IERG6120 Coding for Distributed Storage Systems Lecture 1 - 08/09/2016

Hamming Distance

Lecturer: Kenneth Shum *Scribe: Qiaoqiao Zhou*

We first review some basic materials in coding theory.
According to Shannon, a communication system can be described as

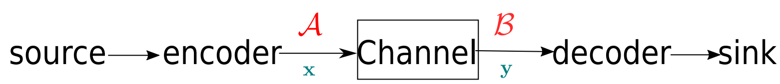

Figure 1: communication system

Discrete Channel:
Channel input alphabet $\mathcal{A} = \{a_1, \dots, a_q\}$
Channel output alphabet $\mathcal{B} = \{b_1, \dots, b_m\}$
Transition probability

$$\Pr(b_j \text{ received} \mid a_i \text{ sent}), \text{ for } i = 1, \dots, q, \quad j = 1, \dots, m. \tag{1}$$

Memoryless:
Consider using the channel n times
The symbols transmitted $\mathbf{x} = (x_1, x_2, \dots, x_n)$
The symbols received $\mathbf{y} = (y_1, y_2, \dots, y_n)$
The transition probability satisfies

$$\Pr(\mathbf{y} \mid \mathbf{x}) = \prod_{t=1}^{n} \Pr(y_t \text{ received} \mid x_t \text{ sent}) \tag{2}$$

Binary symmetric channel (BSC):
$\mathcal{A} = \{0, 1\} = \mathcal{B}$

$$\begin{aligned}
\Pr(1 \text{ received} \mid 1 \text{ transmitted}) &= 1 - \epsilon. \\
\Pr(0 \text{ received} \mid 0 \text{ transmitted}) &= 1 - \epsilon. \\
\Pr(1 \text{ received} \mid 0 \text{ transmitted}) &= \epsilon. \\
\Pr(0 \text{ received} \mid 1 \text{ transmitted}) &= \epsilon.
\end{aligned} \tag{3}$$

Here, ϵ is called the *crossover probability*.

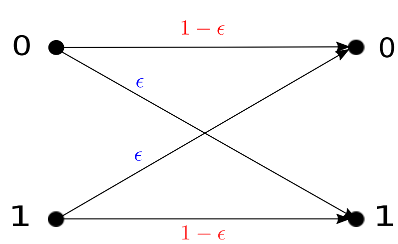

Figure 2: BSC channel

q-ary symmetric channel:
$\mathcal{A} = \{0, 1, \dots, q-1\} = \mathcal{B}$. See Fig. 3.

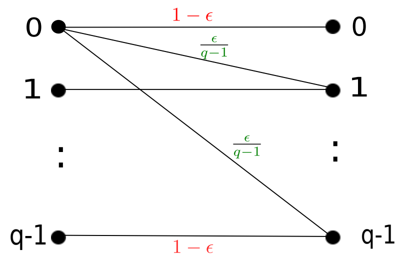

Figure 3: q-ary symmetric channel

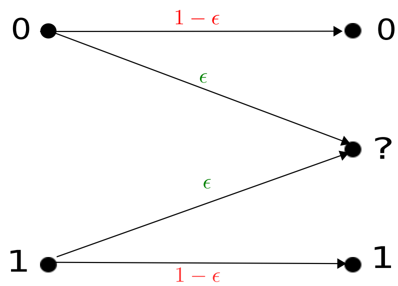

Figure 4: Binary erasure channel

Binary erasure channel (BEC):
$\mathcal{A} = \{0, 1\}, \quad \mathcal{B} = \{0, 1, ?\}$ The transition probabilities are illustrated in Fig. 4.

A *block code*, $\mathcal{C}$, of length n, with alphabet $\mathcal{A} = \{a_1, a_2, \ldots, a_q\}$ is a non-empty collection of $\mathcal{A}^n$. The elements of $\mathcal{C}$ are called the *codewords*.

Example 1 : Repetition code: $\mathcal{A} = \{0, 1\}$, $\mathcal{C} = \{00000, 11111\}$
Encoder :

$$0 \to 00000$$
$$1 \to 11111$$

Majority-vote decoder:

$$\begin{aligned}
00000 &\to 0\\
00001 &\to 0\\
00010 &\to 0\\
00100 &\to 0\\
01000 &\to 0\\
10000 &\to 0\\
\underbrace{\ldots}_{2\,1's} &\to 0\\
\underbrace{\ldots}_{3\,1's} &\to 1\\
\underbrace{\ldots}_{4\,1's} &\to 1\\
\underbrace{\ldots}_{5\,1's} &\to 1
\end{aligned}$$

$$\begin{aligned}
&\Pr(\text{error} \mid 0\,\text{sent})\\
&= \Pr(3\,\text{or more}\,1's\,\text{in the received vector})\\
&= \binom{5}{3}\epsilon^3(1-\epsilon)^2 + \binom{5}{4}\epsilon^4(1-\epsilon)^1 + \binom{5}{5}\epsilon^5 \qquad (4)
\end{aligned}$$

Maximal likelihood decoder:

