Post-Quantum Cryptography - Codes

Lecture 2: Random Codes

Lecturer: Thomas Debris-Alazard

Introduction

We study in this course *random codes*, *i.e.* codes whose parity-check or generator matrix is drawn uniformly at random. However, in light of the history of error correcting codes which has consisted in finding codes becoming more and more complex and structured, one may wonder but why then are we wasting our time to study random codes? That may come as a surprise but random codes enlighten about what we could expect or not in a simple fashion, and even better what is optimal or not. The most prominent example of the interest of random codes is the famous Shannon theorem about the capacity of some "noisy channels". Roughly speaking, Shannon gave (for some error models) the maximum amount of errors that "can" be theoretically decoded with codes of fixed rate. Shannon made his proof by using random codes and he has shown that they are precisely those which reach the optimality.

Our aim in these lecture notes is to study carefully these kind of codes and to show that they enable to answer many questions like for instance:

- How many vectors of Hamming weight t do we expect in a code?
- What is the typical minimum distance of a code?
- etc...

Our study will have an important consequence for cryptographic purposes: a better understanding of the Decoding Problem $\mathsf{DP}(n, q, R, \tau)$ that was defined in lecture notes 1. We will be able to predict with a very good accuracy the number of solutions of this problem as a function of its parameters. This will be particularly helpful to understand the behaviour of algorithms solving it.

1. Prerequisites

Basic notation. In all these lecture notes, q will denote a fixed field size while R will be a *constant* in $(0, 1)$. On the other hand, $\tau(n)$ will denote any function of n taking its values in $(0, 1)$. To simplify notation, since n is clear from the context, we will drop the dependency in n and simply write τ. Furthermore, parameters k and t will always (even implicitly) be defined as $k \stackrel{\text{def}}{=} \lfloor Rn \rfloor$ and $t \stackrel{\text{def}}{=} \lfloor \tau n \rfloor$. A function $f(n)$ is said to be negligible, and we denote this by $f \in \mathrm{negl}(n)$, if for all polynomial $p(n)$, $|f(n)| < |p(n)|^{-1}$ for all sufficiently large n.

Many asymptotic results will be given. As all our parameters are functions of n, our asymptotic results will always hold for:

$$n \longrightarrow +\infty.$$

The parameter n is in most cryptographic applications roughly given by several thousands.

The following function h_q, known as the *$q-$ary entropy*, will play an important role:

$$h_q : x \in [0, 1] \longmapsto -x \log_q \left(\frac{x}{q-1} \right) - (1-x) \log_q (1-x) \quad \text{(extended by continuity in 0 and 1).}$$

It is equal to the entropy of a random variable e over $\mathbb{F}_q$ distributed like the error for a $q-$ary symmetric channel of crossover probability x, *i.e.* $\mathbb{P}(e = 0) = 1 - x$ and $\mathbb{P}(e = \alpha) = \frac{x}{q-1}$ for any $\alpha \in \mathbb{F}_q^\star$.

It can be verified that h_q is an increasing function over $\left[0, \frac{q-1}{q}\right]$ and a decreasing function over $\left[\frac{q-1}{q}, 1\right]$. The $q-$ary entropy is involved in the estimation of $\sharp \mathscr{S}_t$ where $\mathscr{S}_t$ is defined as the sphere of radius t for the Hamming distance $|\cdot|$, namely

$$\mathscr{S}_t \stackrel{\text{def}}{=} \left\{ \mathbf{x} \in \mathbb{F}_q^n \ : \ |\mathbf{x}| = t \right\}.$$

The following elementary lemma will be at the core of most of our asymptotic results.

Lemma 1. *Let $t \stackrel{def}{=} \lfloor \tau n \rfloor$. We have $\sharp \mathscr{S}_t = \binom{n}{t}(q-1)^t$ and*

$$q^{n\left(h_q(\tau)+O\left(\frac{\log_q(n)}{n}\right)\right)} \leq \binom{n}{t}(q-1)^t \leq q^{nh_q(\tau)}.$$

Asymptotically,

$$\frac{1}{n}\log_q\left(\binom{n}{t}(q-1)^t\right) = h_q(\tau) + O\left(\frac{\log_q n}{n}\right).$$

Probabilistic notation. During these lecture notes we wish to emphasize on which probability space the probabilities or the expectations are taken. Therefore we will denote by a subscript the random variable specifying the associated probability space over which the probabilities or expectations are taken. For instance the probability $\mathbb{P}_X(A)$ of the event A is taken over Ω the probability space over which the random variable X is defined, *i.e.* if X is for instance a real random variable, X is a function from a probability space Ω to $\mathbb{R}$, and the aforementioned probability is taken according to the measure chosen for Ω.

Statistical distance. An essential tool for many cryptographic applications is the *statistical distance*, sometimes called the *total variational distance*. It is a distance for probability distributions, which in the case where X and Y are two random variables taking their values in a same finite space $\mathscr{E}$ is defined as

$$\Delta(X,Y) \stackrel{\text{def}}{=} \frac{1}{2}\sum_{a \in \mathscr{E}} \left|\mathbb{P}\left(X=a\right) - \mathbb{P}\left(Y=a\right)\right|. \tag{1}$$

An equivalent definition is given by

$$\Delta\left(X,Y\right) \stackrel{\text{def}}{=} \max_{A \subseteq \mathscr{E}} \left|\mathbb{P}_X(A) - \mathbb{P}_Y(A)\right|. \tag{2}$$

Depending on the context, (1) or (2) is the most useful. A direct consequence of (2) is that given any event A, we have $|\mathbb{P}_X(A) - \mathbb{P}_Y(A)| \leq \Delta(X,Y)$. Therefore, computing probabilities over X or Y will differ by at most $\Delta(X,Y)$. Furthermore, given a single observation, coming from X or Y with probability $1/2$, we will be able to guess which with probability at most $1/2 + \Delta(X,Y)/2$ and there is a strategy to reach this probability of guessing correctly.

The statistical distance enjoys many interesting properties. Among others, it cannot increase by applying a function f,

$$\Delta(f(X), f(Y)) \leq \Delta(X,Y) \quad (\textit{data processing inequality}).$$

The function f can be randomized, but its internal randomness has to be independent from X and Y for the data processing inequality to hold. In particular, it implies that the "success" probability of any algorithm $\mathscr{A}$ for inputs distributed according to X or Y, can only differ by at most $\Delta(X,Y)$.

In our applications we will focus on distributions X, Y such that their statistical distance is negligible. It will show (as a consequence of the data processing inequality) that X and Y are computationally indistinguishable[(1)] without requiring any computational argument with a reduction.

[(1)] See here for a definition: https://www.cs.princeton.edu/courses/archive/spr10/cos433/lec4.pdf

One can define various other distances for capturing in a cryptographic context the differences between two distributions. For instance, we can cite the family Renyi divergences, but this is out of the scope of these lecture notes.

2. Random Codes

The model of random codes. In these lecture notes we will use two probabilistic models that will be referred to as *random* $[n,k]_q$*-codes*. The first one is by choosing a code $\mathcal{C}$ by picking uniformly at random a generator matrix $\mathbf{G} \in \mathbb{F}_q^{k\times n}$ (*i.e.* $\mathcal{C} \stackrel{\text{def}}{=} \left\{\mathbf{mG} : \mathbf{m} \in \mathbb{F}_q^k\right\}$). However, all the probabilistic results of these lecture notes are easier to prove if, instead, we choose $\mathcal{C}$ by picking uniformly at random a parity-check matrix $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ (*i.e.* $\mathcal{C} \stackrel{\text{def}}{=} \left\{\mathbf{c} \in \mathbb{F}_q^n : \mathbf{Hc}^\intercal = \mathbf{0}\right\}$). We will denote $\mathbb{P}_{\mathbf{G}}$ and $\mathbb{P}_{\mathbf{H}}$ respectively the probabilities in these two models.

It may be pointed out that in both models we don't pick uniformly at random an $[n,k]_q$-code. Indeed, the first model always produces codes of dimension $\leq k$ whereas in the second model codes are always of dimension $\geq k$. One may wonder why don't we pick $\mathbf{G}$ (*resp.* $\mathbf{H}$) uniformly at random among the $k \times n$ (*resp.* $(n-k)\times n$) matrices of rank k (*resp.* $n-k$)? First, computations are much more complicated in this "exact" model. Furthermore, it turns out that it is pointless. Roughly speaking, the $\mathbf{G}$-model produces codes of dimension $= k$ with probability $1 - O(q^{-(n-k)})$ while in the $\mathbf{H}$-model we get a code of dimension $= k$ with probability $1 - O(q^{-k})$. As shown in the following lemma this result can even be expressed in a stronger way, our probabilistic models are exponentially close, *for the statistical distance*, to the "exact" model. Therefore all our computations in the $\mathbf{G}$ or $\mathbf{H}$ models can really be thought as by picking uniformly at random an $[n,k]_q$-code $\mathcal{C}$.

