Foundations of Probabilistic Proofs (Fall 2022)
Note 7: Linearity Testing
Date: 2022.10.11

This note contains definitions, theorems, facts, *etc.* that are not fully explained in lectures due to limited time. If you think there are anything missing or any mistakes, please contact ziyi.guan@epfl.ch.

Some of the definitions and the exercise presented in this note are adapted from the course materials of *Algebraic Error Correcting Codes* taught by Professor Mary Wootters at Stanford University. We direct interesting readers to refer to the course website `https://web.stanford.edu/~marykw/classes/CS250_W18/` for more information.

1 Coding Theory

1.1 Code

In the lecture, we go over the BLR-test, which determines if a function is linear or not. We mention that it implicitly rely on the Hadamard code. Here we give a brief introduction to the coding theory.

A basic problem in the coding theory is how to encode messages efficiently and how to deal with data corruption during communication. For example, we prove in lecture that given a function f that is very close to a linear function f_{LIN}, it is possible to recover f_{LIN} from f.

Formally, given a finite alphabet Σ and $n \in \mathbb{Z}_{>0}$, a code C with ***block length*** n over the Alphabet Σ is a subset of Σ^n. We introduce other characteristics of code C in the followings:

- A ***codeword*** of the code C is an element $c \in \mathrm{C}$.
- The ***message length*** of the code C is $k \coloneqq \log_{|\Sigma|} |\mathrm{C}|$.
  - Note that different messages should have different encodings, thus $|\Sigma|^k = |\mathrm{C}|$.
- The ***rate*** of the code C is denoted as R, and $\mathrm{R} \coloneqq \frac{k}{n}$.
- The ***distance*** of the code C is the minimum Hamming distance between codewords, that is, $d \coloneqq \min_{c \neq c' \in \mathrm{C}} \Delta(c, c')$.
  - Hamming distance: $\Delta(x, y) \coloneqq \sum_{i=1}^{n} \mathbb{1}\ (x_i \neq y_i)$
  - Relative Hamming distance: $\delta(x, y) \coloneqq \frac{1}{n} \sum_{i=1}^{n} \mathbb{1}\ (x_i \neq y_i) = \frac{\Delta(x,y)}{n}$
- A $(n, k, d)_{|\Sigma|}$ code: code with block length n, message length k, distance d, and alphabet Σ.

Example 1. $\mathrm{C} \coloneqq \{(0,0,0,0),(1,0,0,1),(0,1,0,1),(0,0,1,1),(0,1,1,0),(1,0,1,0),(1,1,0,0),(1,1,1,1)\}$ *is a code of length* 4 *over* $\Sigma = \{0, 1\}$ *(in which case we call* C *a binary code). To be concrete,* C *is a* $(4, 3, 2)_2$ *code. (One can verify this easily.)*

Intuitively, we would like a code with small block length, i.e. we want R to be close to 1. However, we can show that R can not be any close to 1. To illustrate it we need to define the Hamming ball first:

Definition 1. *The* **Hamming ball** *in* Σ^n *of radius* r *for* $x \in \Sigma^n$ *is*

$$\mathrm{B}_{\Sigma^n}(x, r) \coloneqq \{y \in \Sigma^n \colon \Delta(x, y) \leq r\}$$

The volume of $\mathrm{B}_{\Sigma^n}(x, r)$ *is* $\mathrm{Vol}_{|\Sigma|}(r, n) \coloneqq |\mathrm{B}_{\Sigma^n}(x, r)|$.

The definition of the volume makes sense because $|\mathrm{B}_{\Sigma^n}(x, r)|$ does not depend on x. Now we can use the Hamming Ball to find a bound for the rate, which is called the ***Hamming bound***.

Theorem 1. *For a* $(n, k, d)_{|\Sigma|}$ *code, the Hamming bound of its rate* R *is:*

$$\mathrm{R} \leq 1 - \frac{\log_{|\Sigma|}(\mathrm{Vol}_{|\Sigma|}(\left\lfloor \frac{d-1}{2} \right\rfloor, n))}{n}$$

Proof. We prove by checking the Hamming Ball of each codeword.

- The $(n, k, d)_{|\Sigma|}$ code C is a subset of Σ^n, thus:

$$|\mathrm{C}| \cdot \mathrm{Vol}_{|\Sigma|}(\left\lfloor \frac{d-1}{2} \right\rfloor, n) \leq |\Sigma|^n .$$

- By taking log of both sides, we have:

$$\mathrm{R} = \frac{k}{n} \leq 1 - \frac{\log_{|\Sigma|}(\mathrm{Vol}_{|\Sigma|}(\left\lfloor \frac{d-1}{2} \right\rfloor, n))}{n}$$

□

Example 2. *Suppose* $\mathsf{Enc} \colon \{0,1\}^4 \to \{0,1\}^7$ *with* $\mathsf{Enc}(x_1, x_2, x_3, x_4) \coloneqq (x_1, x_2, x_3, x_4, x_2 + x_3 + x_4, x_1 + x_3 + x_4, x_1 + x_2 + x_4)$. *Let* $\mathrm{C} \coloneqq \mathrm{Img}(\mathsf{Enc})$. *Thus* $\mathrm{C} \subseteq \{0,1\}^7$ *is a* $(7, 4, 3)_2$ *code with tight Hamming bound, since* $\mathrm{Vol}_2(1, 7) = 1 + \binom{7}{1} \times 1 = 8$. *This code is called a Hamming code.*

Exercise 1. *Show that the code* C *in Example 2 has distance* 3*, thus it has tight Hamming bound.*

Solution. It suffices to show that $\min_{c \in \mathrm{C} \setminus \{0\}} \mathrm{wt}(c) \geq 3$ where $\mathrm{wt}(c)$ denotes the number of non-zero elements in c. ■

1.2 Hadamard code

The Hadamard code is a linear code with some nice properties illustrated in the lecture. That is, we can efficiently check if f is a Hadamard codeword (or linear function) and recover a Hadamard codeword from small corruption.

