ECE 587 / STA 563: Lecture 6 – Channel Coding

Information Theory
Duke University, Fall 2023

Author: Galen Reeves
Last Modified: August 24, 2023

Outline of lecture:

6.1 Introduction to Channel Coding

6.1.1 Motivation

- Communication problems play an important role in many applications:
  - Wireless communication between your laptop and a router
  - Wireless communication between your cell phone and a cell tower (e.g, GSM and CDMA)
  - Fiber-optic communication across the internet
  - Data storage on a hard drive (communicating from the past to the future).
  - Communication between Hubble Space Telescope and Earth.
  - Statistics: communicating the true state of nature θ through the observed data X
- The primary goals of communication are:

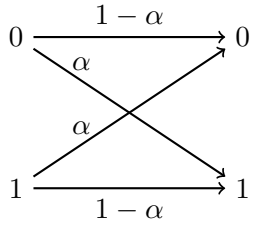
(1) **Reliability:** We want the received message to be equal to the transmitted message (at least with very high probability)

$$\text{Error probability:} \quad \mathbb{P}[\text{received message} \neq \text{transmitted message}]$$

(2) **Efficiency:** We want to communicate each message as quickly as possible. Equivalently, this means that given a set of amount of time, we want to communicate as many messages as possible.

$$\text{Rate:} \quad R = \text{average number of information bits sent per unit time}$$

- Unfortunately, these two goals are fundamentally opposed.
- **Example:** (Binary Symmetric Channel) Suppose we want to send a single bit $W \in \{0,1\}$. To communicate this bit, we can use binary-symmetric channel (BSC).
  - The BSC has binary input $X \in \{0,1\}$ and binary output $Y \in \{0,1\}$. The probability that Y is equal to X is $1-\alpha$ and the probability the the channel flips the bit is α.



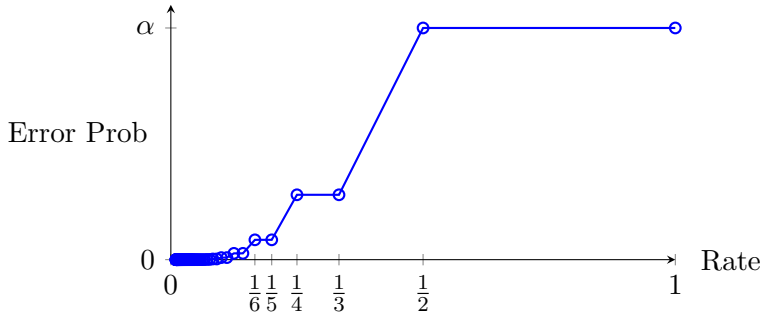
  - If both inputs are equally likely, the error probability for a single use of the channel equal to α.
  - To reduce the error probability, we use the channel multiple times. Each use of the channel consists of one unit of time. Assume the channel is memoryless, i.e., the channel outputs are conditionally independent given the inputs.
  - Encode using a repetition code:

$$\begin{aligned} W = 0 \quad &\Longrightarrow \quad \text{send: } X^n = \overbrace{000\cdots0}^{n} \\ W = 1 \quad &\Longrightarrow \quad \text{send: } X^n = \underbrace{111\cdots1}_{n} \end{aligned}$$

  - Given output $Y_1, Y_2, \cdots, Y_n$, decode using maximum likelihood (ML) decision rule:

$$\hat{W} = \begin{cases} 0, & \#\text{ observed '0's greater than } \#\text{ of '1's} \\ 1, & \#\text{ observed '1's greater than } \#\text{ of '0's} \end{cases}$$

  - The error probability after n channel uses obeys $P_e \geq \alpha^n$. This probability is small, but strictly greater than zero.
  - The rate after n channel uses is $R = 1/n$.
  - Tradeoff between error probability and rate for $\alpha = 0.1$

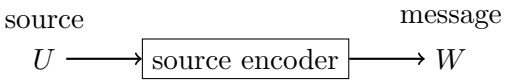


- The contributions of Claude Shannon:
  - Shannon's key insight was that one can overcome the tradeoff between reliability and efficiency by coding across multiple bits. It is possible to reduce the error probability *without reducing the rate*!
  - He chose to focus on information, ignoring computation, and found fundamental limits
  - Over the last 60 years, we have figured out how to achieve these limits. All modern communication systems are based on information theory.

