Summary of basic notions from previous lecture:

- "error correcting code" $C \subseteq \mathcal{X}^n$ with $\begin{cases} \mathcal{X}: \text{ finite alphabet} \\ n \in \mathbb{N}: \text{ "length" of the code} \end{cases}$
- "rate" of an ECC, $R(C) := \frac{\log|C|}{n \log|\mathcal{X}|} \sim \frac{\text{length of message}}{\text{length of its codeword}}$
  $\sim$ fraction of non-redundant info
- "distance" of an ECC: $d(C) := \min_{\substack{c,c' \in C \\ c \neq c'}} d(c-c')$ = min. Hamming distance between two codewords
- "relative distance": $\frac{d(C)}{n}$

remember: an ECC with $d := d(C)$ allows to correct $\lfloor \frac{d-1}{2} \rfloor$ errors or $(d-1)$ symbol erasures

V.2. Linear codes

Def.: If $\mathcal{X}$ is a field and $C \subseteq \mathcal{X}^n$ a subspace, then C is called a "linear code".

remarks:
- $|\mathcal{X}| < \infty$ implies that $\mathcal{X} = GF(q)$ is a "Galois field" with $q := |\mathcal{X}| = p^m$ for some prime p and $m \in \mathbb{N}$.
- A subspace $C \subseteq GF(q)^n$ admits a basis $c_1, \dots, c_k$ so that $\boxed{|C| = q^k}$ & thus $\boxed{R(C) = \frac{k}{n}}$
- for real world applications we often have $n \sim 10^3 - 10^4$

Def.:
- $G \in GF(q)^{n \times k}$ is called a "generator matrix" for a linear code $C \subseteq GF(q)^n$ if its columns form a basis of C.
- C is then called an "$[n,k]$-code" or "$[n,k,d]$-code" if $d = d(C)$.

remark: the encoding map $E: GF(q)^k \to GF(q)^n$ of a linear code is then just $E: x \mapsto Gx$.

Lemma: For any linear code $C \subseteq GF(q)^n$ we have

$$\boxed{d(C) = \min_{c \in C \setminus \{0\}} d(c)}.$$

proof:

- let $c_1, c_2 \in C$ be such that $d(c_1 - c_2) = d(C)$.

  $\tilde{c} := c_1 - c_2 \in C \setminus \{0\}$ then implies $d(C) = d(\tilde{c}) \geq \min_{c \in C \setminus \{0\}} d(c)$.

- conversely, if $c_1 \in C \setminus \{0\}$ s.t. $d(c_1) = \min_{c \in C \setminus \{0\}} d(c)$, then for $c_2 := 0$

  $$d(C) \leq d(c_1 - c_2) = d(c_1) = \text{-"-}.$$

□

Def.: A generator matrix $G \in GF(q)^{n \times k}$ is said to be "in systematic form" if $G = (\mathbb{1}_k \; P)^T$ for some $k \times (n-k)$ matrix P. The encoding $x \mapsto Gx$ is then also called "systematic"

remarks:

- for every code with generator matrix G' we can by linear operations construct one with generator matrix G in sys. form s.t. the two codes are "equivalent" in the sense that their lengths, rates & min. distances coincides.
- the codewords of a syst. encoding contain the raw message in the first k components followed by $(n-k)$ symbols introducing redundancy.

Prop.: Let $G = (\mathbb{1}_k \; P)^T$ be the generator matrix of a linear code $C \subseteq GF(q)^n$. Then $\forall c \in GF(q)^n$:

$$\boxed{c \in C \Leftrightarrow Hc = 0} \quad \text{for } H := (-P^T \; \mathbb{1}_{n-k}).$$

proof:

- $c \in C \Rightarrow \exists x \in GF(q)^k : c = Gx$

  $\Rightarrow Hc = HGx = (-P^T \; \mathbb{1}) \begin{pmatrix} \mathbb{1} \\ P^T \end{pmatrix} x = (P^T - P^T)x = 0$ ✓

