Lecture 3: The Volume Bound and Finite Fields

Anup Rao

October 1, 2019

We have seen several bounds giving tradeoffs between the rate and relative distance of codes. They are:

Singleton bound $R + \delta \leq 1 + O(\log(n)/n)$.

Hamming bound If $\delta \leq 1 - 1/q$, then $R + h_q(\delta/2) \leq 1 + O(\log(n)/n)$.

Gilbert-Varshamov bound If $\delta \leq 1 - 1/q$, then there is a code with $R + h_q(\delta) \geq 1$.

Elias-Bassalygo bound When $q = 2$ and $\delta \leq 1/2$, then

$$R + h_2\Big(\frac{1 - \sqrt{1 - 2\delta}}{2}\Big) \leq 1 + O(\log(n)/n).$$

Many of these bounds relied on our estimates for the volume of balls in the Hamming metric. The estimate says that when $r/n \leq 1 - 1/q$,

$$q^{nh_q(r/n) + O(\log n)} \leq \mathsf{vol}(r) \leq q^{nh_q(r/n)}.$$

Today, we start by proving these estimates on the volume.

Entropy

A basic formula of immense consequence to the topic of coding theory (as well as many other areas) is the definition of Shannon's entropy function. Given a random variable A supported on a finite set, the entropy of A is

Here, and everywhere not specified, logarithm is computed base 2.

$$H(A) = \sum_a p(A = a) \log(1/p(A = a)).$$

The entropy is always non-negative, and maximized when A is uniform, and then it is $\log |\mathsf{supp}(A)|$. Moreover, we have:

$$H(AB) = H(A) + H(B|A),$$

where here $H(B|A) = \mathbb{E}_{a \leftarrow A}\left[H(B|A = a)\right]$. The entropy is subadditive:

$$H(AB) \leq H(A) + H(B).$$

Subadditivity implies that the conditional entropy can only be smaller than the entropy:

$$H(A|B) = H(AB) - H(B) \leq H(A) + H(B) - H(B) = H(A).$$

Let $|\Sigma| = q$. We shall give a bound on $\mathsf{vol}(r)$ when $r/n \leq 1 - 1/q$. When $r/n > 1 - 1/q$, the ball $B(x,r)$ will actually occupy a constant fraction of the whole space. We shall prove that when $r/n \leq 1 - 1/q$, $\mathsf{vol}(r) \leq q^{h_q(r/n)}$, where

$$h_q(\epsilon) = (1-\epsilon)\log_q\left(\frac{1}{1-\epsilon}\right) + \epsilon\log_q\left(\frac{q-1}{\epsilon}\right).$$

Observe that $h_2(\epsilon)$ is just Shannon's entropy function.

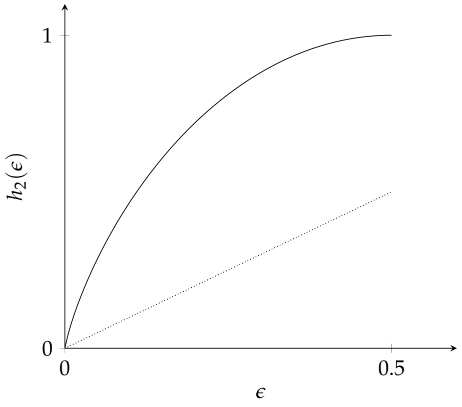

Figure 1: $h_2(\epsilon)$.

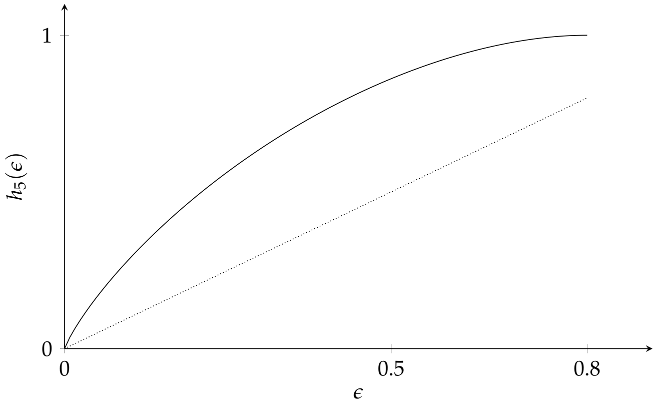

Figure 2: $h_5(\epsilon)$.