6.1.2 Problem Setup

- Communication across a channel

  - The **message** $W \in \{1, 2, \cdots, M\}$ is one of M possible numbers that we want to communicate. This message is generated from a random source and is distributed uniformly over all possibilities.



  - An (M, n) **coding scheme** consists of an encoder (or codebook) $\mathcal{E}$ that maps the message W to an n-length sequence of channel inputs X^n:

$$\mathcal{E} : \{1, \cdots, M\} \to \mathcal{X}^n$$

    and a decoder that maps n-length sequences of channel outputs Y^n to an estimate $\hat{W}$ of the message.

$$\mathcal{D} : \mathcal{Y}^n \to \{1, 2, \cdots, M\}$$

  - The **channel** specifies the (probabilistic) transformation from inputs to outputs:

$$\mathbb{P}[Y^n = y^n | X^n = x^n] = p_{Y^n|X^n}(y^n|x^n)$$

    The channel is **memoryless** if the outputs between channel uses are conditionally independent given the input, i.e.,

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$$

- The **rate** R of an (M, n) coding scheme is defined as

$$R = \frac{\log_2 M}{n} \quad \text{bits/transmission}$$

  Alternatively, the number of messages for a given rate R and block-length n is given by

$$M = 2^{nR}$$

  To specify a rate R code, we write $(2^{nR}, n)$ instead of (M, n).

- The **conditional error probability** is defined by

$$P_e^{(n)}(w) = \mathbb{P}\Big[W \neq \hat{W} | W = w\Big]$$

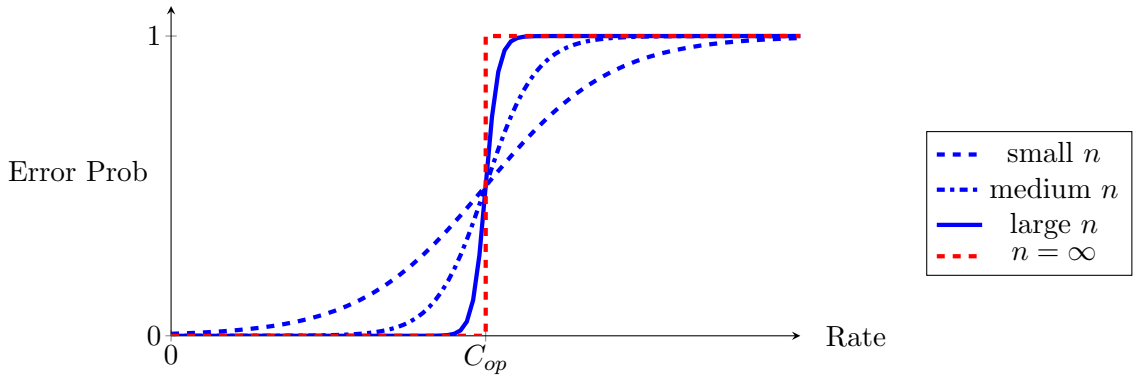
  The **average error probability** is

$$P_e^{(n)} = \mathbb{P}\Big[\hat{W} \neq W\Big] = \frac{1}{M} \sum_{w=1}^{M} P_e^{(n)}(w)$$

