CS 880: Advanced Complexity Theory 3/1/06

Lecture 14: Coding Theory

Instructor: Dieter van Melkebeek Scribe: Michael Kowalczyk

This lecture comprises a review of coding theory, which will be necessary background for later lectures. Relevant definitions are introduced, Shannon's noisy and noiseless coding theorems are presented, and basic facts about data compression and error correcting codes are reviewed. Some upper bounds and lower bounds are also discussed.

1 Information Transmission Model

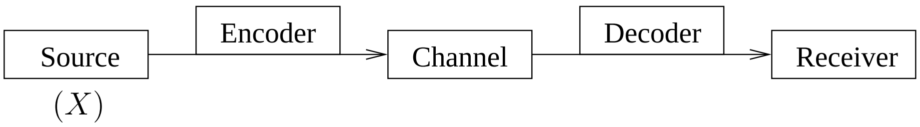

Figure 1: Transmission of information

Figure 1 illustrates the basic problem of information transmission. In general, we imagine some source which sends messages to a receiver. The source encodes a message into a codeword which is subsequently transmitted over a channel, and finally decoded and consumed by the receiver. Of course, we want the receiever to get the same message that was sent, but we list three other properties which are desirable for such a system.

1. The transmission should be kept as small as possible.
2. The system should be robust with respect to noise in the channel.
3. The encoding and decoding transformations should be efficient.

This is a very general model and has many applications. For instance, the channel could be a physical transmission line, computer memory, or a spacecraft uplink. Also note that in some applications, it may be acceptable if the receiver gets something close to the original message, such as in JPEG compression for digital images.

2 Data Compression

When we completely disregard the goal of robustness, we are left with the problem of data compression. We denote the source as a probability distribution X on length k strings over alphabet Σ. We can consider the problem of optimizing the worst case over X, but this is trivially accomplished using $\lg_2 |\Sigma^k|$ bits. We can do better by exploiting additional properties of X.

We start with the simplest setting for X which is that of a memoryless source. By this we mean that each of the k symbols is drawn independently from some fixed distribution, X'. Shannon's noiseless coding theorem renders this completely analyzed. In particular, the best possible compression factor is asymptotically $H(X')$ where $X = (X')^k$ and H denotes information entropy.

Example: If we take X' to be the uniform distribution on $\{0, 1\}$, we have $H(X') = 1$ and no compression can be achieved. If $X' = 0$ then we have $H(X') = 0$, as no information is being sent. ⊠

The bound given in Shannon's theorem is tight, as it can be realized with an arithmetic encoding. This strategy works by associating each message with an interval of size proportional to its probability under X. Then messages with high probability can be represented by rationals with short description length that fall within the interval. Note that restricting this strategy to work locally (under X') results in Huffman encoding, which is optimal for single symbol encoding. In practice there are codes such as Lempel-Ziv which exploit dependence to get even better performance in the non-memoryless setting.

3 Error Correcting Codes

3.1 Definitions and Parameters

The other extreme, where we focus entirely on robustness and impose no requirement with regards to message size, is not interesting since we can just repeat the message infinitely often. We want to strike a balance between redundancy and error correction ability.

Our performance metric will be the worst case for the source. In other words, we want to optimize performance for the worst case over all possible messages. We can think of the original message already being in compressed form so that no errors should reach the receiever after decoding. We will be looking at codes of the form $\Sigma^k \mapsto \Sigma^n$ where $\Sigma = \{0, 1\}$. The parameter n is called the block length.

Again, we can consider many models for the channel, and the memoryless model is simplest. This simply means that the channel transforms each input symbol according to a fixed probability distribution. A special case is the binary symmetric channel with parameter p, which works over the alphabet $\Sigma = \{0, 1\}$ and flips each bit with probablity p, irrespective to its initial value. Shannon's noisy theorem tells us how good we can do in this setting. The theorem is stated in terms of the rate, which is inuitively how much initial message data is contained in each bit after encoding.

Definition 1 (Rate)**.** *A code $f : \Sigma^k \mapsto \Sigma^n$ has rate k/n.*

Theorem 1 (Shannon's Noisy Coding Theorem)**.** *For any channel, there exists a constant c (called the capacity) such that for any $r < c$, there exists a family of codes with rate at least r such that for any message x,* $\Pr[\text{Decoding}(\text{Encoding}(x)) \neq x] \leq 2^{-\Omega(n)}]$ *for block length n. Furthermore, given $r > c$ and any family of codes with rate at least r, there exists an x such that* $\Pr[\text{Decoding}(\text{Encoding}(x)) = x] \leq 2^{-\Omega(n)}]$ *for block length n.*

Example: Given a binary symmetric channel with parameter p, $c = 1 - H(p)$ where H is information entropy. This special case of information entropy is known as binary entropy, and $H(p) = p \lg_2(\frac{1}{p}) + q \lg_2(\frac{1}{q})$ where $q = 1 - p$. ⊠

Another model for the channel is a worst case version with parameter t. In this case, the channel can introduce any error pattern in which at most t bits are changed. Since we want to be able to correct all errors, the errors can be viewed as being choosen adversarily.