In this lecture, a "decoding error" means block error, i.e., $\mathrm{Dec}(\mathbf{y}) \neq \mathbf{c}$ but $\mathbf{c}$ is sent. Assume $\Pr(\mathbf{c} \text{ is sent}) = \frac{1}{M}$, $\forall \mathbf{c} \in \mathcal{C}$, where $M = |\mathcal{C}|$ is the size of the code.

$$\begin{aligned}
&\Pr(\text{correct decoding})\\
&= \sum_{\mathbf{y}\in\mathcal{A}^n} \Pr(\text{correct decoding} \mid \mathbf{y} \text{ received})\Pr(\mathbf{y}\,\text{received})\\
&= \sum_{\mathbf{y}\in\mathcal{A}^n} \Pr(\mathrm{Dec}(\mathbf{y}) \text{ is sent} \mid \mathbf{y} \text{ received})\Pr(\mathbf{y} \text{ received}) \qquad (5)
\end{aligned}$$

Pick a decoder Dec such that

$$\begin{aligned}
&\Pr(\mathrm{Dec}(\mathbf{y}) \text{ is sent} \mid \mathbf{y} \text{ received})\\
&= \max_{\mathbf{c}\in\mathcal{C}} \Pr(\mathbf{c} \text{ is sent} \mid \mathbf{y} \text{ received}) \qquad (6)
\end{aligned}$$

According to Bayes' rule

$$\begin{aligned}
&\Pr(\mathbf{c}\,\text{is sent} \mid \mathbf{y} \text{ received})\\
&= \frac{\Pr(\mathbf{y} \text{ received} \mid \mathbf{c} \text{ is sent})\,\overbrace{\Pr(\mathbf{c}\,\text{sent})}^{=\frac{1}{M}}}{\sum_{\mathbf{c}'\in\mathcal{C}} \Pr(\mathbf{y} \text{ received} \mid \mathbf{c}' \text{ is sent})\,\underbrace{\Pr(\mathbf{c}'\,\text{sent})}_{=\frac{1}{M}}}\\
&= \frac{\Pr(\mathbf{y} \text{ received} \mid \mathbf{c} \text{ is sent})}{\sum_{\mathbf{c}'\in\mathcal{C}} \Pr(\mathbf{y} \text{ received} \mid \mathbf{c}' \text{ is sent})} \qquad (7)
\end{aligned}$$

Hence, the maximization of $\Pr(\mathbf{c}\,\text{is sent} \mid \mathbf{y}\,\text{received})$ is equivalent to the maximization of $\Pr(\mathbf{y}\,\text{received} \mid \mathbf{c}\ \text{is sent})$, provided that the codewords are transmitted with equal probability. We have thus proved the following

Theorem 1. ***(ML decoding):*** *Suppose that the codewords in a code $\mathcal{C}$ are transmitted with the same probability. If we choose the decoder Dec such that*

$$\mathrm{Dec}_{\mathrm{ML}}(\mathbf{y}) = \operatorname*{argmax}_{\mathbf{c}\in\mathcal{C}} \Pr(\mathbf{y}\ \textit{received} \mid \mathbf{c}\ \textit{is sent}), \tag{8}$$

with ties broken arbitrarily, then the probability of error is minimized.

Example 2: Consider the BSC with repetition code $\mathcal{C} = \{00000, 11111\}$. Suppose 11000 is received.

$$\begin{aligned} \Pr(11000\,\text{received} \mid 00000\,\text{sent}) &= \epsilon^2(1-\epsilon)^3 \\ \Pr(11000\,\text{received} \mid 11111\,\text{sent}) &= \epsilon^3(1-\epsilon)^2 \end{aligned} \tag{9}$$

For $\epsilon < \frac{1}{2}$, we have $\epsilon^2(1-\epsilon)^3 > \epsilon^3(1-\epsilon)^2$. Therefore, $\mathrm{Dec}_{\mathrm{ML}}(11000) = 00000$.

Now, consider the q-ary symmetric channel, i.e., $\mathcal{A} = \mathcal{B} = \{0, \ldots, q-1\}$, with block length n.

$$\begin{aligned} \Pr(\text{no error}) &= (1-\epsilon)^n \\ \Pr(\text{error at the } i^{th}\ \text{location}) &= \epsilon(1-\epsilon)^{n-1} \\ \Pr(\text{error at the } i^{th} \text{and}\, j^{th}\ \text{location}) &= \epsilon^2(1-\epsilon)^{n-2} \end{aligned} \tag{10}$$

We can see that the probability of error is independent of the exactly error location, it only depends on the number of errors.

Definition 2. *Let $\mathbf{u}$ and $\mathbf{v}$ be n-tuples in $\mathcal{A}^n$. Define the* **Hamming distance** *between $\mathbf{u}$ and $\mathbf{v}$ as the number of locations in which $\mathbf{u}$ and $\mathbf{v}$ are different,*

$$d_H(\mathbf{u}, \mathbf{v}) := \Big|\{i \colon u_i \neq v_i\}\Big| \tag{11}$$

where $\mathbf{u} = (u_1, \ldots, u_n)$ and $\mathbf{v} = (v_1, \ldots, v_n)$.