Lemma 2. *Let $\mathbf{G} \in \mathbb{F}_q^{k\times n}$ (resp. $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$) be a uniformly random matrix and $\mathbf{G}_k \in \mathbb{F}_q^{k\times n}$ (resp. $\mathbf{H}_{n-k} \in \mathbb{F}_q^{(n-k)\times n}$) be a uniformly random matrix of rank k (resp. $n-k$). We have:*

$$\Delta\left(\mathbf{G}, \mathbf{G}_k\right) = O\left(q^{-(n-k)}\right) \quad \left(\textit{resp. } \Delta\left(\mathbf{H}, \mathbf{H}_{n-k}\right) = O\left(q^{-k}\right)\right).$$

Proof. Let us prove the lemma for $(\mathbf{G}, \mathbf{G}_k)$, the other case will be similar. First it is a classical fact that the density of rank k matrices among $\mathbb{F}_q^{k\times n}$ is equal to $1 - O\left(q^{-(n-k)}\right)$. Therefore, given some rank k matrix $\mathbf{R} \in \mathbb{F}_q^{k\times n}$, we have:

$$\mathbb{P}\left(\mathbf{G}_k = \mathbf{R}\right) = \frac{1}{q^{k\times n}\left(1 - O(q^{-(n-k)})\right)}.$$

It leads to the following computation:

$$\begin{aligned}
2\Delta\left(\mathbf{G}, \mathbf{G}_k\right) &= \sum_{\substack{\mathbf{R}\in\mathbb{F}_q^{k\times n}\\ \text{rank}(\mathbf{R})=k}} \left|\mathbb{P}(\mathbf{G}=\mathbf{R}) - \mathbb{P}(\mathbf{G}_k=\mathbf{R})\right| + \sum_{\substack{\mathbf{R}\in\mathbb{F}_q^{k\times n}\\ \text{rank}(\mathbf{R})\neq k}} \mathbb{P}(\mathbf{G}=\mathbf{R}) \\
&= \sum_{\substack{\mathbf{R}\in\mathbb{F}_q^{k\times n}\\ \text{rank}(\mathbf{R})=k}} \left|\frac{1}{q^{k\times n}}\left(1 - \frac{1}{1-O\left(q^{-(n-k)}\right)}\right)\right| + \sum_{\substack{\mathbf{R}\in\mathbb{F}_q^{k\times n}\\ \text{rank}(\mathbf{R})\neq k}} \frac{1}{q^{k\times n}} \\
&= O\left(q^{-(n-k)}\right)
\end{aligned}$$

which concludes the proof. □

Now one may wonder why do we consider two models for random $[n,k]_q$-codes? It turns out that depending of the context, computations might be easier and/or more natural in one model rather than in the other one. In addition, for the same reasons as those given in the previous lemma, $\mathbf{G}$ and $\mathbf{H}$ models are closely related,

computations in both probabilistic models will outcome the same results up to an additive exponentially small factor.

Lemma 3. *Let $\mathscr{E}$ be a set of linear codes of length n in $\mathbb{F}_q$ which is defined as an event. We have,*

$$|\mathbb{P}_{\mathbf{G}}(\mathscr{E}) - \mathbb{P}_{\mathbf{H}}(\mathscr{E})| = O\left(q^{-\min(k,n-k)}\right).$$

Proof. Let $\mathbf{G}_k$ and $\mathbf{H}_{n-k}$ be defined as in Lemma 2. Notice that $\mathbb{P}_{\mathbf{G}_k}(\mathscr{E}) = \mathbb{P}_{\mathbf{H}_{n-k}}(\mathscr{E})$, in both models, we exactly pick uniformly at random an $[n,k]_q$-code. It leads to the following computation:

$$\begin{aligned}|\mathbb{P}_{\mathbf{G}}(\mathscr{E}) - \mathbb{P}_{\mathbf{H}}(\mathscr{E})| &\leq |\mathbb{P}_{\mathbf{G}}(\mathscr{E}) - \mathbb{P}_{\mathbf{G}_k}(\mathscr{E})| + \left|\mathbb{P}_{\mathbf{H}_{n-k}}(\mathscr{E}) - \mathbb{P}_{\mathbf{H}}(\mathscr{E})\right| \\ &\leq \Delta(\mathbf{G}, \mathbf{G}_k) + \Delta(\mathbf{H}, \mathbf{H}_{n-k})\end{aligned}$$

where in the last line we used Equation (2). It concludes the proof by using Lemma 2. □

Exercise 1. *Let us introduce the following variant of* DP *(which has been introduced in Lecture 1) with generator matrices instead of parity-check matrices*

$\mathsf{DP}'(n,q,R,\tau)$. *Let $k \stackrel{def}{=} \lfloor Rn \rfloor$ and $t \stackrel{def}{=} \lfloor \tau n \rfloor$.*

- Input: $(\mathbf{G}, \mathbf{y} \stackrel{def}{=} \mathbf{sG} + \mathbf{x})$ *where $\mathbf{G}, \mathbf{s}$ and $\mathbf{x}$ are uniformly distributed over $\mathbb{F}_q^{k\times n}$, $\mathbb{F}_q^k$ and words of Hamming weight t in $\mathbb{F}_q^n$.*
- Output: *an error $\mathbf{e} \in \mathbb{F}_q^n$ of Hamming weight t such that $\mathbf{y} - \mathbf{e} = \mathbf{mG}$ for some $\mathbf{m} \in \mathbb{F}_q^k$.*

Show that for any algorithm $\mathscr{A}$ solving this problem with probability ε and time T, there exists an algorithm $\mathscr{B}$ which solves $\mathsf{DP}(n,q,R,\tau)$ in time $O\left(n^3 + T\right)$ with probability $\geq \varepsilon - O\left(q^{-\min(k,n-k)}\right)$. Show that we can exchange DP' by DP *in the previous question.*

Remark 1. *The above exercise shows that defining* DP *with generator or parity-check matrices is just a matter of personal taste, it does not change the average hardness.*

A first computation with random codes. Now that random codes are well defined, we are ready to make our first computation in this probabilistic model. The following elementary lemma gives the probability (over the codes) that a fixed non-zero word $\mathbf{y}$ reaches some syndrome $\mathbf{s}$ according to the code. In particular, by setting $\mathbf{s}$ to $\mathbf{0}$, we obtain the probability that $\mathbf{y}$ belongs to the code. This lemma will be at the core of all our results about random codes.

Lemma 4. *Given $\mathbf{s} \in \mathbb{F}_q^{n-k}$ and $\mathbf{y} \in \mathbb{F}_q^n$ such that $\mathbf{y} \neq \mathbf{0}$, we have for $\mathbf{H}$ being uniformly distributed at random in $\mathbb{F}_q^{(n-k)\times n}$,*

$$\mathbb{P}_{\mathbf{H}}(\mathbf{yH}^\intercal = \mathbf{s}) = \frac{1}{q^{n-k}}.$$

Proof. Let $h_{i,j}$ be the coefficient of $\mathbf{H}$ at position (i,j). Without loss of generality, we can suppose that $y_1 = 1$ (by permuting $\mathbf{y}$ if $y_1 = 0$ and then multiplying by y_1^{-1} which is possible as we work in $\mathbb{F}_q$). The probability that we are looking for is the probability of the following event:

$$\forall i \in [\![1, n-k]\!], \quad h_{i,1} = s_i - \sum_{j=2}^{n} h_{i,j} y_j$$

Recall that $\mathbf{H}$ is uniformly distributed: the $h_{i,j}$'s are independent and equidistributed. Therefore the above $n-k$ equations will be independently true with probability $1/q$ which concludes the proof. □

Exercise 2. *Show that for any non-zero $\mathbf{y} \in \mathbb{F}_q^n$,*

$$\mathbb{P}_{\mathbf{G}}(\mathbf{y} \in \mathcal{C}^*) = \frac{1}{q^k}.$$

3. Weight Distribution of Cosets of Random Codes

The aim of this section is to answer the following question: given a random code $\mathcal{C}$ and a fixed vector $\mathbf{y} \in \mathbb{F}_q^n$, how many codewords $\mathbf{c} \in \mathcal{C}$ do we expect to be at Hamming distance t from $\mathbf{y}$? Or equivalently, given a parity-check matrix of our random code $\mathcal{C}$ and a fixed syndrome $\mathbf{s}$, how many vectors $\mathbf{e}$ of Hamming weight t do we expect to reach the syndrome $\mathbf{s}$ according to $\mathbf{H}$? Notice that deriving an answer to these questions in the particular cases $\mathbf{y} = \mathbf{0}$ and $\mathbf{s} = \mathbf{0}$ enables to compute the expected number of codewords of weight t in $\mathcal{C}$. It will be useful to compute the expected minimum distance of a code.