Check that $\lim_{q\to\infty} h_q(\epsilon) = \epsilon$.

Now we are ready to bound $|\mathsf{vol}(r)|$. Let X be a uniformly random element of $B(1^n, r)$. The entropy of X is exactly $\log|\mathsf{vol}(r)|$. We have $X = (X_1, X_2, \ldots, X_n)$. Let I be a uniformly random element of $[n]$, independent of X, and let $Y = X_I$ be the Ith coordinate of X. Then we have

$$\begin{aligned} H(X) &\leq H(X_1) + \cdots + H(X_n) \\ &= n \cdot H(Y|I) \\ &\leq n \cdot H(Y). \end{aligned}$$

by subadditivity

How large can $H(Y)$ be? If $\epsilon = p(Y \neq 1)$, then we have $\epsilon \leq r/n$. Let Z denote the random variable that indicates whether or not $Y = 1$.

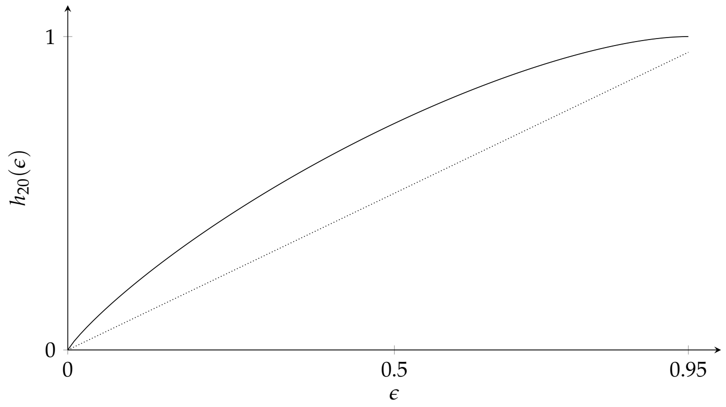

Figure 3: $h_{20}(\epsilon)$.

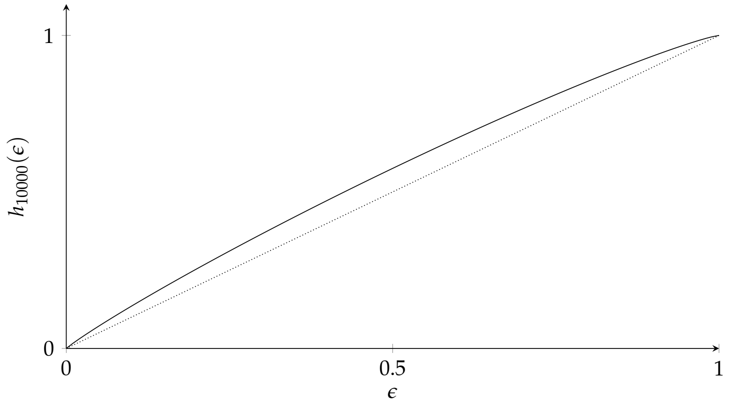

Figure 4: $h_{10000}(\epsilon)$. Although it looks like the two curves touch near $\epsilon = 1$, there is a gap between them. We always have $h_q(1 - 1/q) = 1$.

Using the fact that $H(Y|Z)$ is maximized when Y is uniform on the elements that are not 1, we get

$$\begin{aligned} H(Y) = H(YZ) = H(Z) + H(Y|Z) \\ &\leq (1-\epsilon)\log\left(\frac{1}{1-\epsilon}\right) + \epsilon\log\left(\frac{1}{\epsilon}\right) + \epsilon\log(q-1) \\ &= (1-\epsilon)\log\left(\frac{1}{1-\epsilon}\right) + \epsilon\log\left(\frac{q-1}{\epsilon}\right). \end{aligned}$$

