CSE 190, Great ideas in algorithms: Error correcting codes - basics

1 Definitions

An error correcting code encodes messages into longer codewords, such that even in the presence of errors, it can decode the original message. Here, we focus on "worst case errors", where we make no assumptions on the distribution of errors, but instead limit the number of errors.

Definition 1.1. *Let Σ be a finite alphabet, $k \geq 1$ be the message length, $n \geq k$ be the codeword length (also called the block length). An error correcting code is simply a subset*

$$\mathcal{C} \subset \Sigma^n$$

of size $|\mathcal{C}| = |\Sigma|^k$, together with a one-to-one encoding map

$$E : \Sigma^k \to \mathcal{C}$$

and decoding map

$$D : \Sigma^n \to \Sigma^k.$$

To describe the error correcting capability of a code, define the distance of $x, y \in \Sigma^n$ as the number of coordinates where they differ,

$$\text{dist}(x, y) = |\{i \in [n] : x_i \neq y_i\}.$$

Definition 1.2 (Error correction capability of a code)**.** *A code $(\mathcal{C}, E, D)$ can correct up to e errors if for any message $m \in \Sigma^k$ and any $x \in \Sigma^n$ such that $\text{dist}(E(m), x) \leq e$, it holds that then $D(x) = m$.*

Example 1.3 (Repetition code)**.** *Let $\Sigma = \{0, 1\}, k = 1, n = 3$. Define $\mathcal{C} = \{000, 111\}$, $E : \{0, 1\} \to \{0, 1\}^3$ by $E(0) = 000, E(1) = 111$. Define $D : \{0, 1\}^3 \to \{0, 1\}$ by $D(x_1, x_2, x_3) = \text{Majority}(x_1, x_2, x_3)$. Then $(\mathcal{C}, E, D)$ can recover from up to 1 error.*

If we just care about combinatorial bounds (that is, ignore algorithmic aspects), then a code is defined by its codewords. We can define $E : \Sigma^k \to \mathcal{C}$ is any one-to-one way, and $D : \Sigma^n \to \Sigma^k$ by mapping $x \in \Sigma^n$ to the closest codeword $E(m)$, breaking ties arbitrarily. So from now on, we simply describe codes by describing the set of codewords $\mathcal{C}$. Once we start discussing algorithms, we will revisit this assumption.

Definition 1.4. *The minimal distance of $\mathcal{C}$ is the minimal distance of any two distinct codewords,*

$$\mathrm{dist}_{\min}(\mathcal{C}) = \min_{x \neq y \in \mathcal{C}} \mathrm{dist}(x, y).$$

Definition 1.5. *An (n, k, d) code over Σ is a set of codewords $\mathcal{C} \subset \Sigma^n$ of size $|\mathcal{C}| = |\Sigma|^k$ and minimal distance $\geq d$.*

Lemma 1.6. *Let $\mathcal{C}$ be an $(n, k, 2e+1)$ code. Then it can decode from e errors.*

Proof. Let $x \in \mathcal{C}$, $y \in \Sigma^n$ be such that $\mathrm{dist}(x, y) \leq e$. We claim that x is the unique closest codeword to y. Assume not, that is there is another $x' \in \mathcal{C}$ with $\mathrm{dist}(x', y) \leq e$. Then by the triangle inequality, $\mathrm{dist}(x, x') \leq \mathrm{dist}(x, y) + \mathrm{dist}(y, x') \leq 2e$, which contradicts the assumption that the minimal distance of $\mathcal{C}$ is $2e + 1$. □

Moreover, if $\mathcal{C}$ has minimal distance d, then there exist $x, x' \in \mathcal{C}$ and $y \in \Sigma^n$ such that $\mathrm{dist}(x, y) + \mathrm{dist}(x', y) = d$, so the bound is tight. So, we can restrict our study to the existence of (n, k, d) codes.

