**6.895 Essential Coding Theory** September 20, 2004

# Lecture 4

*Lecturer: Madhu Sudan* *Scribe: Victor Chen*

## 1 Overview

In this lecture, we shall examine the asymptotic performance of codes, prove some simple negative results concerning the parameters of codes, and analyze random and greedy codes.

## 2 Parameters of a Code

Let us recall the main parameters of a code from Lecture 3. For a code $C = (n, k, d)_q$, $n$ denotes the blocklength, $k$ denotes the message or information length, $d$ denotes the minimum distance of $C$, and $q$ denotes the alphabet size. When $C$ is linear, we use a matrix like notation using square brackets and write $C = [n, k, d]_q$.

In practice, it is possible that the message can be processed by a streaming encoding scheme. However, in this course, we can always assume that the message length and blocklength are fixed.

To study the asymptotics of a code, we normalize the parameters.

**Definition 1** *For an infinite family of code $C = \{(n_i, k_i, d_i)_q\}_{i=1}^{\infty}$, the rate of $C$ is defined to be $R = \lim \inf_{i\to\infty}\{\frac{k_i}{n_i}\}$ and the relative distance is defined to be $\delta = \lim \inf_{i\to\infty}\{\frac{d_i}{n_i}\}$.*

We use lim inf to guarantee that these two limits exist. Also, note that $q_i$ does not depend on $n_i$ in this definition. However, in certain cases, it is advantageous to construct an intermediary code where $q_i$ actually depends on $n_i$.

## 3 Some Negative Results

We now examine some negative results. The first we consider is the Singleton bound, due to R. C. Singleton. It perhaps is more appropriate to call it the projection bound as the following simple proof illustrates.

**Theorem 1 (Singleton Bound)** *For a code $C = (n, k, d)_q$, $d \le n - k + 1$. Asymptotically, $R + \delta \le 1$.*

**Proof** $C$ has $q^k$ codewords, and we project all codewords of $C$ onto the first $k-1$ coordinates. By the Pigeonhole Principle, two codewords must have the same projection since the total number of possible projections is $q^{k-1}$. These two codewords must agree on the first $k-1$ coordinates, and hence, they differ in at most $n - (k-1)$ coordinates. ■

One may think that a more careful analysis may yield a tighter bound, but we shall see codes that meet the Singleton Bound in a later lecture. However, note that $q$ does not play a role in this bound. To bring $q$ into the picture, we re-examine the Hamming bound studied in a previous lecture. Recall that a code $C = (n, k, d)_q$ is $(d-1)/2$ error correcting. So balls of radius $(d-1)/2$ around each codewords in the space $\Sigma^n$ do not overlap. Define $\mathrm{Vol}(r, n)$ to be the number of points in a ball of radius $r$ in $\Sigma^n$. Then clearly $q^k \cdot Vol_q(\frac{d-1}{2}, n) \le q^n$. Consider the binary case when $q = 2$:

$$\begin{aligned} 2^k \cdot Vol_2(\frac{d-1}{2}, n) &\le 2^n, \\ 2^k \cdot 2^{H_2(\frac{d-1}{2n})n} &\approx\le 2^n, \end{aligned}$$

$$\begin{aligned} k + H_2(\frac{d-1}{2n})n &\approx\leq n + o(n) \\ R + H_2(\delta/2) &\leq 1, \end{aligned}$$

where the second line follows for $p \leq 1/2$ and Stirling's formula for factorial. Hence, we have the following binary Hamming bound (also known as the volume or packing bound):

**Theorem 2 (Hamming Bound)** *For an infinite family of binary code with rate $R$ and relative distance $\delta$, $R + H_2(\delta/2) \leq 1$.*

Now we examine the $q$-ary Hamming bound $q^k \cdot Vol_q(\frac{d-1}{2}, n) \leq q^n$ again. First observe that $Vol_q(pn, n) = \sum_{i=0}^{pn} \binom{n}{i}(q-1)^i$, which is dominated by $\binom{n}{pn}(q-1)^{pn}$ for $0 < p < 1/2$. (Madhu conjectured in class that this is also true for $0 < p < (q-1)/q$.) Define $H_q(p) = -p \log_q p - (1-p) \log_q (1-p) + p \log_q (q-1)$, which is 1 at $p = 1 - 1/q$. (See the below figure for its graph.) Taking log and dividing both sides by $n$, we obtain

$$R + H_q(\delta/2) \leq 1.$$

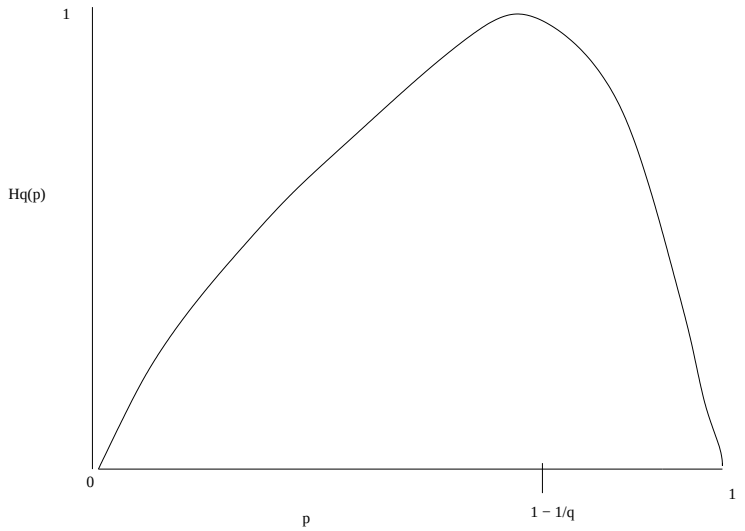


**Figure 1**: the entropy function $H_q(p)$

Note that when $q = 2$, the Hamming bound strictly dominates the Singleton bound. For small values of $q$, the Hamming bound intersects the Singleton bound at $\delta$ close to 1. For large enough $q$, the Singleton bound dominates the Hamming bound.