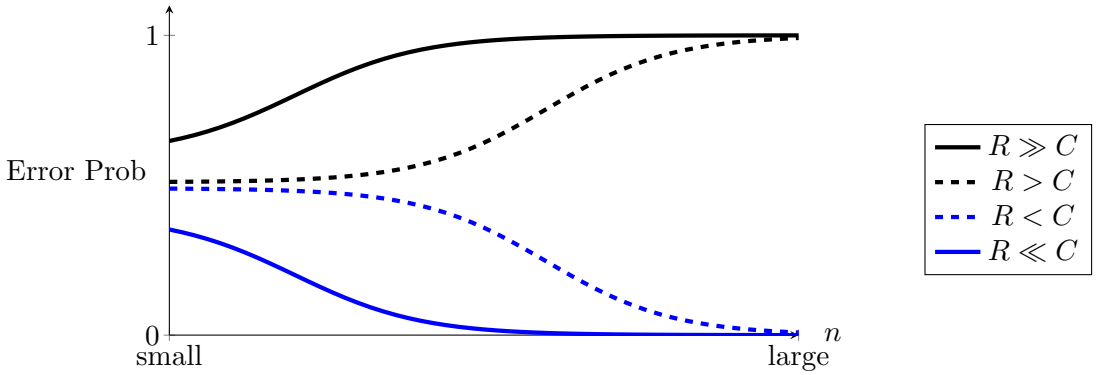
  The **maximum error probability** is

$$P_{\text{e,max}}^{(n)} = \max_{w \in \{1, \cdots, M\}} P_e^{(n)}(w)$$

- The key objective is to find an $(2^{nR}, n)$ coding scheme with $P_{\text{e,max}}^{(n)}$ as small as possible.
- Tradeoff between rate and error probability for fixed n



- Tradeoff between error probability and n for fixed rate



- **Definition:** A rate R is **achievable** for given discrete memoryless channel $p(y|x)$ if there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ coding schemes such that maximum error probability $P_{\text{e,max}}^{(n)}$ tends to zero as $n \to \infty$.
- The **operational capacity** C_{op} is the supremum over the set of achievable rates.

$$C_{\text{op}} = \sup\{R : R \text{ is achievable}\}$$

6.1.3 Channel Coding Theorem

- A **discrete memoryless channel** (DMC) is a channel specified by
  - an input alphabet $\mathcal{X}$; which is a set of symbols the channel accepts as input;
  - an output alphabet $\mathcal{Y}$; which is a set of symbols that the channel can produce at its output; and
  - a conditional probability distribution $p_{Y|X}(\cdot|x)$ for all $x \in \mathcal{X}$ such that the channel outputs are conditionally independent given the input, i.e.,

$$p_{Y^n|X^n}(y^n|x^n) = \prod_{i=1}^{n} p_{Y|X}(y_i|x_i)$$

- The **information capacity** of a discrete memoryless channel is defined as

$$C = \max_{p_X(x)} I(X;Y)$$

  or equivalently

$$C = \max_{p(x)} \sum_{x\in\mathcal{X}, y\in\mathcal{Y}} p(x)p(y|x) \log\left(\frac{p(y|x)}{\sum_{x'} p(y|x')p(x')}\right)$$

- The function $I(X;Y)$ is *concave* in $p_X(x)$. This means we can find a maximizing distribution (and that one exists). Characterizing this distribution can be tricky, but there are some nice examples.

- **Channel Coding Theorem:** The operational capacity of a discrete memoryless channel is equal to the information capacity,

$$C_{\text{op}} = \max_{p(x)} I(X;Y)$$

  This equality consists of two statements:
  - **Achievability:** Every rate $R < C$ is achievable, i.e., there exists a sequence of $(2^{nR}, n)$ coding schemes such that the maximum error probability $P_{\text{e,max}}^{(n)}$ converges to zero as the block-length n increases:

$$R < C \quad \Longrightarrow \quad R \text{ is achievable}$$

  - **Converse:** Any sequence of $(2^{nR}, n)$ coding schemes with maximum error probability $P_{\text{e,max}}^{(n)}$ converging to zero as the block-length n increase must have rate $R \leq C$.