2 Basic bounds

Lemma 2.1 (Singleton bound)**.** *Let $\mathcal{C}$ be an (n, k, d)-code over Σ. Then $k \leq n - d + 1$.*

Proof. Let $\mathcal{C}'$ be a code obtained by deleting the first $d - 1$ coordinates of all codewords in $\mathcal{C}$. Note that all codewords remain distinct, as we assume the minimal distance is at least d. So, $\mathcal{C}' \subset \Sigma^{n-d+1}$ has $|\Sigma|^k$ distinct elements. This implies $k \leq n - d + 1$. □

An MDS code (Maximal Distance Separable) is a code for which $k = n - d + 1$. We will later see an example of such a code (Reed-Solomon code).

Lemma 2.2 (Hamming bound)**.** *Let $\mathcal{C}$ be an $(n, k, 2e+1)$-code over Σ with $|\Sigma| = q$. Then*

$$q^k \sum_{i=0}^{e} \binom{n}{i} (q-i)^i \leq q^n.$$

Proof. For each codeword $x \in \mathcal{C}$ define the ball of radius e around it,

$$B(x) = \{y \in \Sigma^n : \mathrm{dist}(x, y) \leq e\}.$$

These ball cannot intersect by the minimal distance requirement. Each ball contains $\sum_{i=0}^{e} \binom{n}{i}(q-1)^i$ elements, and there are q^k such balls. The lemma follows. □

In different scenarios, these bounds provide different bounds, and none is always superior to the other.

Example 2.3. *Let $d = 3$, corresponding to correcting 1 error. The singleton bound gives $k \leq n - 2$ over any alphabet. The hamming bound gives (for $e = 1$)*

$$q^k(1 + (q-1)n) \leq q^n.$$

For binary codes, eg $q = 2$, it gives $2^k(n+1) \leq 2^n$, which gives $k \leq n - \log_2(n+1)$. However, if we take $q \to \infty$, then it translates to $k \leq n - 1 + o(1)$.

3 Existence of asymptotically good codes

An (n, k, d) code is said to be asymptotically good (or simply good) if $k = \alpha n, d = \beta n$ for some constants $\alpha, \beta > 0$. More precisely, we consider families of codes with growing n and fixed $\alpha, \beta > 0$. The singleton bound implies that $\alpha + \beta \leq 1$, and MDS codes achieve that. It is unknown if over binary alphabet it is achievable, and it is one of the major open problems in coding theory. Here, we will show that good codes exist for some constants α, β, without trying to optimize them.

Lemma 3.1. *There exists a family $\mathcal{C} \subset \{0,1\}^n$ of size $2^{n/10}$ such that $\mathrm{dist}_{\min}(\mathcal{C}) \geq n/10$.*

Proof. The proof is probabilistic. For $N = 2^{n/10}$ let $x_1, \ldots, x_N \in \{0,1\}^n$ be uniformly chosen. We claim that with high probability, $\mathcal{C} = \{x_1, \ldots, x_N\}$ is as claimed. To see that, lets consider the probability that $\mathrm{dist}(x_i, x_j) \leq n/10$ for some fixed $1 \leq i < j \leq N$. The number of choices for x_i is 2^n. Given x_i, the number of choices for x_j of distance at most $n/10$ from x_i is $\sum_{i=0}^{n/10} \binom{n}{i}$. This should be divided by the total number of pairs, which is 2^{2n}. So,

$$\Pr[\mathrm{dist}(x_i, x_j) \leq n/10] \leq \frac{2^n \sum_{i=0}^{n/10} \binom{n}{i}}{2^{2n}} \leq \frac{n\binom{n}{n/10}}{2^n}.$$

We need some estimates for the binomial coefficient. A useful one is

$$\left(\frac{n}{m}\right)^m \leq \binom{n}{m} \leq \left(\frac{en}{m}\right)^m.$$

So,

$$\binom{n}{n/10} \leq \left(\frac{en}{n/10}\right)^{n/10} \leq ((10e)^{1/10})^n \leq (1.4)^n.$$