If we set $h_q(\epsilon)$ to be

$$(1-\epsilon)\log_q\left(\frac{1}{1-\epsilon}\right) + \epsilon\log_q\left(\frac{q-1}{\epsilon}\right),$$

then we get that $H(Y) \leq \log q \cdot h_q(\epsilon)$. The derivative of $h_q(\epsilon)$ with respect to ϵ is

$$\begin{aligned} &\log_q(e) \cdot (\ln(1-\epsilon) + 1 + \ln\left(\frac{1}{\epsilon}\right) - 1 + \ln(q-1)) \\ &= \log_q(e) \cdot (\ln\left(\frac{1-\epsilon}{\epsilon} \cdot (q-1)\right)). \end{aligned}$$

This quantity is positive when $\epsilon < 1 - 1/q$, equal to 0 when $\epsilon = 1 - 1/q$, and negative otherwise. So, the maximum is achieved when $\epsilon = r/n$.

To prove the other side of the inequality, note that Stirling's approximation gives:

The bound we are using here is: $\sqrt{2\pi} \cdot n^{n+1/2} \cdot e^{-n} \leq n! \leq e \cdot n^{n+1/2} \cdot e^{-n}$.

$$\begin{aligned} \mathsf{vol}(r) &\geq \binom{n}{r} \cdot (q-1)^r \\ &= \frac{n!}{r!(n-r)!} \cdot (q-1)^r \\ &\geq \sqrt{\frac{2\pi n}{e^4 r(n-r)}} \cdot q^{n \cdot h_q(r/n)} \\ &\geq q^{n \cdot h_q(r/n) - \log_q(10(r/n)\cdot\sqrt{n})}. \end{aligned}$$

so the upper bound is very close to the truth.

Finite Fields

We have been talking about the tradeoffs for the parameters of error correcting codes. An equally important topic is about how to find *efficiently* encodable and decodable codes. The way to do this is to leverage linear algebra. We shall insist that the codes we are using are linear subspaces of Σ^n. For this to make sense, we need to consider vector spaces of finite size. A *finite field* is a finite set endowed with the operations of addition, subtraction, mulitplication and division. Before we can discuss linear codes, we need to establish some facts about finite fields.

These notes of Keith Conrad are a great reference: `https://kconrad.math.uconn.edu/blurbs/galoistheory/finitefields.pdf`.

We begin our whirlwind tour of finite fields now. Here, I am going to gather some very basic facts to show you how finite fields can be constructed. The notes will move at a very fast pace, and I will often sketch arguments. If you are motivated, you can try to flesh out the sketchy arguments into proofs — this is a good way to learn the material. If you are comfortable taking the existence of finite fields and their properties on faith, it is reasonable to skip over the sketches.

One of the reasons I was first attracted to learning about coding theory is that it is a beautiful example of how deep math can have extremely practical and surprising applications. When finite fields were studied by Galois in the early 19th century, he could not possibly have imagined the far reaching consequences he would have had on communications and data storage (and many other fields) much later.

The simplest example of a finite field is the set $\mathbb{F}_p$ for a prime p: it is the set of integers modulo a prime number p. In other words, it is the set $\{0, 1, 2, \ldots, p-1\}$ where all operations are carried out modulo p. It is easy to see how to add, subtract and multiply elements modulo p. Division is a little more tricky (and the only one that requires p to be prime).

We start with some basic facts about integers and polynomials with coefficients coming from any field. Try to prove these facts using Euclid's gcd algorithm and the division algorithm:

- If a, b are two positive integers, then you can always express $a = bq + r$, where q, r are integers, and $r < b$.
- If $a(X), b(X)$ are two polynomials, then you can always express $a(X) = b(X)u(X) + r(X)$, where $u(X), r(X)$ are polynomials and the degree of $r(X)$ is less than the degree of $b(X)$.
- If a, b are two positive integers, then you can always express the greatest common divisor d of a, b as $d = au + bv$, for some integers u, v.
- If $a(X), b(X)$ are two polynomials, you can always express the common diviosor $d(X)$ of $a(X), b(X)$ of largest degree as $d(X) = a(X)u(X) + b(X)v(X)$, for some polynomials $u(X), v(X)$.

Now, suppose a is an integer and p is a prime. Then, the above facts imply that either $a = 0 \mod p$, or $1 = au + pv$ for some integers u, v. In the second case, we have $au = 1 \mod p$, so u is the inverse of a in the finite field. This shows that every non-zero element $a \in \mathbb{F}_p$ has a multiplicative inverse a^{-1}.