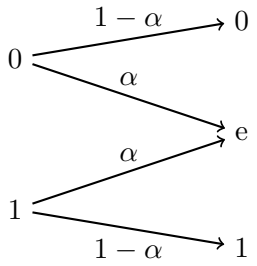
$$R \text{ is achievable} \quad \Longrightarrow \quad R \leq C$$

6.2 Examples of Discrete Memoryless Channels

6.2.1 Binary Erasure Channel (BEC)

- Encoder can send binary inputs $\mathcal{X} = \{0, 1\}$. Channel output is equal to the input with probability $1 - \alpha$ and equal to the erasure symbol 'e' with probability α.

$$Y = \begin{cases} X & \text{with probability } 1 - \alpha \\ e & \text{with probability } \alpha \end{cases}$$



- Computation of capacity
  - Let $E = \{Y = e\}$ be the event that an erasure has occurred and note that $\mathbb{P}[E] = \alpha$.
  - The conditional entropy of Y given X is

$$\begin{aligned} H(Y|X) &= H(E, Y|X) \\ &= H(E|X) + \underbrace{H(Y|X, E)}_{=0} \\ &= H_b(\alpha) \end{aligned}$$

  - Let $\beta = \mathbb{P}[X = 1]$. The entropy of Y is

$$\begin{aligned} H(Y) &= H(Y, E) \\ &= H(E) + H(Y|E) \\ &= H_b(\alpha) + (1 - \alpha) H_b(\beta) \end{aligned}$$

  - Starting with the definition of capacity, we have

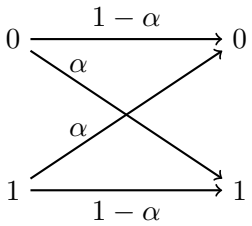
$$\begin{aligned} C &= \max_{p(x)} I(X; Y) \\ &= \max_{p(x)} \Big(H(Y) - H(Y|X) \Big) \\ &= \max_{0 \le \beta \le 1} (1 - \alpha) H_b(\beta) \\ &= (1 - \alpha) H_b(1/2) \\ &= (1 - \alpha) \end{aligned}$$

- **Intuition:** Times slots where input is erased cannot carry any information. If we knew where these were, we could easily achieve $R = 1 - \alpha$. The channel coding theorem tells us that, in the limit of long block lengths, there is no penalty for no knowing the locations of the erasures at the encoder!

6.2.2 Binary Symmetric Channel (BSC)

- Encoder can send binary inputs, $\mathcal{X} = \{0, 1\}$. Channel flips the input with probability α.

$$Y = \begin{cases} X & \text{with probability } 1-\alpha \\ 1-X & \text{with probability } \alpha \end{cases}$$



- Computation of capacity
  - Let $E = \{Y = 1 - X\}$ be the event that the input is flipped and observe that $\mathbb{P}[E] = \alpha$.
  - Letting $\beta = \mathbb{P}[X = 1]$, we have

$$\begin{aligned} C &= \max_{\beta \in [0,1]} I(X;Y) \\ &= \max_{\beta \in [0,1]} \Big(H(Y) - H(Y|X) \Big) \\ &= \max_{\beta \in [0,1]} \Big(H_b(\beta(1-\alpha) + (1-\beta)\alpha) - H_b(\alpha) \Big) \\ &= H_b\big(\tfrac{1}{2}(1-\alpha) + \tfrac{1}{2}\alpha\big) - H_b(\alpha) \\ &= 1 - H_b(\alpha) \end{aligned}$$

  where the last step is achieved when $p_X(0) = 1/2$, i.e. X is Bernoulli(1/2)

- **Intuition:** If $\alpha < 1/2$, then flipping bits is more damaging than an erasure.

6.3 Converse to the Channel Coding Theorem:

- We now prove the converse to the channel coding theorem: Any sequence of $(2^{nR}, n)$ coding schemes with maximum error probability $P_{\text{e,max}}^{(n)}$ converging to zero a the block-length n increases must have rate $R \leq C$.