We first give a definition of the linear code:

Definition 2 (Linear code). *A linear code is a code for which any linear combination of codewords is also a codeword.*

Example 3. *The code* C *defined in Example 2 is a linear code, since for any codewords* $c_1 \coloneqq \mathsf{Enc}(x_1, x_2, x_3, x_4)$ *and* $c_2 \coloneqq \mathsf{Enc}(y_1, y_2, y_3, y_4)$*, we have* $ac_1 + bc_2 = \mathsf{Enc}(ax + by) \in \mathrm{C}$.

Exercise 2. *Show that a linear code of block length* n *and message length* k *over a finite field* $\mathbb{F}$ *(which means that the alphabet is* $\mathbb{F}$*) is a* k*-dimensional linear subspace of* $\mathbb{F}^n$.

Solution. We will prove by contradiction:

- A linear code C is a linear subspace by definition, we need to show that its dimension is k.
- Assume that the dimension of the subspace is $k+1$, that is, the subspace has $k+1$ linearly independent basis $\{c_i\}_{i\in[k]}$ and every codeword c can be written as $\sum_{i=0}^{k} b_i c_i$.
- Note that two codeword $c = c'$ iff their corresponding representations $\{b_i\}_{i\in[k]}$, $\{b'_i\}_{i\in[k]}$ are the same. That is to say, there is a bijection from $c \in \mathrm{C}$ to $\{b_i\}_{i\in[k]}, b_i \in \mathbb{F}$.
- $|\mathrm{C}| = |\mathbb{F}|^{k+1}$, which is a contradiction with the message length being k.

■

As mentioned in the lecture, the set of all linear functions is precisely the Hadamard code. Assume without loss of generality that $\mathbb{F} = \mathbb{F}_2$, we introduce Hadamard code and its properties.

Definition 3 (Hadamard code)**.** *The Hadamard code is a subset* $\mathrm{C} \subseteq \{0,1\}^{2^n}$ *which is the image of the encoding function* $\mathrm{Had}\colon \{0,1\}^n \to \{0,1\}^{2^n}$*. The encoding function* Had *encodes a message* $u \in \{0,1\}^n$ *to the sequence of all inner product with* u*. That is,*

$$\mathrm{Had}(u) \coloneqq (\langle u, a\rangle)_{a\in\{0,1\}^n}.$$

We have the following observations:

- The Hadamard code's codeword block length is 2^n with message length n.
- The Hadamard code is a linear code.
  - Any linear combination of Hadamard codewords is also a Hadamard codeword because:

$$\begin{aligned}\alpha\mathrm{Had}(u) + \beta\mathrm{Had}(v) &= \alpha(\langle u, a\rangle)_{a\in\{0,1\}^n} + \beta(\langle u, a\rangle)_{a\in\{0,1\}^n} \\ &= (\langle \alpha u + \beta v, a\rangle)_{a\in\{0,1\}^n} \\ &= \mathrm{Had}(\alpha u + \beta v)\end{aligned}$$

- The relative distance of the Hadamard code is $\frac{1}{2}$ (for general $\mathbb{F}$, it will be $1 - \frac{1}{|\mathbb{F}|}$).
- The Hadamard code is the truth table of LIN, which is the set of all linear functions from $\{0,1\}^n$ to $\{0,1\}$.
  - As defined in the lecture, a function $f\colon \mathbb{F}^n \to \mathbb{F}$ is linear iff there exists $c \in \mathbb{F}^n$ such that for every $x \in \mathbb{F}^n$, $f(x) = \sum_{i=1}^{n} c_i x_i$.
  - The truth table of LIN is a $2^n \times 2^n$ table where the rows are indexed by input values and the columns are indexed by the linear functions.
  - A Hadamard codeword encoding from the message u is precisely the truth table of the linear function $f(x) \coloneqq \langle u, x\rangle = \sum_{i=1}^{n} u_i x_i$, that is, the values of f over every possible x.
  - There is a natural bijection from every $u \in \mathbb{F}^n$ to every $f \in \mathrm{LIN}$: $u \to f(x) = \sum_{i=1}^{n} u_i x_i$.

Example 4. *We show the Hadamard code for the case* $n = 2$*:*

x	$\text{Had}(0,0)$	$\text{Had}(0,1)$	$\text{Had}(1,0)$	$\text{Had}(1,1)$
$(0,0)$	0	0	0	0
$(0,1)$	0	1	0	1
$(1,0)$	0	0	1	1
$(1,1)$	0	1	1	0

Every column is the truth table of a linear function $f(x) := \sum_{i=1}^{n} u_i x_i$.

Based on the observations above, it can be clearly seen that the Hadamard code has two nice properties as we mention at the beginning of the subsection:

- Local testablity: Given a function $f\colon \{0,1\}^n \to \{0,1\}$, we are able to know whether there exists $u \in \{0,1\}^n$ such that $f(x) = \text{Had}(u)_x$ for all $x \in \{0,1\}^n$ (or just $f = \text{Had}(u)$ for simplicity). In other words, we can know if $f \in \text{LIN}$.
  - Simply use V_{BLR}.
- Local decodability: Given a function f that is close to some linear function $\widehat{f}$, we are able to learn $\widehat{f}(x)$ for any $x \in \{0,1\}^n$.
  - Local correction: Sample $y \in \{0,1\}^n$ and return $f(x+y) - f(y)$.