Similarly, suppose $f(X) \in \mathbb{F}[X]$ is a polynomial of degree k with coefficients coming from some field $\mathbb{F}$. Suppose further that $f(X)$ is *irreducible* — it cannot be factored into two polynomials of smaller degree. Then the gcd algorithm establishes that every polynomial $a(X)$ is either divisible by $f(X)$, or we have $1 = a(X)u(X) + f(X)v(X)$, so $a(X)u(X) = 1 \mod f(X)$. Thus, $u(x)$ is the inverse of

$a(x)$ modulo $f(X)$, and the set of polynomials modulo $f(X)$ is itself a field! If $\mathbb{F} = \mathbb{F}_p$, then we obtain a finite field of size p^k — there are p^k possible remainder polynomials when you divide by $f(X)$.

In fact, the above two constructions capture *all possible* finite fields. Upto renaming the elements, there is a unique finite field of size p^k, for every prime power p^k.

The fact that finite fields are unique is, as far as I know, a fact that requires significant machinery from Galois theory. It is however, easier to see that there is only one finite field of size p for prime p.

Many of the definitions you may have learnt in linear algebra carry over to finite fields. For example, vectors are linearly independent or not, and every subspace has a dimension, with a basis of vectors etc etc. One thing to be careful about is that the idea of orthogonality does not have the same meaning over finite fields. For example, there are many vectors $x \in \mathbb{F}_2^n$ such that $\langle x, x \rangle = \sum_{i=1}^n x_i^2 = 0$, something that does not happen over the reals. That said, it does make sense to define the orthogonal complement of a subspace $V \subseteq \mathbb{F}^n$ as $V^\perp = \{x \in \mathbb{F}^n : \forall y \in V, \langle x, y \rangle = 0\}$. You can prove:

Fact 1. *The dimension of V plus the dimension of $V^\perp$ is always exactly n.*

If you take a non-zero element $\gamma \in \mathbb{F}$ in a finite field and start powering it $\gamma, \gamma^2, \gamma^3, \ldots$, you must eventually get to $\gamma^h = 1$. By considering the map that takes every non-zero element of the field α to $\gamma\alpha$, we see that this map must partition the non-zero elements into disjoint sets $S_1, S_2, \ldots, S_r$, where $\gamma \cdot S_i = S_i$ for all i, and $|S_i| = h$ for all i. This proves that h must divide $|\mathbb{F}| - 1$. The above discussion implies that:

Fact 2. *If $\mathbb{F}$ is a field of size q, then the polynomial $X^q - X$ evaluates to 0 on every element of the field.*

Another useful fact:

Fact 3. *If $f(X) \in \mathbb{F}[X]$ is a polynomial and $f(\alpha) = 0$, then $X - \alpha$ divides $f(X)$.*

Proof. We can always write $f(X) = (X - \alpha)u(X) + r(X)$, where $r(X)$ is a polynomial of degree 0, namely a constant. But since $f(\alpha) = 0$, we see that $r(\alpha) = 0$, so $r = 0$. □

As a consequence, we get

Fact 4. *If $f(X) \in \mathbb{F}[X]$ is a non-zero polynomial of degree d, then it has at most d roots.*

Proof. If $\alpha_1, \ldots, \alpha_{d+1}$ are distinct roots of f, then $f(X)$ must be divisible by $(X - \alpha_1)(X - \alpha_2) \ldots (X - \alpha_{d+1})$. So, the degree of f must be at least $d + 1$, a contradiction. □

Note that it is not necessary that f has even a single root in $\mathbb{F}$. However, you can show that there is always a larger field containing $\mathbb{F}$ where f has d roots (possibly counted with multiplicity).

Fact 5. *If $\mathbb{F}$ is a field of size q, then $X^q - X = \prod_{\alpha \in \mathbb{F}} (X - \alpha)$.*

Finally, an non-trivial fact that is extremely useful:

Fact 6. *Given any finite field of size q, there is an element γ such that $\gamma, \gamma^2, \ldots, \gamma^{q-1} = 1$ are all the non-zero elements of the field.*