Definition 2 (Relative distance)**.** *The distance, d, of a code $f : \Sigma^k \mapsto \Sigma^n$ is the minimum Hamming distance over all distinct codewords. The relative distance is defined as d/n.*

Note that a code can correct up to t errors if and only if it has distance $d \geq 2t + 1$. Similarly, it can detect up to t errors if and only if it has distance $d \geq t + 1$.

If a code $f : \Sigma^k \mapsto \Sigma^n$ has distance d, then we say f is an (n, k, d)-code. If in addition it is a linear code (f can be represented as a linear transformation), then we write it as an $[n, k, d]$-code. We would like to realize codes with both rate $r = k/n$ and relative distance $\delta = d/n$ as large as possible.

Linear codes have some desirable properties. Specifically, the description length is small as we only need to write down the transformation matrix, which is nk bits. In contrast, an arbitrary code might require $n \cdot 2^k$ bits to describe it. Furthermore, the encoding step for any linear code can be realized in quadratic time. Decoding a general code, on the other hand, is $\mathsf{NP\text{-}hard}$. As we will see in later lectures, expander codes yield an explicit family of codes with constant r and δ where both encoding and decoding are linear time operations.

Much study has been directed towards tradeoffs between δ and r. We will study these parameters asymptotically, which provokes the following definition.

Definition 3. *Given a family of codes, $\mathcal{C}$, we denote*

$$\begin{aligned} r(\mathcal{C}) &= \liminf_{n\to\infty} r(\mathcal{C}_n) \\ \delta(\mathcal{C}) &= \liminf_{n\to\infty} \delta(\mathcal{C}_n) \end{aligned}$$

where n is the block length and $\mathcal{C}_n$ is the code with block length n.

Since we are interested in the relationship between δ and r, we define

$$R(\delta) = \sup(\{r \mid \text{there exists a family } \mathcal{C} \text{ with } r(\mathcal{C}) \geq r \text{ and } \delta(\mathcal{C}) \geq \delta\})$$

as a way of measuring the relationship.

3.2 Upper and Lower Bounds

The Gilbert-Varshamov upper bound gives us $R(\delta) \geq 1 - H(\delta)$.

Theorem 2. *For sufficiently large block size, a random linear code has rate $r \geq 1 - H(\delta)$ with high probability.*

Proof. First, we observe that the distance of a code is equal to the minimum weight over all nonzero codewords. So it suffices to show that it is unlikely that a random code will have some codeword with small Hamming weight. Denote Hamming weight of x by $h(x)$.

Pick $G \in \{0,1\}^{n\times k}$ uniformly at random. For any fixed $x \neq 0$ in $\{0,1\}^k$, Gx is uniformly distributed over $\{0,1\}^n$. Applying the union bound to this, we get

$$\begin{aligned}
\Pr[(\exists x \neq 0)(h(Gx) \leq \delta n)] &\leq 2^k \sum_{i=0}^{\delta n} \frac{\binom{n}{i}}{2^n} \\
&\leq \frac{2^k}{2^n} \cdot 2^{H(\delta)\cdot n} \\
&= 2^{(H(\delta)+r-1)\cdot n}.
\end{aligned}$$

This inequality $\sum_{i=0}^{\delta n} \binom{n}{i} \leq 2^{H(\delta)\cdot n}$ is asympototically tight and left as an exercise. So for any $r < H(\delta) - 1$, the probability limits to zero, which yields the theorem. $\square$

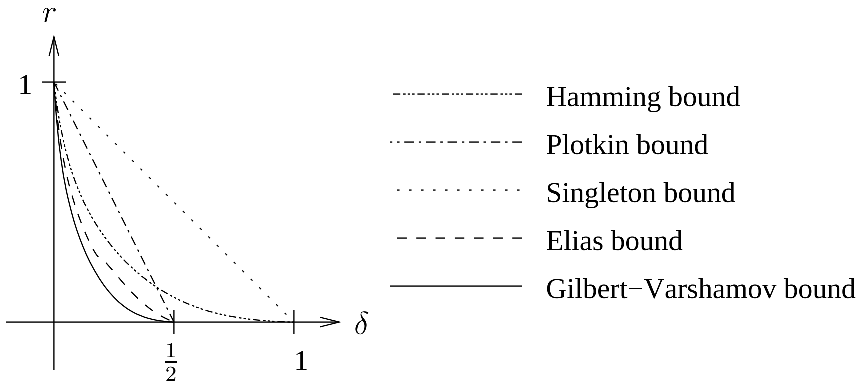

Figure 2: Upper and lower bounds on $R(\delta)$

We illustrate the singleton bound, Hamming bound, Plotkin bound, and Elias bound in Figure 3.2. These will be covered next lecture.