$$R \text{ is achievable} \quad \Longrightarrow \quad R \leq C$$

- **Lemma:** For any input distribution $p_{X^n}(x^n)$, the mutual information between the input X^n and output Y^n of a discrete memoryless channel with capacity C obeys

$$I(X^n; Y^n) \leq nC.$$

- **Proof:** Fix any distribution on X^n and observe that

$$\begin{aligned} I(X^n;Y^n) &= H(Y^n) - H(Y^n|X^n) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1^{i-1}, X^n) \qquad \text{(Chain rule)} \end{aligned}$$

$$
\begin{aligned}
&= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) && \text{(Channel is Memoryless)} \\
&= \sum_{i=1}^{n} H(Y_i|Y_1^{i-1}) - \sum_{i=1}^{n} H(Y_i|X_i) && \text{(Chain rule)} \\
&\leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) && \text{(Conditioning cannot increase entropy)} \\
&= \sum_{i=1}^{n} I(X_i;Y_i) \\
&\leq nC && \text{(Definition of } C\text{)}
\end{aligned}
$$

- Now we would like to show that $R \leq C$ for any sequence of channel codes with vanishing probability of error, $P_{\mathrm{e}}^{(n)} \to 0$ as $n \to \infty$.
- **Lemma:** (Fano's Inequality) Let $\hat{W}$ be an estimate of the message $W \in \{1, \cdots, 2^{nR}\}$. The conditional entropy of W given $\hat{W}$ is related to the average error probability $P_e^{(n)} = \mathbf{P}(W \neq \hat{W})$ via the inequality
$$H(W|\hat{W}) \leq 1 + P_e^{(n)} nR$$
- **Proof:** By Fano's inequality,
$$P_e^{(n)} = \mathbf{P}(W \neq \hat{W}) \geq \frac{H(W|\hat{W}) - 1}{\log(2^{nR})} = \frac{H(W|\hat{W}) - 1}{nR}$$
- Proof of converse to the channel coding theorem
  - Starting with the fact that W is uniformly distributed on $\{1, \cdots, 2^{nR}\}$, we have
$$
\begin{aligned}
nR &= H(W) \\
&= H(W|\hat{W}) + I(W;\hat{W}) \\
&\leq 1 + P_e^{(n)} nR + I(W;\hat{W}) && \text{(Fano's inequality)} \\
&\leq 1 + P_e^{(n)} nR + I(X^n;Y^n) && \text{(Data processing inequality)} \\
&\leq 1 + P_e^{(n)} nR + nC && \text{(Lemma: } I(X^n;Y^n) \leq nC\text{)}
\end{aligned}
$$
  - Dividing both sides by n and rearranging leads to
$$P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}$$
  - Itf $R > C$ then the right-hand side is strictly positive for large values of n, and thus R is not achievable. We have shown that
$$R > C \quad \Longrightarrow \quad R \text{ is not achievable,}$$
which concludes the proof of the converse to the coding theorem.
- The converse we proved is called a *weak* converse since it says that the error probability is bounded away from zero, but does not rule out the probability that the error is small. Using more advanced techniques, it is possible to prove a *strong* converse which says that if $R > C$ then the probability of error converges to one

6.4 Proof of Channel Coding Theorem via Random Coding

- We prove the achievability part of the channel coding theorem: Every rate $R < C$ is achievable, i.e., there exists a sequence of $(2^{nR}, n)$ coding schemes such that the maximum error probability $P_{\text{e,max}}^{(n)}$ converges to zero as the block-length n increases:

$$R < C \quad \Longrightarrow \quad R \text{ is achievable}$$

- Any $(2^{nR}, n)$ encoder $\mathcal{E}$ can be represented by a *codebook* $\mathcal{C}$, i.e., a massive lookup table whose rows are length-n vectors:

$$\mathcal{C} = \begin{bmatrix} x^n(1) \\ x^n(2) \\ \vdots \\ x^n(2^{nR}) \end{bmatrix} = \begin{bmatrix} x_1(1) & x_2(1) & \cdots & x_n(1) \\ x_1(2) & x_2(2) & \cdots & x_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ x_1(2^{nR}) & x_2(2^{nR}) & \cdots & x_n(2^{nR}) \end{bmatrix}$$

  To communicate message w, the encoder sends the w'th codeword:

$$\mathcal{E}(w) = x^n(w)$$

- Properties of the optimal decoder

  - The the probability of error is minimized by the maximum a posteriori (MAP) decoder:

$$\begin{aligned} \mathcal{D}^{\text{MAP}}(y^n) &= \arg \max_{w \in \{1, \cdots, 2^{nR}\}} p_{W|Y^n}(w \mid y^n) \\ &= \arg \max_{w \in \{1, \cdots, 2^{nR}\}} p_W(w) p_{Y^n|W}(y^n|w). \end{aligned}$$

    The second line follows by Bayes' theorem and the fact that $p_{Y^n}(y^n)$ does not depend on w.