So,

$$\Pr[\mathrm{dist}(x_i, x_j) \leq n/10] \leq \frac{n(1.4)^n}{2^n} = n(0.7)^n.$$

Now, the probability that there exists some pair $1 \leq i < j \leq N$ such that $\mathrm{dist}(x_i, x_j) \leq n/10$ can be upper bounded by the union bound,

$$\begin{aligned}\Pr[\exists 1 \leq i < j \leq N, \mathrm{dist}(x_i, x_j) \leq n/10] &\leq \sum_{1 \leq i < j \leq N} \Pr[\mathrm{dist}(x_i, x_j) \leq n/10] \\ &\leq N^2 n(0.7)^n = 2^{2n/10} n(0.7)^n \leq n(0.81)^n.\end{aligned}$$

So, the probability that $\mathrm{dist}_{\min}(\mathcal{C}) \leq n/10$ is at most most $n(0.81)^n$, which is exponentially small. Hence, with very high probability, the randomly chosen code will be a good code. □

We can get the same bounds without using probability. Consider the following process for choosing $x_1, x_2, \ldots, x_N \in \{0,1\}^n$. Pick x_1 arbitrarily, and delete all points of distance $\leq n/10$ from it; pick x_2 from the remaining points, and delete all points of distance $\leq n/10$

from it; pick x_3 from the remaining points, and so on. Continue in such a way until all points are exhausted. The number of points chosen N satisfies that

$$N \geq \frac{2^n}{\sum_{i=0}^{n/10} \binom{n}{i}}.$$

This is because we have initially a total of 2^n points, and at each point we delete at most $\sum_{i=0}^{n/10} \binom{n}{i}$ undelete points. The same calculations as before show that $N \geq 2^{n/10}$.

4 Linear codes

A special family of codes are *linear codes*. Let $\mathbb{F}$ be a finite field. In a linear code, $\Sigma = \mathbb{F}$ and $\mathcal{C} \subset \mathbb{F}^n$ is a k-dimensional subspace. The encoding map is a linear map: $E(x) = Ax$ where A is a $n \times k$ matrix over $\mathbb{F}$. Note that $\text{rank}(A) = k$, as otherwise the set of codewords will have dimension less than k. In practice, nearly all codes are linear, as the encoding map is easy to define. However, the decoding map needs inherently to be nonlinear, and is usually the hardest to compute.

Claim 4.1. *Let $\mathcal{C}$ be a linear code. Then* $\text{dist}_{\min}(\mathcal{C}) = \min_{0 \neq x \in \mathcal{C}} \text{dist}(0, x)$.

Proof. If $x_1, x_2 \in \mathcal{C}$ have the minimal distance, then $\text{dist}(x_1, x_2) = \text{dist}(0, x_1 - x_2)$ and $x_1 - x_2 \in \mathcal{C}$. □

We can view the decoding problem from either erasures or errors as a linear algebra problem. Let A be an $n \times k$ matrix. Codewords are $\mathcal{C} = \{Ax : x \in \mathbb{F}^k\}$, or equivalently the subspace spanned by the columns of A.

Decoding from erasures. The problem of decoding from erasures is equivalently the following problem: given $y \in (\mathbb{F} \cup \{?\})^n$, find $x \in \mathbb{F}^k$ such that $(Ax)_i = y_i$ for all $y_i \neq ?$. Equivalently, we want the sub-matrix formed by keeping only the rows $\{i \in [n] : y_i \neq ?\}$ to form a rank k matrix. So, the requirement that a linear code can be uniquely decoded from e errors, is equivalent to the requirement that if any e rows are deleted in the matrix, it still has rank k. Clearly, $e \leq n - k$. We will see a code achieving this bound, the Reed-Solomon code.

Decoding from errors. The problem of decoding from e errors is the following problem: given $y \in \mathbb{F}^n$, find $x \in \mathbb{F}^k$ such that $(Ax)_i \neq y_i$ for at most e coordinates. Equivalently, we want to find a vector spanned by the columns of A, which agrees with y in at least $n - e$ coordinates. If the code has minimal distance d, then we know that this is mathematically possible whenever $e < d/2$; however, finding this vector is in general hard. We will see a code where this is possible, and which moreover has the best minimal distance, $d = n - k + 1$. Again, it will be Reed-Solomon code.