These results will have an important consequence: a better understanding of the Decoding Problem (DP) that we recall now.

Problem 1 (Decoding Problem - $\mathsf{DP}(n, q, R, \tau)$)**.** *Let $k \stackrel{def}{=} \lfloor Rn \rfloor$ and $t \stackrel{def}{=} \lfloor \tau n \rfloor$.*

- Input : $(\mathbf{H}, \mathbf{s} \stackrel{def}{=} \mathbf{x}\mathbf{H}^\intercal)$ *where* $\mathbf{H}$ *(resp.* $\mathbf{x}$*) is uniformly distributed over* $\mathbb{F}_q^{(n-k)\times n}$ *(resp.* $\mathscr{S}_t$*, words of Hamming weight* t *in* $\mathbb{F}_q^n$*).*
- Output : *an error* $\mathbf{e} \in \mathbb{F}_q^n$ *of Hamming weight* t *such that* $\mathbf{e}\mathbf{H}^\intercal = \mathbf{s}$.

According to our probabilistic model, this problem really corresponds to decode a random $[n, k]_q$-code of parity-check matrix $\mathbf{H}$. In that case it is natural to wonder how many vectors $\mathbf{e} \in \mathscr{S}_t$ are expected to reach the syndrome $\mathbf{s}$ according to $\mathbf{H}$, but why? To understand this let us take a toy example. A trivial solution to solve DP is to pick a random error $\mathbf{e} \in \mathscr{S}_t$ with the hope that it gives a solution. By definition there is a solution to our problem (here $\mathbf{x}$). If there is exactly one solution, our success probability is given by $\frac{1}{\binom{n}{t}(q-1)^t}$. But now imagine that we expect N solutions to our problem. In that case we would expect our success probability to be equal to $\approx \frac{N}{\binom{n}{t}(q-1)^t}$. It is therefore important to know the value of N to be able to predict the running time of our algorithm. It is the aim of what follows.

Notation. Given $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ and $\mathbf{s} \in \mathbb{F}_q^{n-k}$, let

$$N_t(\mathcal{C}, \mathbf{s}) \stackrel{\text{def}}{=} \sharp \left\{ \mathbf{e} \in \mathscr{S}_t : \mathbf{e}\mathbf{H}^\intercal = \mathbf{s} \right\}$$

where implicitly $\mathcal{C}$ is defined as $\{\mathbf{c} \in \mathbb{F}_q^n : \mathbf{H}\mathbf{c}^\intercal = \mathbf{0}\}$. Notice that $N_t(\mathcal{C}, \mathbf{s})$ is a random variable that gives the number of solutions of DP with input $(\mathbf{H}, \mathbf{s})$ (where $\mathbf{s} \in \mathbb{F}_q^{n-k}$ is fixed and not necessarily computed as some $\mathbf{x}\mathbf{H}^\intercal$ for $\mathbf{x} \in \mathscr{S}_t$). On the other hand, $N_t(\mathcal{C}, \mathbf{0})$ is a random variable that gives the number of codewords $\mathbf{c} \in \mathcal{C}$ of Hamming weight t.

Expected weight distribution of cosets. From now on, our main objective is to compute the expected number of Hamming weight t vectors in a given coset, namely to compute the expectation of $N_t(\mathcal{C}, \mathbf{s})$ over $\mathcal{C}$. To avoid any suspense, we will prove that for any syndrome $\mathbf{s} \in \mathbb{F}_q^{n-k}$, the expectation of $N_t(\mathcal{C}, \mathbf{s})$ is given by $\binom{n}{t}(q-1)^t/q^{n-k}$. However, before showing this result, let us start to understand how this quantity behaves as function of it parameters, namely $\tau = t/n$ and $R = k/n$. By Lemma 1, we have

$$\frac{1}{n}\log_q \binom{n}{t}(q-1)^t/q^{n-k} = h_q(\tau) - (1-R) + O\left(\frac{\log_q n}{n}\right). \tag{3}$$

Recall now that $x \in [0,1] \mapsto h_q(x)$ is an increasing function over $\left[0, \frac{q-1}{q}\right]$ and a decreasing function over $\left[\frac{q-1}{q}, 1\right]$. Furthermore, $h_q(0) = 0$ and $h_q(1) = \log_q(q-1)$. This shows that $N_t(\mathcal{C}, \mathbf{s})$ is expected (according to τ) to be exponentially small or large (in n) at the exception of one value τ^- and potentially a second one in the case where $(1-R) \geq \log_q(q-1)$, that we will denote τ^+. We summarize the picture by drawing in Figure 1 the logarithm in basis 3 (for n large enough) of $\binom{n}{t}(q-1)^t/q^{n-k}$ when $q = 3$ and $k/n = 1/4$.

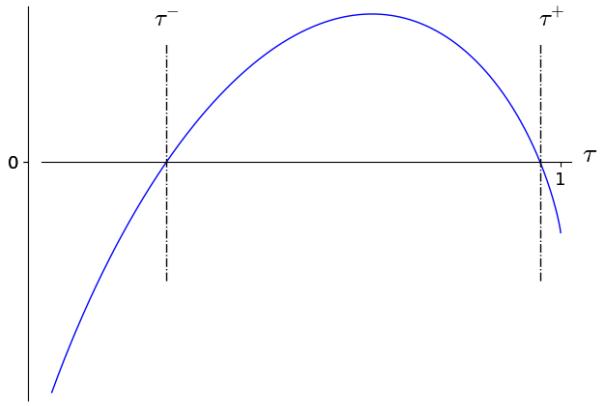

FIGURE 1. $\lim_{n\to+\infty} \frac{1}{n} \log_q \binom{n}{t}(q-1)^t/q^{n-k}$ when $q=3$ and $k/n=1/4$ as function of $\tau = t/n$.

It turns out that an analytic expression of τ^- and τ^+ can be given,

$$\tau^- \stackrel{\text{def}}{=} g_q^-(1-R) \quad \text{and} \quad \tau^+ \stackrel{\text{def}}{=} g_q^+(1-R) \text{ when } R \leq 1-\log_q(q-1) \tag{4}$$

where g_q^- (*resp.* g_q^+) denotes the inverse of h_q over $\left[0, \frac{q-1}{q}\right]$ $\left(\text{resp. } \left[\frac{q-1}{q}, 1\right]\right)$.

Remark 2. *As we will see in Section 4, τ^- is commonly called the relative Gilbert-Varshamov distance or bound.*

Quantities τ^- and τ^+ give the boundaries between which we expect $N_t(\mathcal{C}, \mathbf{s})$ to be exponentially large as we show now.

Proposition 1. *Let $k \stackrel{def}{=} \lfloor Rn \rfloor$, $t \stackrel{def}{=} \lfloor \tau n \rfloor$ and $\mathbf{s} \in \mathbb{F}_q^{n-k}$. We have:*

$$\mathbb{E}_{\mathbf{H}}(N_t(\mathcal{C}, \mathbf{s})) = \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}. \tag{5}$$

When $\tau \in \begin{cases} (\tau^-, \tau^+) & \text{if } R \leq 1-\log_q(q-1) \\ (\tau^-, 1) & \text{otherwise} \end{cases}$, *we expect $N_t(\mathcal{C}, \mathbf{s})$ to be exponentially large:*

$$\mathbb{E}_{\mathbf{H}}(N_t(\mathcal{C}, \mathbf{s})) = q^{\alpha n(1+o(1))} \quad \text{where} \quad \alpha \stackrel{def}{=} h_q(\tau) - 1 + R > 0.$$

In the case where $\tau = \begin{cases} \tau^- \text{ or } \tau^+ & \text{if } R \leq 1-\log_q(q-1) \\ \tau^- & \text{otherwise.} \end{cases}$, *the expectation of $N_t(\mathcal{C}, \mathbf{s})$ equals $P(n)(1+o(1))$ for some polynomial P.*

Proof. Let $\mathbb{1}_{\mathbf{e}}$ be the indicator function of the event "$\mathbf{e}\mathbf{H}^\intercal = \mathbf{s}$". It is readily verified that by definition,

$$N_t(\mathcal{C}, \mathbf{s}) = \sum_{\mathbf{e} \in \mathscr{S}_t} \mathbb{1}_{\mathbf{e}}.$$