- $Hc = 0 \Rightarrow 0 = (-P^T \; \mathbb{1}) \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = c_2 - P^T c_1 \Rightarrow c_2 = P^T c_1$

$\Rightarrow c = \begin{pmatrix} c_1 \\ c_2 \end{pmatrix} = \begin{pmatrix} c_1 \\ P^T c_1 \end{pmatrix} = \begin{pmatrix} \mathbb{1} \\ P^T \end{pmatrix} c_1 = G c_1$ ✓ □

remarks:
- if $c \in C$ is corrupted via $c \mapsto c' := c + e$, then $Hc' = He$ is independent of the original codeword.
- He is called "syndrome" & H is called "parity check matrix"
- a possible decoding strategy is then to infer/guess e from the syndrome.

V.3. Bounds on the performance of error correcting codes

Prop.: (Hamming bound) Let $C \subseteq \mathcal{X}^n$ be a code with $|\mathcal{X}| = q$, distance $d(C) := d$ and $m := \lfloor \frac{d-1}{2} \rfloor$ (= # of errors which can be corrected).

Then $$\boxed{|C| \leq \frac{q^n}{V(q,n,m)}}$$ where $V(q,n,m) := \sum_{i=0}^{m} \binom{n}{i} (q-1)^i$

proof: For each $c \in C$ define a neighborhood $B_m(c) := \{ y \in \mathcal{X}^n \mid d(y,c) \leq m \}$.

Then $B_m(c) \cap B_m(c') = \emptyset$ for $c, c' \in C$ with $c \neq c'$ and $|B_m(c)| = V(q,n,m)$.

So $$|\mathcal{X}^n| = q^n \geq \Big| \bigcup_{c \in C} B_m(c) \Big| = \sum_{c \in C} |B_m(c)| = |C| \, V(q,n,m)$$

($\uparrow$ $B(c)$'s disjoint) □

remark: if '=' holds in the Hamming bound, then we have a perfect packing of non-overlapping Hammingballs that cover the full space.

Def.: A code for which '=' holds in the Hamming bound is called „perfect".

Thm.: (Tietavainen / van Lint '70ies) The following are all perfect binary codes (i.e. $q=2$):

(i) $[2^r-1,\ 2^r-1-r,\ 3]$ Hamming codes (e.g. $[7,4]$ for $r=3$)

(ii) the "$[23,12,7]$ Golay code"

(iii) trivial codes (meaning $|C| \in \{1, 2^n\}$)

(iv) repetition codes $x_i \mapsto \underbrace{(x_i, \dots, x_i)}_{n \text{ times}}$ for odd n

Thm.: (Gilbert-Varshamov bound)

For every triple $(q,n,d) \in \mathbb{N}^3$ there exist a code $C \subseteq \mathcal{X}^n$ with $|\mathcal{X}| = q$ and distance $d(C) = d$ s.t.

$$\boxed{|C| \geq \frac{q^n}{V(q,n,d-1)}}$$

where $V(q,n,d-1) = \sum_{i=0}^{d-1} \binom{n}{i} (q-1)^i$

= Volume of $B_{d-1} \in \mathcal{X}^n$

proof: construct the code step-by-step via:

(i) start with arbitrary first codeword

(ii) add any point as a codeword which has Hamming distance at least d from all previously chosen codewords,

(iii) iterate (ii) until the Hamming balls of radius $(d-1)$ around the codewords cover all of $\mathcal{X}^n$.

The constructed code then satisfies $|C| \cdot V(q,n,d-1) \geq q^n$.

□

remarks:

- There are linear codes satisfying this bound. In fact, random linear codes do the job for large enough n.

- computing (even approximating) the distance of a linear code is NP-hard
  → picking a random code & checking whether it has good distance is not feasible.
- for prime powers ≥ 49 there are explicit constructions based on algebraic geometry which satisfy the GV bound.
- for $q=2$ no explicit construction is known.