Consider the following question: can we have three binary codewords of length $n$ with pairwise difference of $0.9n$? A simple analysis shows that this is impossible. However, neither the Singleton bound nor the binary Hamming bound rules this case out. So we seek a tighter bound.

# 4 Random Code

Suppose we pick codewords $c_1, \ldots, c_k$ at random from $\{0,1\}^n$. We want the minimum distance of these codewords to be $d = \delta n$. How large can $k$ be?

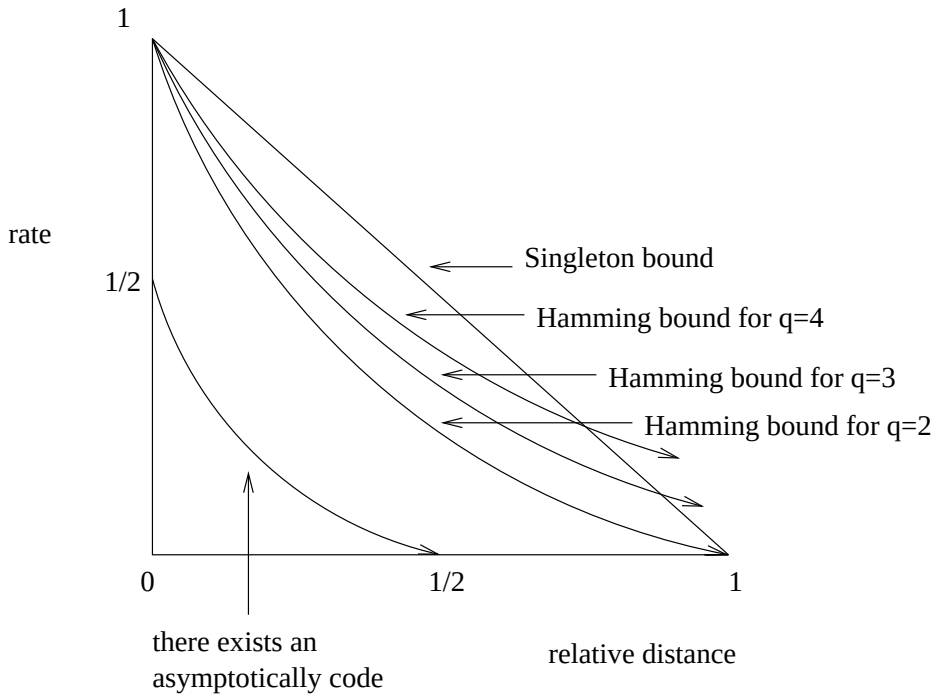


**Figure 2**: $R$ versus $\delta$

We use the probabilistic method. Suppose we are picking $c_i$ now. Let $E_i$ be the bad event that there is some $j < i$ such that $\Delta(c_i, c_j) < d$. Then

$$\begin{aligned}
\Pr[\cup_{i=1}^{k} E_i] &\leq \sum_{i=0}^{k} \Pr[E_i] \\
&\leq \sum_{i=0}^{k} \frac{i \cdot 2^{H(\delta)n}}{2^n} \\
&= \binom{k}{2} 2^{(H(\delta)-1)n} \\
&= 2^{(2R+H(\delta)-1)n},
\end{aligned}$$

by writing $k = 2^{Rn}$. For any $2R + H_2(\delta) < 1$, the above probability is strictly less than 1. Hence, there is a nonzero probability that none of the bad events occur, i.e., all codewords picked are far from each other. Hence, there exists a binary code with rate $R$ and relative distance $\delta$ for any $2R+H_2(\delta) < 1$.

The union bound gives only a crude estimate, and we would like to remove the factor of 2 in front of $R$. We assumed in our analysis that no pairwise codewords are at distance less than $d$. However, even if a few number of codewords are too close, we can delete these bad codewords and still hope to maintain a minimum distance of $d$ and large $k$. For a simple analysis, suppose $k$ is fixed such that $E_k \leq 1/10$. Let $X_i$ be 1 if $E_i$ occurs and 0 otherwise. Then by Markov's Inequality,

$$\Pr[\sum_{i=1}^{k} X_i \geq k/2] \leq \frac{E[\sum_{i=1}^{k} X_i]}{k/2} \leq 1/5.$$

So the probability that many bad events occur is small. We will use this idea of deleting words in the next section.

# 5 Greedy/Maximal Code

A code with minimum distance is maximal if no more codeword can be added while maintaining minimum distance $d$.

**Claim 3** *For every maximal code* $C = (n, k, d)_2$, $2^k \cdot Vol_2(d-1, n) \geq 2^n$.

**Sketch of Proof** Consider a greedy algorithm (may run in exponential time) that adds a codeword and deletes a ball or radius $d-1$ around it. The minimum distance will be maintained at each step, and the algorithm stops until no more possible codeword can be picked. When the algorithm stops, each vector in $\{0,1\}^n$ was picked or deleted within a region, with volume at most $Vol_2(n, d-1)$. The claim then follows. ■