Some explanations about the histograms

Charles Meyer-Hilfiger

Inria de Paris & Sorbonne Université

1 Notation

We use the same notation as in arXiv version of the Statistical Decoding 2.0 article.
Let $\mathbf{y} = \mathbf{c} + \mathbf{e}$ be a noisy codeword at distance t from an $[n, k]$ linear code $\mathscr{C}$, namely $\mathbf{c} \in \mathscr{C}$ and $|\mathbf{e}| = t$. Let $\mathscr{P} \subset [\![1; n]\!]$ such that $\#\mathscr{P} = s$ and define $\mathscr{N} = [\![1; n]\!] \setminus \mathscr{P}$.
Let $\widetilde{\mathscr{H}}$ be a set of N parity-checks of $\mathscr{C}$ of weight w on $\mathscr{N}$ and so that their restriction to $\mathscr{P}$ leads to a set $\mathscr{H}$ of N different vectors of $\mathbb{F}_2^s$, we let for $\mathbf{a} \in \mathscr{H}$, $\widetilde{\mathbf{a}}$ be the unique parity-check in $\widetilde{\mathscr{H}}$ such that $\widetilde{\mathbf{a}}_{\mathscr{P}} = \mathbf{a}$.
We denote by $\mathscr{D}$ the code

$$\mathscr{D} \stackrel{\triangle}{=} \{\mathbf{c}_{\mathbf{x}},\ \mathbf{x} \in \mathbb{F}_2^s\} \text{ where } \mathbf{c}_{\mathbf{x}} \stackrel{\triangle}{=} (\langle \mathbf{x}, \mathbf{a} \rangle)_{\mathbf{a} \in \mathscr{H}},$$

and by $\mathbf{u}_{\mathbf{y},\mathscr{H}}$ the word we want to decode in $\mathscr{D}$, namely: $\mathbf{u}_{\mathbf{y},\mathscr{H}} = (\langle \mathbf{y}, \widetilde{\mathbf{a}} \rangle)_{\mathbf{a} \in \mathscr{H}}$.
We define

$$\begin{aligned} f_{\mathbf{y},\mathscr{H}} &: \mathbb{F}_2^s \to \mathbb{R} \\ \mathbf{a} &\mapsto \begin{cases} (-1)^{\langle \mathbf{y}, \widetilde{\mathbf{a}} \rangle} & \text{if } \mathbf{a} \in \mathscr{H} \\ 0 & \text{otherwise} \end{cases} \end{aligned} \tag{1.1}$$

and denote by

$$\widehat{f_{\mathbf{y},\mathscr{H}}}(\mathbf{x}) = \#\mathscr{H} - 2|\mathbf{u}_{\mathbf{y},\mathscr{H}} - \mathbf{c}_{\mathbf{x}}|$$

the Fourier transform of $f_{\mathbf{y},\mathscr{H}}$.
We denote by $GV(k, n)$ the smallest integer d such that $\sum_{i=0}^{d} \binom{n}{i} > 2^{n-k}$.

2 Experimental procedure

Each file **histogram_w_s_k_n_u_t.pdf** shows 4 histograms. Each histogram is made from the values of $\widehat{f_{\mathbf{y},\mathscr{H}}}$. We get the 4 samples as follows:
We choose an $[n, k]$ linear code $\mathscr{C}$ at random and choose a random codeword $\mathbf{c}$ of $\mathscr{C}$. Then we iterate 4 times the following procedure:

- $\mathscr{P} \stackrel{\$}{\leftarrow} \{\mathscr{I} \subseteq [\![1, n]\!] \ :\ \#\mathscr{I} = s\}$
- $\mathbf{e} \stackrel{\$}{\leftarrow} \{\mathbf{e} \in \mathbb{F}_2^n \ : |\mathbf{e}_{\mathscr{P}}| = t - u \text{ and } |\mathbf{e}_{\mathscr{N}}| = u\}$
- $\mathbf{y} \leftarrow \mathbf{c} + \mathbf{e}$
- Compute the set $\widetilde{\mathscr{H}}' = \{\mathbf{h} \in \mathscr{C}^{\perp} \ : |\mathbf{h}_{\mathscr{N}}| = w\}$
- Compute $\widetilde{\mathscr{H}}$ by choosing for each $\mathbf{a} \in \mathbb{F}_2^s$ a random element in $\{\mathbf{h} \in \widetilde{\mathscr{H}}' : \mathbf{h}_{\mathscr{P}} = \mathbf{a}\}$. This set can be empty in which case $\mathbf{a}$ does not appear in $\mathscr{H}$.
- $\textsc{Store}(\widehat{f_{\mathbf{y},\mathscr{H}}})$

3 Reading the files

histogram_w_s_k_n_u_t.pdf shows 4 histograms corresponding to the values of $\widehat{f_{\mathbf{y},\mathscr{H}}}$ for the 4 iterations of our experimental procedure.
histogram_w_s_k_n_u_t_zoom.pdf shows the same histograms but where we only consider the values of $\widehat{f_{\mathbf{y},\mathscr{H}}}$ greater than $0.6 * GV(s, \#\mathscr{H})$.
For the sake of readability in both files we only show on the histograms the values of $\widehat{f_{\mathbf{y},\mathscr{H}}}$ that are below $2 * GV(s, \#\mathscr{H})$.
The following values are present in these files:

- $\mathcal{F}(GV) = \#\mathcal{H} - 2 * GV(s, \#\mathcal{H})$. It is the theoretical value of the Walsh transform of a word at distance $GV(s, \#\mathcal{H})$ from the code $\mathscr{D}$.
- $\mathcal{F}(\epsilon) = \#\mathcal{H} - 2\lfloor \frac{1-\epsilon}{2} * \#\mathcal{H} \rfloor$. It is the theoretical value of the Walsh transform of a word at distance $\lfloor \frac{1-\epsilon}{2} * \#\mathcal{H} \rfloor$ from the code $\mathscr{D}$ where $\epsilon = \frac{K_w^{n-s}(u)}{\binom{n-s}{w}}$ is the bias of the LPN samples.
- $\frac{\mathcal{F}(GV)+\mathcal{F}(\epsilon)}{2}$ represents the threshold of acceptation of $\mathbf{x}_0 = \arg\max \widehat{f_{\mathbf{y},\mathscr{H}}}$ (it is basically the one we use in Algorithm 3.1).
  If $\widehat{f_{\mathbf{y},\mathscr{H}}}(\mathbf{x}_0) > \frac{\mathcal{F}(GV)+\mathcal{F}(\epsilon)}{2}$ then $\mathbf{x}_0$ is considered as a valid solution.

- $\mathcal{F}(e_P) = \widehat{f_{\mathbf{y},\mathscr{H}}}(\mathbf{e}_{\mathscr{P}})$. It is the experimental value of the Walsh transform of $\mathbf{e}_{\mathscr{P}}$.
- The second highest Walsh coefficient. It is the experimental value: $\max_{\mathbf{x} \neq \mathbf{x}_0} \widehat{f_{\mathbf{y},\mathscr{H}}}$