We have the following computation,

$$\begin{aligned}\mathbb{E}_{\mathbf{H}}(N_t(\mathcal{C},\mathbf{s})) &= \mathbb{E}_{\mathbf{H}}\left(\sum_{\mathbf{e}\in\mathscr{S}_t}\mathbb{1}_{\mathbf{e}}\right)\\ &= \sum_{\mathbf{e}\in\mathscr{S}_t}\mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{e}}) \quad \text{(by linearity of the expectation)}\\ &= \sum_{\mathbf{e}\in\mathscr{S}_t}\mathbb{P}_{\mathbf{H}}\left(\mathbf{e}\mathbf{H}^{\intercal}=\mathbf{s}\right)\end{aligned}$$

which gives (5) by using Lemma 4. The second part of the proposition is a consequence of Lemma 1. □

Exercise 3. *Show that the average number of solutions of* $\mathsf{DP}(n,q,R,\tau)$ *(over the input distribution) is given by*

$$1+\frac{\binom{n}{t}(q-1)^t-1}{q^{n-k}}$$

Remark 3. *Proposition 1 can actually be stated more generally. Given some set* $\mathscr{E}\subseteq\mathbb{F}_q^n$*, we can show that the expected number of vectors* $\mathbf{e}\in\mathscr{E}$ *that reach some syndrome for a random* $[n,k]_q$*-code is given by* $\sharp\mathscr{E}/q^{n-k}$*.*

Remark 4. *The expected number of codewords of weight* t *in a random* $[n,k]_q$*-code is given by* $\binom{n}{t}(q-1)^t/q^{n-k}$ *(by setting* $\mathbf{s}$ *to* $\mathbf{0}$ *in Proposition 1). This statement, in the same manner as in the previous remark, can be generalized to give the expected number of codewords in any set* $\mathscr{E}\subseteq\mathbb{F}_q^n$*. In particular, it can be used to obtain the expected number of codewords of "weight"* t *that belong to a random code, for any notion of weight and therefore any metric.*

Exercise 4. *Show that* $(t>0)$*,*

$$\mathbb{E}_{\mathbf{G}}\left(\sharp\left\{\mathbf{m}\in\mathbb{F}_q^k:|\mathbf{m}\mathbf{G}|=t\right\}\right)=\frac{q^k-1}{q^n}\binom{n}{t}(q-1)^t \quad \textit{and} \quad \mathbb{E}_{\mathbf{H}}\left(\sharp\left\{\mathbf{c}\in\mathcal{C}:|\mathbf{c}| \textit{ is odd}\right\}\right)=\frac{1}{2}\,\frac{q^n-(2-q)^n}{q^{n-k}}.$$

Hint: For the first part of the exercise first show that $\mathbf{m}\mathbf{G}$ is uniformly distributed over $\mathbb{F}_q^n$ when $\mathbf{m}\in\mathbb{F}_q^k\backslash\{\mathbf{0}\}$.

At this point our work has given the *expected* number of solutions of $\mathsf{DP}(n,q,R,\tau)$. Situation is depicted in Figure 2.

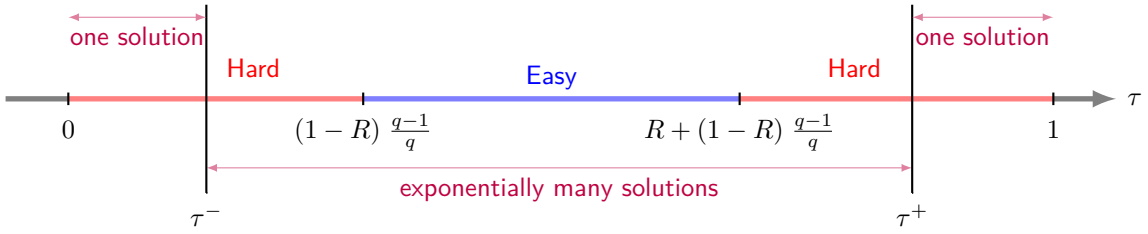

FIGURE 2. Hardness and expected number of solutions of $\mathsf{DP}(n,q,R,\tau)$ as function of τ.

However, can we be much more precise? For instance, can we give with an overwhelming probability, and therefore for almost all codes, the number of solutions of DP? The answer is yes. Below is given two techniques for achieving this, the first one uses Markov's inequality (*first moment technique*) and the second one, which is more accurate, uses Bienaymé-Tchebychev's inequality (*second moment technique*).

Proposition 2 (First Moment Technique)**.** *Let* $\mathbf{s} \in \mathbb{F}_q^{n-k}$*. For any* $a > 0$*, we have*

$$\mathbb{P}_{\mathbf{H}}(N_t(\mathcal{C}, \mathbf{s}) > a) \leq \frac{1}{a} \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}.$$

Proof. By Proposition 1, $\mathbb{E}_{\mathbf{H}}(N_t(\mathcal{C}, \mathbf{s})) = \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}$. It remains to apply Markov's inequality to conclude the proof. □

Notice that Proposition 2 is not very accurate. One has to choose $a \gg \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}$ to obtain a negligible probability that $N_t(\mathcal{C}, \mathbf{s})$ is larger than a. But the expectation of $N_t(\mathcal{C}, \mathbf{s})$ is exactly given by $\frac{\binom{n}{t}(q-1)^t}{q^{n-k}}$. A meaningful result would be

$$\mathbb{P}_{\mathbf{H}}\left(\left|N_t(\mathcal{C}, \mathbf{s}) - \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}\right| > a\right) < \varepsilon,$$

for some $\varepsilon \in \mathrm{negl}(n)$ and a be such that $a \in \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}\mathrm{negl}(n)$. It is precisely the aim of the following proposition.

Proposition 3 (Second Moment Technique)**.** *Let* $\mathbf{s} \in \mathbb{F}_q^{n-k}$*. For any* $a > 0$*, we have*

$$\mathbb{P}_{\mathbf{H}}\left(\left|N_t(\mathcal{C}, \mathbf{s}) - \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}\right| \geq a\right) \leq \frac{(q-1)\binom{n}{t}(q-1)^t}{a^2 q^{n-k}}.$$

Proof. Let $\mathbb{1}_{\mathbf{e}}$ be the indicator function of the event "$\mathbf{eH}^\intercal = \mathbf{s}$". By using Bienaymé-Tchebychevs inequality with the random variable $N_t(\mathcal{C}, \mathbf{s}) = \sum_{\mathbf{e} \in \mathscr{S}_t} \mathbb{1}_{\mathbf{e}}$, we obtain

$$\begin{aligned}
\mathbb{P}_{\mathbf{H}}\left(\left|N_t(\mathcal{C}, \mathbf{s}) - \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}\right| \geq a\right) &\leq \frac{\mathbf{Var}_{\mathbf{H}}(N_t(\mathcal{C}, \mathbf{s}))}{a^2} \\
&= \frac{1}{a^2}\left(\sum_{\mathbf{e} \in \mathscr{S}_t} \mathbf{Var}_{\mathbf{H}}(\mathbb{1}_{\mathbf{e}}) + \sum_{\substack{\mathbf{x},\mathbf{y} \in \mathscr{S}_t \\ \mathbf{x} \neq \mathbf{y}}} \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}}\mathbb{1}_{\mathbf{y}}) - \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}})\mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{y}})\right) \\
&\leq \frac{1}{a^2}\left(\sum_{\mathbf{e} \in \mathscr{S}_t} \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{e}}) + \sum_{\substack{\mathbf{x},\mathbf{y} \in \mathscr{S}_t \\ \mathbf{x} \neq \mathbf{y}}} \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}}\mathbb{1}_{\mathbf{y}}) - \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}})\mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{y}})\right) \\
&= \frac{1}{a^2}\left(\frac{\binom{n}{t}(q-1)^t}{q^{n-k}} + \sum_{\substack{\mathbf{x},\mathbf{y} \in \mathscr{S}_t \\ \mathbf{x} \neq \mathbf{y}}} \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}}\mathbb{1}_{\mathbf{y}}) - \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}})\mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{y}})\right) \qquad (6)
\end{aligned}$$

where we used that $\mathbf{Var}_{\mathbf{H}}(\mathbb{1}_{\mathbf{e}}) \leq \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{e}}^2) = \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{e}})$. Let us now upper-bound the second term of the inequality. To this aim let us prove the following lemma.