Theorem 3. *For q-ary symmetric channel with $\epsilon \ll 1$, the ML decoder outputs the codeword $\mathbf{c}$ such that $d_H(\mathbf{c}, \mathbf{y})$ is minimized.*

When the probability of channel error is sufficiently small, the ML decoder is the same as the *nearest-neighbor decoder*

$$\mathrm{Dec}_{\mathrm{NN}}(\mathbf{y}) = \operatorname*{argmax}_{\mathbf{c}\in\mathcal{C}} d_H(\mathbf{y}, \mathbf{c}).$$

Example 3: $\mathcal{C} = \{11111, 11000, 00110, 00001\}$. The received vector is $\mathbf{y} = 01000$. Then,

$$\begin{aligned} d_H(11111, 01000) &= 4 \\ d_H(11000, 01000) &= 1 \\ d_H(00110, 01000) &= 3 \\ d_H(00001, 01000) &= 2 \end{aligned}$$

Therefore, according to Theorem 3, the ML decoder will decode the received vector to 11000.

Properties of Hamming distance:

$$d_H(\mathbf{u},\mathbf{v}) \geq 0 \quad \text{with equality iff } \mathbf{u}=\mathbf{v} \tag{12}$$
$$d_H(\mathbf{u},\mathbf{v}) = d_H(\mathbf{v},\mathbf{u}) \tag{13}$$
$$d_H(\mathbf{u},\mathbf{v}) \leq d_H(\mathbf{u},\mathbf{w}) + d_H(\mathbf{w},\mathbf{v}) \tag{14}$$

Hamming sphere: Given $\mathbf{u} \in \mathcal{A}$ and $r \geq 0$, define the *Hamming sphere* with radius r and center $\mathbf{u}$ as the set

$$B(\mathbf{u}, r) := \{\mathbf{v} \in \mathcal{A}^n : d_H(\mathbf{u},\mathbf{v}) \leq r\}. \tag{15}$$

Theorem 4. *A code $\mathcal{C}$ can correct t errors under nearest-neighbor decoding iff $B(\mathbf{c}, t)$ for all $\mathbf{c} \in \mathcal{C}$ are disjoint.*

Proof ($\Leftarrow$) Assume that $B(\mathbf{c}, t)$ and $B(\mathbf{c}', t)$ are disjoint for $\mathbf{c}, \mathbf{c}' \in \mathcal{C}$ and $\mathbf{c} \neq \mathbf{c}'$. Suppose $\mathbf{c} \in \mathcal{C}$ is sent, and the channel introduces no more than t errors. The received vector $\mathbf{y}$ satisfies $d_H(\mathbf{c}, \mathbf{y}) \leq t$. Hence $\mathbf{y} \in B(\mathbf{c}, t)$ by the definition of Hamming sphere. Consider a codeword $\mathbf{c}'$ which is not equal to $\mathbf{c}$. Since the Hamming spheres $B(\mathbf{c}, t)$ and $B(\mathbf{c}', t)$ are disjoint, we have

$$\mathbf{y} \notin B(\mathbf{c}', t),$$

which means that $d_H(\mathbf{y}, \mathbf{c}') > t$. As this is true for all $\mathbf{c}' \neq \mathbf{c}$, we get $d_H(\mathbf{y}, \mathbf{c}) = \min_{\mathbf{c}' \in \mathcal{C}} d_H(\mathbf{y}, \mathbf{c}')$. The codeword $\mathbf{c}$ is outputted by the nearest-neighbor decoder correctly.

($\Rightarrow$) Suppose that there are two distinct codewords $\mathbf{c}$ and $\mathbf{c}'$ such that $B(\mathbf{c}, t)$ and $B(\mathbf{c}', t)$ are not disjoint. Let $\mathbf{y}$ be a vector in the intersection of $B(\mathbf{c}, t)$ and $B(\mathbf{c}', t)$, i.e., $d_H(\mathbf{y}, \mathbf{c}) \leq t$ and $d_H(\mathbf{y}, \mathbf{c}') \leq t$. We consider three cases.
Case 1, $\mathrm{Dec}_{\mathrm{NN}}(\mathbf{y}) = \mathbf{c}$.
Case 2, $\mathrm{Dec}_{\mathrm{NN}}(\mathbf{y}) = \mathbf{c}'$.
Case 3, $\mathrm{Dec}_{\mathrm{NN}}(\mathbf{y})$ is not equal to $\mathbf{c}$ or $\mathbf{c}'$.

In case 1, we have a decoding error if $\mathbf{c}'$ is transmitted and $\mathbf{y}$ is received. In case 2, we have a decoding error if $\mathbf{c}$ is transmitted and $\mathbf{y}$ is received. In case 3, we have a decoding error if $\mathbf{y}$ is received but the transmitted codeword is $\mathbf{c}$ or $\mathbf{c}'$. In all three cases, an erroneous codeword is returned by the decoder even though the number of channel errors is no more than t. □

Exercise: Show that the code in Example 3 with block length 5 can correct 1 error. Find all 5-tuples in $\{0,1\}^5$ that are at Hamming distance at least 2 from all codewords.