  - Because the message is uniform, the MAP decoder is equal to the maximum likelihood (ML) decoder:

$$\mathcal{D}^{\text{ML}}(y^n) = \arg \max_{w \in \{1, \cdots, 2^{nR}\}} p_{Y^n|W}(y^n|w).$$

  - Furthermore, because $W \to \mathcal{E}(W) \to Y^n$ forms a Markov chain, we have

$$p_{Y^n|W}\big(y^n|w\big) = p_{Y^n|X^n}\big(y^n|x^n(w)\big).$$

- For a given R and block length n, the *optimal* coding scheme is the encoding decoding pair $(\mathcal{E}, \mathcal{D})$ that minimizes the probability of error $P_{\text{e,max}}^{(n)}$. However, the design and analysis of optimal coding schemes is challenging due to the fact that the number of messages grows exponentially with the block length.

- For the purposes of proving achievability of the channel coding theorem, it is sufficient to show that there exists a sequence of *good* (although not necessarily optimal) coding schemes which achieve capacity for any rate $R < C$.

- We use the following approach based on the ideas of *random coding*:

  (1) Design a distribution over set of potential codebooks.

  (2) Show that the error probability (averaged over the distribution on codebooks) is small.

  (3) Conclude that there must exists a good (nonrandom) coding scheme.

6.4.1 Encoding & Decoding

- Random codebook construction:

  (1) choose input distribution $p_X(x)$.

  (2) Let each entry in the codebook $\mathcal{C}$ be drawn iid $\sim p_X(x)$. The distribution of the random code book is

$$\mathcal{C} = \begin{bmatrix} X_1(1) & X_2(1) & \cdots & X_n(1) \\ X_1(2) & X_2(2) & \cdots & X_n(2) \\ \vdots & \vdots & \ddots & \vdots \\ X_1(2^{nR}) & X_2(2^{nR}) & \cdots & X_n(2^{nR}) \end{bmatrix}, \qquad p(\mathcal{C}) = \prod_{w=1}^{2^{nR}} \prod_{k=1}^{n} p_X(x_k(w))$$

  (3) Reveal the resulting codebook to the encoder and decoder.

- **Encoding:** The source generates a random message W distributed uniformly on $\{1, 2 \cdots, 2^{nR}\}$. The encoder looks up the corresponding row in the codebook and sends $X^n(w)$.

- **Decoding:**

  - By construction, each row of the codebook has has the same distribution:

$$p_{X^n}(x^n) = \prod_{k=1}^{n} p_X(x_k)$$

    Consequently, the distribution of the output of the channel is given by

$$p_{Y^n}(y^n) = \sum_{x^n \in \mathcal{X}^n} p_{X^n}(x^n) p_{Y^n|X^n}(y^n|x^n) = \prod_{k=1}^{n} p_Y(y)$$

    where

$$p_Y(y) = \sum_{x \in \mathcal{X}} p_X(x) p_{Y|X}(y|x)$$

  - The distributions described above arise from the randomness in both the codebook and the message. These distributions are different from the conditional distributions of the input and output corresponding to a specific realization of the codebook, which are given by

$$P_{X^n|\mathcal{C}}(x^n) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} \mathbf{1}(x^n = x^n(w))$$

$$p_{Y^n|\mathcal{C}}(y^n|\mathcal{C}) = \frac{1}{2^{nR}} \sum_{w=1}^{2^{nR}} p_{Y^n|X^n}(y^n|x^n(w))$$

  - The **information density** associated with the joint distribution of (X^n, Y^n) is defined as

$$i(x^n; y^n) = \log\left(\frac{p_{X^n,Y^n}(x^n, y^n)}{p_{X^n}(x^n) p_{Y^n}(y^n)}\right) = \log\left(\frac{p_{Y^n|X^n}(y^n|x^n)}{p_{Y^n}(y^n)}\right).$$