Lemma 5. *We have*

$$\mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}}\mathbb{1}_{\mathbf{y}}) \leq \begin{cases} 1/q^{n-k} & \text{if } \mathbf{x} \text{ and } \mathbf{y} \text{ are colinear} \\ 1/q^{2(n-k)} & \text{otherwise.} \end{cases}$$

Proof. The result is clear when $\mathbf{x}$ and $\mathbf{y}$ are colinear by Lemma 4. Let us suppose that $\mathbf{x}$ and $\mathbf{y}$ are not colinear and define

$$\varphi : \mathbf{h} \in \mathbb{F}_q^n \longmapsto (\mathbf{h} \cdot \mathbf{x}, \mathbf{h} \cdot \mathbf{y}).$$

It is readily verified that this linear application has a kernel of $\mathbb{F}_q$-dimension $n-2$. Therefore, for any $a, b \in \mathbb{F}_q$ and $\mathbf{h} \in \mathbb{F}_q^n$ being uniformly distributed,

$$\mathbb{P}_{\mathbf{h}}\left(\varphi(\mathbf{h}) = (a,b)\right) = \frac{q^{n-2}}{q^n} = \frac{1}{q^2} \tag{7}$$

Let us remark now that

$$\begin{aligned}\mathbb{E}_{\mathbf{H}}\left(\mathbb{1}_{\mathbf{x}}\mathbb{1}_{\mathbf{y}}\right) &= \mathbb{P}_{\mathbf{H}}\left(\mathbf{x}\mathbf{H}^{\intercal} = \mathbf{s} \ \textit{and} \ \mathbf{y}\mathbf{H}^{\intercal} = \mathbf{s}\right)\\ &= \mathbb{P}_{\mathbf{h}}\left(\varphi(\mathbf{h}) = (a,b)\right)^{n-k} \end{aligned} \tag{8}$$

where in the last line we used that each rows of $\mathbf{H}$ are independent and uniformly distributed. To conclude the proof it remains to plug Equation (7) in Equation (8). □

Lemma 5 enables us to deduce that

$$\begin{aligned}\sum_{\substack{\mathbf{x},\mathbf{y}\in\mathscr{S}_t\\ \mathbf{x}\neq\mathbf{y}}} \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}}\mathbb{1}_{\mathbf{y}}) - \mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{x}})\mathbb{E}_{\mathbf{H}}(\mathbb{1}_{\mathbf{y}}) &\leq \sum_{\mathbf{x}\in\mathscr{S}_t} \sum_{\substack{\mathbf{y}\in\mathscr{S}_t\setminus\mathbf{x}:\\ \text{colinear to } \mathbf{x}}} \frac{1}{q^{n-k}} - \frac{1}{q^{2(n-k)}}\\ &\leq \sum_{\mathbf{x}\in\mathscr{S}_t} \sum_{\substack{\mathbf{y}\in\mathscr{S}_t\setminus\mathbf{x}:\\ \text{colinear to } \mathbf{x}}} \frac{1}{q^{n-k}}\\ &\leq \frac{(q-2)\binom{n}{t}(q-1)^t}{q^{n-k}} \end{aligned} \tag{9}$$

It gives by plugging (9) in (6)

$$\begin{aligned}\mathbb{P}_{\mathbf{H}}\left(\left|N_t(\mathcal{C},\mathbf{s}) - \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}\right| \geq a\right) &\leq \frac{1}{a^2}\left(\frac{\binom{n}{t}(q-1)^t}{q^{n-k}} + \frac{(q-2)\binom{n}{t}(q-1)^t}{q^{n-k}}\right)\\ &= \frac{(q-1)\binom{n}{t}(q-1)^t}{a^2 q^{n-k}}\end{aligned}$$

which concludes the proof. □

The point of this proposition is that, for relative weights $t/n \in (\tau^-, \tau^+)$ the term $\binom{n}{t}(q-1)^t/q^{n-k}$, is exponentially large. Therefore, by carefully setting a in this case we deduce $N_t(\mathcal{C},\mathbf{s})$ with a very good precision. For instance, if $t/n \in (\tau^-, \tau^+)$, let $a = \left(\frac{\binom{n}{t}(q-1)^t}{q^{n-k}}\right)^{3/4}$ to obtain:

$$\mathbb{P}_{\mathbf{H}}\left(\left|N_t(\mathcal{C},\mathbf{s}) - \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}\right| \geq \left(\frac{\binom{n}{t}(q-1)^t}{q^{n-k}}\right)^{3/4}\right) \leq (q-1)\sqrt{\frac{q^{n-k}}{\binom{n}{t}(q-1)^t}} \in \mathrm{negl}(n).$$

The number of errors of weight t that reach some syndrome is with an overwhelming probability equal to its expectation $\binom{n}{t}(q-1)^t/q^{n-k}$ up to an additive factor $\left(\binom{n}{t}(q-1)^t/q^{n-k}\right)^{3/4}$ which is exponentially small with respect to $\binom{n}{t}(q-1)^t/q^{n-k}$.

4. Expected Minimum Distance of Codes

We are now interested in computing the expected minimum distance of a random code. As we will see it is given by the so-called *Gilbert-Varshamov* distance. This result is for cryptographic purposes very important. Suppose that one finds in a code a word of Hamming weight much smaller than it is expected. This would mean that the code is peculiar and maybe even worse, this codeword of small weight may reveal some secret information.

In view of the foregoing, it may be tempting to say that the expected minimum distance of a random $[n,k]_q$-code is given by the largest t such that $\sum_{\ell\leq t}\frac{\binom{n}{\ell}(q-1)^\ell}{q^{n-k}}\leq 1$ and that is indeed what happens. It turns out that this value of t plays an important role in coding theory and is known as the Gilbert-Varshamov distance.

Definition 1 (Gilbert-Varshamov Distance)**.** *Let $k\leq n$ and q be integers. The Gilbert-Varshamov distance $t_{\mathrm{GV}}(q,n,k)$ is defined as the largest integer such that*

$$\sum_{\ell=0}^{t_{\mathrm{GV}}(q,n,k)}\binom{n}{\ell}(q-1)^\ell\leq q^{n-k}.$$

Remark 5. *The Gilbert-Varshamov distance gives the maximum r such that the volume of the ball of radius r is smaller than the inverse density of any $[n,k]_q$-code. Its analogue for lattices (in the context of lattice-based cryptography) is called the Gaussian heuristic.*

It can be verified (by using Lemma 1)

$$\frac{t_{\mathrm{GV}}(q,n,Rn)}{n}=\tau^-+o(1)$$

where τ^- is defined in Equation (4). It explains why τ^- is commonly called the *relative Gilbert-Varshamov distance.* In the following proposition we show that the minimum distance of almost all $[n,k]_q$-codes is given by τ^-. Interestingly, the proof that $d_{\min}(\mathcal{C})/n>\tau^-$ happens with a negligible probability relies on Proposition 3 that used the second moment technique.

Proposition 4. *Let $\varepsilon>0$. We have*

$$\mathbb{P}_{\mathbf{H}}\left((1-\varepsilon)\tau^-<\frac{d_{\min}(\mathcal{C})}{n}<(1+\varepsilon)\tau^-\right)\geq 1-q^{-\alpha n(1+o(1))}$$

where $\alpha\stackrel{def}{=}\min\left((1-R)-h_q\left((1+\varepsilon)\tau^-\right),h_q\left((1-\varepsilon)\tau^-\right)-(1-R)\right)>0$.

Proof. First notice that,

$$\mathbb{P}_{\mathbf{H}}\left(\frac{d_{\min}(\mathcal{C})}{n}\notin\left((1-\varepsilon)\tau^-,(1+\varepsilon)\tau^-\right)\right)\leq\mathbb{P}_{\mathbf{H}}\left(\frac{d_{\min}(\mathcal{C})}{n}\leq(1-\varepsilon)\tau^-\right)+\mathbb{P}_{\mathbf{H}}\left(\frac{d_{\min}(\mathcal{C})}{n}\geq(1+\varepsilon)\tau^-\right).\tag{10}$$

Let us upper-bound independently both terms of the above inequation. First,

$$\begin{aligned}\mathbb{P}_{\mathbf{H}}\left(\frac{d_{\min}(\mathcal{C})}{n}\leq(1-\varepsilon)\tau^-\right)&=\mathbb{P}_{\mathbf{H}}\left(\exists\mathbf{x}\ :\ \frac{|\mathbf{x}|}{n}\leq(1-\varepsilon)\tau^-\text{ and }\mathbf{x}\in\mathcal{C}\right)\\&\leq\sum_{\substack{\mathbf{x}:\\ \frac{|\mathbf{x}|}{n}\leq(1-\varepsilon)\tau^-}}\mathbb{P}_{\mathbf{H}}\left(\mathbf{x}\in\mathcal{C}\right)\\&=\sum_{\ell=0}^{(1-\varepsilon)n\tau^-}\frac{\binom{n}{\ell}(q-1)^\ell}{q^{n-k}}\quad\text{(By Lemma 4.)}\\&=\frac{q^{n\left(h_q\left((1-\varepsilon)\tau^-\right)+o(1)\right)}}{q^{n-k}}\qquad(11)\\&=q^{n\left(h_q\left((1-\varepsilon)\tau^-\right)-(1-R)+o(1)\right)}\qquad(12)\end{aligned}$$

where we used in (11) Lemma 1 and the fact that $x\mapsto h_q(x)$ is an increasing function for $x\in[0,\tau^-]\subseteq\left[0,\frac{q-1}{q}\right]$.

