ are correspondingly 45.533 116.727, 225.669

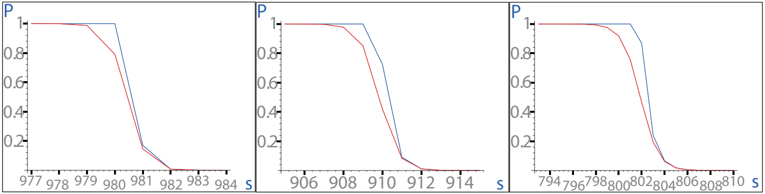
Bounds for $\mathbf{P}\{\mu^*(L) \geqslant s\}$ for $L \subset \mathbf{F}_2^N$, $N = 128$, and some $k = \dim L$.



From left to right: $k = 3N/4 = 96$, $k = N/2 = 64$, $k = N/4 = 32$.

Solutions of the equation $H\left(\frac{s}{N}, \frac{p-1}{p}\right) \approx \frac{1}{N}\,(k \ln p - \ln(4\pi k \ln p))$ are correspondingly 120.371, 110.706, 95.82

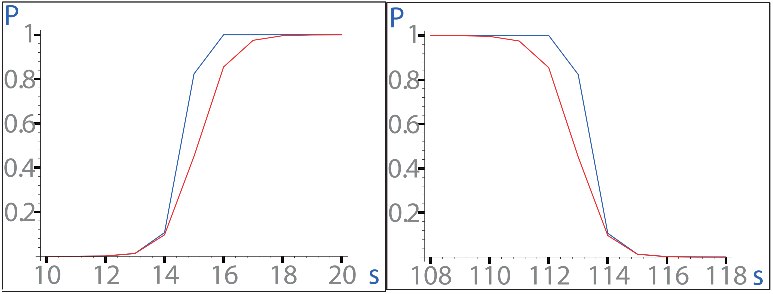
Bounds for $\mathbf{P}\{\mu^*(L) \geqslant s\}$ for $L \subset \mathbf{F}_2^N, N = 1024$, and some $k = \dim L$.



From left to right: $k = 3N/4 = 768$, $k = N/2 = 512$, $k = N/4 = 256$.

Solutions of the equation $H\left(\frac{s}{N}, \frac{p-1}{p}\right) \approx \frac{1}{N}\ (k \ln p - \ln(4\pi k \ln p))$ are correspondingly 978.466, 907.272, 798.330

One may note that graphics for $\mathbf{P}\{\mu_*(L) \leqslant s\}$ and $\mathbf{P}\{\mu^*(L) \geqslant s\}$ are visually similar, for example:



$$N = 128,\ k = 64.$$

Comparing inequalities of theorem 1 and theorem 2 we can note that

$$\mathbf{E}v_{\geqslant s}(L) = \frac{2^k - 1}{2^N - 1} \sum_{r=s}^{N} C_N^r =$$

$$= \frac{2^k - 1}{2^N - 1} \sum_{r=0}^{N-s} C_N^r = \mathbf{E}v_{\leqslant N-s}(L) + \frac{2^k - 1}{2^N - 1},$$

and thus the differences between bounds for probabilities $\mathbf{P}\{\mu_*(L) \leqslant s\}$ and $\mathbf{P}\{\mu^*(L) \geqslant N - s\}$ are very small.

Algorithms for searching codeword of minimal weight in random code:

- 1989: Stern J. A method for finding codewords of small weight.
- 1998: Canteaut A., Chabaud F. A new algorithm for finding minimum-weight words in a linear code: application to McEliece's cryptosystem and to narrow-sense BCH codes of length 511.
- 2011: May A., Meurer A., Thomae E. Decoding random linear codes in $O(2^{0.054n})$.
- 2012: Becker A., Joux A., May A., Meurer A. Decoding random binary linear codes in $2^{n/20}$: how $1+1=0$ improves information set decoding.