Let us now upper-bound the second term of (10). Let $\delta \in (0,\varepsilon)$ and $u \stackrel{\text{def}}{=} (1+\delta)n\tau^-$. Notice now that,

$$\begin{aligned}\mathbb{P}_{\mathbf{H}}\left(\frac{d_{\min}(\mathcal{C})}{n} \geq (1+\varepsilon)\tau^-\right) &\leq \mathbb{P}_{\mathbf{H}}\left(N_u(\mathcal{C},\mathbf{0}) = 0\right) \\ &\leq \mathbb{P}_{\mathbf{H}}\left(\left|N_u(\mathcal{C},\mathbf{0}) - \frac{\binom{n}{u}(q-1)^u}{q^{n-k}}\right| \geq \frac{\binom{n}{u}(q-1)^u}{q^{n-k}}\right) \\ &\leq (q-1)\frac{q^{n-k}}{\binom{n}{u}(q-1)^u} \qquad (13)\end{aligned}$$

where in the last line we used Proposition 3 by setting $a = \frac{\binom{n}{u}(q-1)^u}{q^{n-k}}$. But now by using Lemma 1 and that $u = (1+\delta)\tau^-$ we obtain:

$$\frac{q^{n-k}}{\binom{n}{u}(q-1)^u} = q^{n\left(1-R-h_q\left((1+\delta)\tau^-\right)+o(1)\right)}.$$

By plugging this in Equation (13) we obtain for any $\delta \in (0,\varepsilon)$:

$$\mathbb{P}_{\mathbf{H}}\left(\frac{d_{\min}(\mathcal{C})}{n} \geq (1+\varepsilon)\tau^-\right) \leq q^{n\left(1-R-h_q\left((1+\delta)\tau^-\right)+o(1)\right)}.$$

Therefore, by letting $\delta \to \varepsilon$ we get:

$$\mathbb{P}_{\mathbf{H}}\left(\frac{d_{\min}(\mathcal{C})}{n} \geq (1+\varepsilon)\tau^-\right) \leq q^{n\left(1-R-h_q\left((1+\varepsilon)\tau^-\right)+o(1)\right)}. \qquad (14)$$

To conclude the proof it remains to put together Equations (12) and (14) in Equation (10). □

Remark 6. *In coding theory, the relative Gilbert-Varshamov distance τ^- is known as a "lower-bound", there exists a family of $[n, Rn]_q$-codes with a relative minimum distance $\geq \tau^-$. What we have actually proven is that almost all families of $[n, Rn]_q$-codes have asymptotically a relative minimum distance $= \tau^-$. Let us stress that it does not show that the relative Gilbert-Varshamov distance τ^- is an "upper-bound", all families of $[n, Rn]_q$-codes are such that their asymptotically relative minimum distance is $\leq \tau^-$. This is a widely open conjecture in the case of $\mathbb{F}_2$ but which is not true for $\mathbb{F}_q$ as long as q is a square ≥ 49.*

About the optimality of random codes. We have seen in lecture notes 1 that balls centred at codewords $\mathbf{c} \in \mathcal{C}$ and whose radius is $< \frac{d_{\min}(\mathcal{C})}{2}$ never overlap. This condition over the radius ensures that when an error $\mathbf{e}$ of Hamming weight smaller than $< \frac{d_{\min}(\mathcal{C})}{2}$ occurs, we are sure that computing the closest codeword from $\mathbf{c}+\mathbf{e}$ will outcome $\mathbf{c}$ (we say that the *maximum likelihood decoding* succeeds). However, for random codes this property can be made even stronger. In that case we can show that $\mathbf{c}$ is indeed the closest codeword from $\mathbf{c}+\mathbf{e}$ (with overwhelming probability) if $\mathbf{e} \approx d_{\min}(\mathcal{C})$ where $d_{\min}(\mathcal{C}) \approx \tau^- n$ as we show now in the following proposition (by using Markov's inequality).

Proposition 5. *Let $t \stackrel{def}{=} n(1-\varepsilon)\tau^-$ for some $\varepsilon > 0$ and let $\eta > 0$. We have*

$$\mathbb{P}_{\mathbf{H}}\left(\frac{\sharp\left\{\mathbf{c},\mathbf{c}' \in \mathcal{C},\ \mathbf{e},\mathbf{e}' \in \mathscr{S}_t\ :\ \mathbf{c}+\mathbf{e} = \mathbf{c}'+\mathbf{e}'\right\}}{q^k\binom{n}{t}(q-1)^t} \geq 1+\eta\right) \leq \frac{1}{\eta}\,\frac{\binom{n}{t}(q-1)^t}{q^{n-k}} = \frac{1}{\eta}\,q^{-\alpha n(1+o(1))}$$

where $\alpha \stackrel{def}{=} h_q\left((1-\varepsilon)\tau^-\right) - 1 + R > 0$.

Notice that,

$$\frac{\sharp\left\{\mathbf{c},\mathbf{c}' \in \mathcal{C},\ \mathbf{e},\mathbf{e}' \in \mathscr{S}_t\ :\ \mathbf{c}+\mathbf{e} = \mathbf{c}'+\mathbf{e}'\right\}}{q^k\binom{n}{t}(q-1)^t} = 1$$

means that there are no collisions between noisy codewords with errors of Hamming weight t (and that balls of radius t and centered at codewords do not overlap). Therefore, the above proposition shows that

for random codes, the maximum likelihood decoding will succeed for almost all noisy codewords, up to the Gilbert-Varshamov distance if η is chosen sufficiently small. However, once again, we could be more accurate by using the second moment technique with Bienaymé-Tchebychev's inequality.

Proof. Let,

$$Z \stackrel{\text{def}}{=} \sharp\left\{\mathbf{c}, \mathbf{c}' \in \mathcal{C},\ \mathbf{e}, \mathbf{e}' \in \mathscr{S}_t : \mathbf{c}+\mathbf{e} = \mathbf{c}'+\mathbf{e}'\right\}.$$

We have the following computation

$$\begin{aligned} Z &= \sum_{\mathbf{c}\in\mathcal{C}, \mathbf{e}\in\mathscr{S}_t} 1 + \sum_{\substack{\mathbf{c},\mathbf{c}'\in\mathcal{C}, \mathbf{e},\mathbf{e}'\in\mathscr{S}_t \\ (\mathbf{c},\mathbf{e})\neq(\mathbf{c}',\mathbf{e}') \\ \mathbf{c}+\mathbf{e}=\mathbf{c}'+\mathbf{e}'}} 1 \\ &= q^k \binom{n}{t}(q-1)^t + q^k \sum_{\substack{\mathbf{e},\mathbf{e}'\in\mathscr{S}_t \\ \mathbf{e}\neq\mathbf{e}':(\mathbf{e}-\mathbf{e}')\mathbf{H}^\intercal=\mathbf{0}}} 1 \\ &= q^k \binom{n}{t}(q-1)^t \left(1 + \frac{1}{\binom{n}{t}(q-1)^t} \sum_{\substack{\mathbf{e},\mathbf{e}'\in\mathscr{S}_t \\ \mathbf{e}\neq\mathbf{e}':(\mathbf{e}-\mathbf{e}')\mathbf{H}^\intercal=\mathbf{0}}} 1\right). \end{aligned}$$

Let,

$$X \stackrel{\text{def}}{=} \frac{1}{\binom{n}{t}(q-1)^t} \sum_{\substack{\mathbf{e},\mathbf{e}'\in\mathscr{S}_t \\ \mathbf{e}\neq\mathbf{e}':(\mathbf{e}-\mathbf{e}')\mathbf{H}^\intercal=\mathbf{0}}} 1$$

By using Lemma 4,

$$\begin{aligned} \mathbb{E}_{\mathbf{H}}(X) &= \frac{1}{\binom{n}{t}(q-1)^t} \sum_{\substack{\mathbf{e},\mathbf{e}'\in\mathscr{S}_t \\ \mathbf{e}\neq\mathbf{e}'}} \frac{1}{q^{n-k}} \\ &\leq \frac{1}{\binom{n}{t}(q-1)^t} \frac{\left(\binom{n}{t}(q-1)^t\right)^2}{q^{n-k}} \\ &= \frac{\binom{n}{t}(q-1)^t}{q^{n-k}}. \end{aligned}$$

To conclude the proof it is enough to apply Markov's inequality to upper-bound $\mathbb{P}_{\mathbf{H}}(X > \eta)$. □

5. Uniform Distribution of Syndromes

In lecture notes 1 we have seen that the decoding problem is defined (for cryptographic purposes) with some distribution in input, namely $(\mathbf{H}, \mathbf{x}\mathbf{H}^\intercal)$ where $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ and $\mathbf{x} \in \mathscr{S}_t$ are uniformly distributed. Our aim in what follows is to show that when $t/n \in (\tau^-, \tau^+)$ we could replace $\mathbf{x}\mathbf{H}^\intercal$ by $\mathbf{s} \in \mathbb{F}_q^{n-k}$ being uniformly distributed, without changing our problem. More precisely, we are going to show that distributions $(\mathbf{H}, \mathbf{x}\mathbf{H}^\intercal)$ and $(\mathbf{H}, \mathbf{s})$ are *statistically close* under the condition $t/n \in (\tau^-, \tau^+)$, showing that for any algorithm solving $\mathsf{DP}(n, q, R, \tau)$ we could choose its inputs as $(\mathbf{H}, \mathbf{s})$ without changing at much its success probability. This result may seem useless but in some applications where $\tau \in (\tau^-, \tau^+)$ it is much more comfortable to consider directly $(\mathbf{H}, \mathbf{s})$ rather than $(\mathbf{H}, \mathbf{x}\mathbf{H}^\intercal)$.

Proposition 6. *Let $k \stackrel{def}{=} \lfloor Rn \rfloor$, $t \stackrel{def}{=} \lfloor \tau n \rfloor$ and $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$, $\mathbf{s} \in \mathbb{F}_q^{n-k}$, $\mathbf{e} \in \mathscr{S}_t$ being uniformly distributed. We have*

$$\mathbb{E}_{\mathbf{H}}\left(\Delta\left(\mathbf{e}\mathbf{H}^{\intercal}, \mathbf{s}\right)\right) \leq \frac{1}{2}\sqrt{\frac{q^{n-k}-1}{\binom{n}{t}(q-1)^t}}. \tag{15}$$

In particular, when $\tau \in (\tau^-, \tau^+)$

$$\mathbb{E}_{\mathbf{H}}\left(\Delta\left(\mathbf{e}\mathbf{H}^{\intercal}, \mathbf{s}\right)\right) \leq q^{-\alpha n(1+o(1))} \quad \textit{where} \quad \alpha \stackrel{def}{=} \frac{1}{2}\left(h_q(\tau) - 1 + R\right) > 0.$$

Exercise 5. *Let $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ being uniformly distributed, $\mathbf{s} \in \mathbb{F}_q^{n-k}$, $\mathbf{e} \in \mathscr{S}_t$ be some random variables. Show that*

$$\mathbb{E}_{\mathbf{H}}\left(\Delta\left(\mathbf{e}\mathbf{H}^{\intercal}, \mathbf{s}\right)\right) = \frac{1}{q^{(n-k)\times n}} \sum_{\mathbf{H}_0 \in \mathbb{F}_q^{(n-k)\times n}} \Delta\left(\mathbf{e}\mathbf{H}_0^{\intercal}, \mathbf{s}\right)$$

The proof of Proposition 6 relies on the following lemma which is a rewriting in our context of the result known as the *left over hash lemma*. Roughly speaking, it shows that if the outputs of a function collide with a probability ε-close to the case where they would be randomly distributed, then this function is $\sqrt{\varepsilon}$-close to a random function.

Lemma 6. *Let $\mathscr{H} = (h_i)_{i\in I}$ be a finite family of applications from E in F. Let ε be the "collision bias"*

$$\mathbb{P}_{h,e,e'}(h(e) = h(e')) = \frac{1}{\sharp F}(1+\varepsilon)$$

where h is uniformly drawn in $\mathscr{H}$, e and e' be distributed according to some random variable X taking its values E. Let $\mathscr{U}$ be the uniform distribution over F and $\mathscr{D}(h)$ be the distribution $h(e)$ when e is distributed according to X. We have,

$$\mathbb{E}_h\left(\Delta(\mathscr{D}(h), \mathscr{U})\right) \leq \frac{1}{2}\sqrt{\varepsilon}.$$

Proof. By definition of the statistical distance we have

$$\begin{aligned}
\mathbb{E}_h\left(\Delta(\mathscr{D}(h), \mathscr{U})\right) &= \sum_{h\in\mathscr{H}} \frac{1}{\sharp\mathscr{H}} \Delta\left(\mathscr{D}(h), \mathscr{U}\right) \\
&= \frac{1}{2}\sum_{h\in\mathscr{H}} \frac{1}{\sharp\mathscr{H}} \sum_{f\in F} \left|\mathbb{P}_e(h(e) = f) - \frac{1}{\sharp F}\right| \\
&= \frac{1}{2}\sum_{(h,f)\in\mathscr{H}\times F} \left|\mathbb{P}_{h_0,e}(h_0 = h, h_0(e) = f) - \frac{1}{\sharp\mathscr{H}\ \sharp F}\right| \qquad (16) \\
&= \frac{1}{2}\sum_{(h,f)\in\mathscr{H}\times F} \left|q_{h,f} - \frac{1}{\sharp\mathscr{H}\ \sharp F}\right|. \qquad (17)
\end{aligned}$$

where $q_{h,f} \stackrel{\text{def}}{=} \mathbb{P}_{h_0,e}\left((h_0, h_0(e)) = (h, f)\right)$ with h_0 being uniformly chosen at random in $\mathscr{H}$ and e be distributed according to X. Using the Cauchy-Schwarz inequality, we obtain

$$\sum_{(h,f)\in\mathscr{H}\times F} \left|q_{h,f} - \frac{1}{\sharp\mathscr{H}\ \sharp F}\right| \leq \sqrt{\sum_{(h,f)\in\mathscr{H}\times F} \left(q_{h,f} - \frac{1}{\sharp\mathscr{H}\ \sharp F}\right)^2}\ \sqrt{\sharp\mathscr{H}\ \sharp F}. \tag{18}$$

Let us observe now that

$$\begin{aligned}\sum_{(h,f)\in\mathscr{H}\times F}\left(q_{h,f}-\frac{1}{\sharp\mathscr{H}\ \sharp F}\right)^2 &= \sum_{h,f}\left(q_{h,f}^2-2\frac{q_{h,f}}{\sharp\mathscr{H}\ \sharp F}+\frac{1}{\sharp\mathscr{H}^2\ \sharp F^2}\right)\\ &= \sum_{h,f}q_{h,f}^2-2\frac{\sum_{h,f}q_{h,f}}{\sharp\mathscr{H}\ \sharp F}+\frac{1}{\sharp\mathscr{H}\ \sharp F}\\ &= \sum_{h,f}q_{h,f}^2-\frac{1}{\sharp\mathscr{H}\ \sharp F}. \qquad (19)\end{aligned}$$

Consider for $i\in\{0,1\}$ independent random variables h_i and e_i that are drawn uniformly at random in $\mathscr{H}$ and according to X respectively. We continue this computation by noticing now that

$$\begin{aligned}\sum_{h,f}q_{h,f}^2 &= \sum_{h,f}\mathbb{P}_{h_0,e_0}(h_0=h,h_0(e_0)=f)\mathbb{P}_{h_1,e_1}(h_1=h,h_1(e_1)=f)\\ &= \mathbb{P}_{h_0,h_1,e_0,e_1}\left(h_0=h_1,h_0(e_0)=h_1(e_1)\right)\\ &= \frac{\mathbb{P}_{h_0,e_0,e_1}\left(h_0(e_0)=h_0(e_1)\right)}{\sharp\mathscr{H}}\\ &= \frac{1+\varepsilon}{\sharp\mathscr{H}\ \sharp F}. \qquad (20)\end{aligned}$$

By substituting for $\sum_{h,f}q_{h,f}^2$ the expression obtained in (20) into (19) and then back into (18) we finally obtain

$$\begin{aligned}\sum_{(h,f)\in\mathscr{H}\times F}\left|q_{h,f}-\frac{1}{\sharp\mathscr{H}\ \sharp F}\right| &\le \sqrt{\frac{1+\varepsilon}{\sharp\mathscr{H}\ \sharp F}-\frac{1}{\sharp\mathscr{H}\ \sharp F}}\ \sqrt{\sharp\mathscr{H}\ \sharp F}\\ &= \sqrt{\frac{\varepsilon}{\sharp\mathscr{H}\ \sharp F}}\ \sqrt{\sharp\mathscr{H}\ \sharp F}\\ &= \sqrt{\varepsilon}.\end{aligned}$$

□

We are now ready to prove Proposition 6.