Overbeck, Sendrier: "most binary linear codes of length N and codimension $N-k$ have a minimum distance very close to the Gilbert-Varshamov distance d_0", where d_0 is defined as the largest integer such that

$$\sum_{i=0}^{d_0-1} C_N^i \le 2^{N-k}.$$

If $p=2$, then it follows from Theorem 1 that

$$\mathbf{E}v_{\leqslant s}(L)=\frac{2^k-1}{2^N-1}\sum\nolimits_{u=1}^{s}C_N^u,$$

$$\frac{1}{1+\frac{2^N-2^k}{2^N-1}(\mathbf{E}v_{\leqslant s}(L))^{-1}}\leqslant\mathbf{P}\{\mu_*(L)\leqslant s\}\leqslant\mathbf{E}v_{\leqslant s}(L).$$

So, typical values of $\mu_*(L)$ correspond to values of s such that $\mathbf{E}v_{\leqslant s}(L)\approx\frac{1}{2^{N-k}}\sum_{u=1}^{s}C_N^u\approx 1$ is separated from 0 and ∞.

This corresponds the *Gilbert-Varshamov distance* d_0 which is the largest integer such that $\sum_{i=0}^{d_0-1}C_N^i\leq 2^{N-k}$, but our inequalities also give estimates for fractions of codes with atypical minimal codeword weight.

According to the equation

$$H\left(\frac{s}{N}, \frac{p-1}{p}\right) \approx \frac{1}{N}\, \left(k \ln p - \ln(4\pi k \ln p)\right),$$

for random linear codes L in $\mathbf{F}_2^N$ of dimension $k = N/2$ the typical values of minimal non-zero codeword weight $\mu_*(L)$ are concentrated near the value

$$0.1100\ldots \cdot N$$

and typical values of maximal codeword weight $\mu^*(L)$ are concentrated near the value

$$0.8899\ldots \cdot N.$$

N	k	d_0	min root
64	32	7	10.043
128	64	15	17.293
256	128	29	31.634
512	256	57	60.088
768	384	85	88.431
1024	512	113	116.727
1536	768	170	173.244
2048	1024	226	229.710

Thank you!

Finding codeword Vs. Decoding.

One can decode a linear code by finding a low-weight codeword in a slightly larger code.
If L is a code over $\mathbf{F}_2$, and $y \in \mathbf{F}_2^N$ has distance w from a codeword $x \in L$, then $y - x$ is a weight-w element of the code $L + \{0, y\}$.
Conversely, if L is a code over $\mathbf{F}_2$ with minimum distance larger than w, then a weight-w element $e \in L + \{0, y\}$ cannot be in L, so it must be in $e \in L + \{y\}$ and thus $y - e$ is an element of L with distance w from y.
If $\dim L = k$ and $y \notin L$ then $\dim L + \{0, y\} = k + 1$.

McEliece cryptosystem.

- Keysetting: select $n \times n$ permutation binary matrix P, nonsingular $k \times k$ binary matrix S, select an irreducible polynomial $g \in \mathbf{F}_{2^d}[x]$ of degree t and fix generator matrix G of corresponding Goppa code of dimension $k = n - td$.
- Public key is SGP, private key is (S, G, P), values n, k, t are also public parameters.

McEliece cryptosystem.

- Encryption: for message $m \in \mathbf{F}_2^k$ select random error vector $e \in \mathbf{F}_2^N$ of weight $w(e) = t$ and compute cyphertext $c = mSGP \oplus e \in \mathbf{F}_2^N$.

- Decryption: for cyphertext $c = mSGP \oplus e$ compute $mP^{-1} = mSG + eP^{-1}$. Note that mSG is a codeword in Γ and $w(eP^{-1}) = t$, so we can recover mSG and therefore m.

- Eavesdropper faces NP-hard problem of correcting error e for seemingly random linear code with generator matrix SGP.

Goppa codes.

Fix a finite field $\mathbf{F}_{2^d}$, a basis of $\mathbf{F}_{2^d}$ over $\mathbf{F}_2$, and a set of n distinct elements $\alpha_1, \ldots, \alpha_n \in \mathbf{F}_{2^d}$. Fix an irreducible polynomial $g \in \mathbf{F}_{2^d}[x]$ of degree t, where $2 \leq t \leq (n-1)/d$.

The Goppa code $\Gamma = \Gamma(\alpha_1, \ldots, \alpha_n, g)$ consists of all elements $\mathbf{c} = (c_1, \ldots, c_n)$ in $\mathbf{F}_2^n$ satisfying

$$\sum\nolimits_{i=1}^{n} \frac{c_i}{x - \alpha_i} = 0 \quad \text{in} \quad \mathbf{F}_{2^d}[x]/g.$$

The dimension of Γ is at least $n - td$ and typically is exactly $n - td$. The minimum distance of Γ is at least $2t + 1$.

Следовало сравнить кодовое расстояние кода Гоппы с нашими оценками.