Proof of Proposition 6. Let us compute the "collision bias" in our case. We have:

$$\begin{aligned}\mathbb{P}_{\mathbf{H},\mathbf{e},\mathbf{e}'}\left(\mathbf{e}\mathbf{H}^\intercal=\mathbf{e}'\mathbf{H}^\intercal\right) &= \mathbb{P}_{\mathbf{H},\mathbf{e},\mathbf{e}'}\left((\mathbf{e}-\mathbf{e}')\mathbf{H}^\intercal=\mathbf{0}\right)\\ &= \mathbb{P}_{\mathbf{H},\mathbf{e},\mathbf{e}'}\left((\mathbf{e}-\mathbf{e}')\mathbf{H}^\intercal=\mathbf{0}\mid\mathbf{e}\neq\mathbf{e}'\right)\mathbb{P}\left(\mathbf{e}\neq\mathbf{e}'\right)+\mathbb{P}_{\mathbf{e},\mathbf{e}'}(\mathbf{e}=\mathbf{e}')\\ &= \frac{1}{q^{n-k}}\left(1-\frac{1}{\binom{n}{t}(q-1)^t}\right)+\frac{1}{\binom{n}{t}(q-1)^t}\quad\text{(by Lemma 4)}\\ &= \frac{1}{q^{n-k}}\left(1+\varepsilon\right)\quad\text{where}\quad\varepsilon\stackrel{\text{def}}{=}\frac{q^{n-k}-1}{\binom{n}{t}(q-1)^t}\end{aligned}$$

which shows (15) by using Lemma 6. The second part of the proposition easily follows from the definition of τ^- and τ^+. □

Exercise 6. *Show that if one replaces in the binary case ($q=2$) $\mathbf{e}$ in Proposition 6 by $\mathbf{e}^{\mathrm{Ber}}$, where the e_i^{Ber}'s are distributed according to a Bernoulli distribution of parameter τ, we would obtain*

$$\mathbb{E}_{\mathbf{H}}\left(\Delta\left(\mathbf{e}^{\mathrm{Ber}}\mathbf{H}^\intercal,\mathbf{s}\right)\right)\le\frac{1}{2}\sqrt{2^{-k}\left(1+(1-2\tau)^2\right)^n}.$$

What can you deduce when comparing both results with $\mathbf{e}$ *or* $\mathbf{e}^{\mathrm{Ber}}$*? What is (according to Proposition 6) the "best" choice of error* $\mathbf{x}$ *to ensure that* $\mathbf{x}\mathbf{H}^\intercal$ *is uniformly distributed?*

Exercise 7. *Let* $\mathcal{C}$ *be a fixed* $[n,k]_q$*-code of parity-check matrix* $\mathbf{H}$ *and* $\mathbf{y},\mathbf{s},\mathbf{e} \in \mathbb{F}_q^n \times \mathbb{F}_q^{n-k} \times \mathscr{S}_t$ *be uniformly distributed. Our aim in this exercise is to show that* $\Delta(\mathbf{c}+\mathbf{e},\mathbf{y}) = \Delta(\mathbf{e}\mathbf{H}^\intercal,\mathbf{s})$.

1. *Given* $\mathbf{s} \in \mathbb{F}_q^{n-k}$*, let* $\mathbf{y}(\mathbf{s}) \in \mathbb{F}_q^n$ *be such that* $\mathbf{y}(\mathbf{s})\mathbf{H}^\intercal = \mathbf{s}$*. Show that*
$$\sum_{\mathbf{y}\in\mathbb{F}_q^n} \left| \mathbb{P}_{\mathbf{e},\mathbf{c}}(\mathbf{c}+\mathbf{e}=\mathbf{y}) - \frac{1}{q^n} \right| = \sum_{\mathbf{s}\in\mathbb{F}_q^{n-k}} \sum_{\mathbf{c}'\in\mathcal{C}} \left| \mathbb{P}_{\mathbf{e},\mathbf{c}}(\mathbf{c}+\mathbf{e}=\mathbf{y}(\mathbf{s})+\mathbf{c}') - \frac{1}{q^n} \right|.$$
2. *Deduce that* $\Delta(\mathbf{c}+\mathbf{e},\mathbf{y}) = \Delta(\mathbf{e}\mathbf{H}^\intercal,\mathbf{s})$.

Up to now, we have proven in Proposition 6 that syndromes $\mathbf{e}\mathbf{H}^\intercal$ are statistically close to the uniform distribution when $\mathbf{e}$ is picked uniformly at random in $\mathscr{S}_t$ with t larger than the Gilbert-Varshamov distance (more precisely, when $t/n \in (\tau^-,\tau^+)$) *but* in average over $\mathbf{H}$. One may ask if the result still holds for a fixed matrix $\mathbf{H}_0$? Actually we can prove, without too much effort, that the above result is true for almost all matrices but with a loss given by a square root as shown by the following proposition.

Proposition 7. *Let* $k \stackrel{def}{=} \lfloor Rn \rfloor$*,* $t \stackrel{def}{=} \lfloor \tau n \rfloor$*,* $\mathbf{H} \in \mathbb{F}_q^{(n-k)\times n}$ *being uniformly distributed and* $\mathbf{s} \in \mathbb{F}_q^{n-k}$*,* $\mathbf{e} \in \mathbb{F}_q^n$ *be some random variables. Suppose that*
$$\mathbb{E}_{\mathbf{H}}\left(\Delta\left(\mathbf{e}\mathbf{H}^\intercal,\mathbf{s}\right)\right) \leq \varepsilon$$
Then, we have
$$\frac{\sharp\left\{\mathbf{H}_0 \in \mathbb{F}_q^{(n-k)\times n} \ : \ \Delta\left(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s}\right) \geq \sqrt{\varepsilon}\right\}}{q^{(n-k)\times n}} \leq \sqrt{\varepsilon}$$

Proof. First,
$$\mathbb{E}_{\mathbf{H}}\left(\Delta\left(\mathbf{e}\mathbf{H}^\intercal,\mathbf{s}\right)\right) = \frac{1}{q^{(n-k)\times n}} \sum_{\mathbf{H}_0\in\mathbb{F}_q^{(n-k)\times n}} \Delta\left(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s}\right). \tag{21}$$
The idea of the proof is to split in two parts this sum according to the terms that are larger and smaller than $\sqrt{\varepsilon}$
$$\begin{aligned}
\sum_{\mathbf{H}_0\in\mathbb{F}_q^{(n-k)\times n}} \Delta\left(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s}\right) &= \sum_{\substack{\mathbf{H}_0 \ : \\ \Delta(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s})<\sqrt{\varepsilon}}} \Delta\left(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s}\right) + \sum_{\substack{\mathbf{H}_0 \ : \\ \Delta(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s})\geq\sqrt{\varepsilon}}} \Delta\left(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s}\right) \\
&\geq \sum_{\substack{\mathbf{H}_0 \ : \\ \Delta(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s})<\sqrt{\varepsilon}}} \Delta\left(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s}\right) + \sum_{\substack{\mathbf{H}_0 \ : \\ \Delta(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s})\geq\sqrt{\varepsilon}}} \sqrt{\varepsilon} \\
&\geq \sum_{\substack{\mathbf{H}_0 \ : \\ \Delta(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s})\geq\sqrt{\varepsilon}}} \sqrt{\varepsilon} \\
&= \sqrt{\varepsilon}\,\sharp\left\{\mathbf{H}_0 \in \mathbb{F}_q^{(n-k)\times n} \ : \ \Delta\left(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s}\right) \geq \sqrt{\varepsilon}\right\} \qquad (22)
\end{aligned}$$
By plugging Equation (22) in (21), we obtain
$$\mathbb{E}_{\mathbf{H}}\left(\Delta\left(\mathbf{e}\mathbf{H}^\intercal,\mathbf{s}\right)\right) \geq \sqrt{\varepsilon}\,\frac{\sharp\left\{\mathbf{H}_0 \in \mathbb{F}_q^{(n-k)\times n} \ : \ \Delta\left(\mathbf{e}\mathbf{H}_0^\intercal,\mathbf{s}\right) \geq \sqrt{\varepsilon}\right\}}{q^{(n-k)\times n}}$$
But by assumption we have $\mathbb{E}_{\mathbf{H}}\left(\Delta\left(\mathbf{e}\mathbf{H}^\intercal,\mathbf{s}\right)\right) \leq \varepsilon$. It concludes the proof. □