    Because the input is iid and the channel is memoryless, the information density can be decomposed as

$$i(x^n; y^n) = \sum_{k=1}^{n} i(x_k; y_k), \qquad \text{where} \quad i(x; y) = \log\left(\frac{p_{Y|X}(y|x)}{p_Y(y)}\right)$$

- The optimal decoder can be expressed in terms of the information density:

$$\begin{aligned} D^{\text{MAP}}(y^n) &= \arg\max_w p_{Y^n|X^n}(y^n|x^n(w)) \\ &= \arg\max_w \frac{p_{Y^n|X^n}(y^n|x^n(w))}{p_{Y^n}(y^n)} \\ &= \arg\max_w i(x^n(w); y^n) \qquad \text{(because logarithm is increasing)} \end{aligned}$$

- To help with the analysis we will study a sup-optimal thresholding decoder. For a given threshold T we define the decoding rule as follows:

$$\mathcal{D}(y^n) = \begin{cases} \hat{w}, & \text{if } i(x^n(\hat{w}); y^n) > T \text{ and } i(x^n(w); y^n) \le T \text{ for all } w \neq \hat{w} \\ 0, & \text{otherise} \end{cases}$$

6.4.2 Performance Analysis

- There is a decoding error if $\hat{W} \neq W$. We compute the probability of this event with respect to the distribution on the codebook $\mathcal{C}$

$$\begin{aligned} \mathbb{P}\Big[\hat{W} \neq W\Big] &= \sum_{\mathcal{C}'} \mathbb{P}\big[\mathcal{C} = \mathcal{C}'\big] \mathbb{P}\Big[\hat{W} \neq W | \mathcal{C} = \mathcal{C}'\Big] \\ &= \sum_{\mathcal{C}'} \mathbb{P}\big[\mathcal{C} = \mathcal{C}'\big] \frac{1}{2^{nR}} \sum_{w=1}^{2^nR} \mathbb{P}\Big[\hat{W} \neq W | \mathcal{C} = \mathcal{C}', W = w\Big] \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^nR} \sum_{\mathcal{C}'} \mathbb{P}\big[\mathcal{C} = \mathcal{C}'\big] \mathbb{P}\Big[\hat{W} \neq W | \mathcal{C} = \mathcal{C}', W = w\Big] \\ &= \frac{1}{2^{nR}} \sum_{w=1}^{2^nR} \sum_{\mathcal{C}'} \mathbb{P}\big[\mathcal{C} = \mathcal{C}'\big] \mathbb{P}\Big[\hat{W} \neq W | \mathcal{C} = \mathcal{C}', W = 1\Big] \qquad \text{(by symmetry)} \\ &= \sum_{\mathcal{C}'} \mathbb{P}\big[\mathcal{C} = \mathcal{C}'\big] \mathbb{P}\Big[\hat{W} \neq W | \mathcal{C} = \mathcal{C}', W = 1\Big] \\ &= \mathbb{P}\Big[\hat{W} \neq W | W = 1\Big] \end{aligned}$$

  In other words, the average probability of error with respect to the codebook does not depend on which message was sent. So, for the purposes of analysis, it is sufficient condition on the event that the first codeword was sent.

- Henceforth, we will condition on the event $W = 1$. Recall that the columns of the code book $X^n(w)$ have entries iid $\sim p_X(x)$. The received vector Y^n is distributed jointly with $X^n(1)$ and is independent of $X^n(2), \cdots X^n(2^{nR})$.

- Let $\mathcal{A}$ be the event that the information density corresponding to the correct message exceeds the threshold:

$$\mathcal{A} = \{i(X^n(1); Y^n) > T\}$$

  Let $\mathcal{B}$ be the event that none of the information densities corresponding to the incorrect messages (i.e., the "imposters") exceed the threshold:

$$\mathcal{B} = \bigcap_{w=2}^{2^{nR}} \underbrace{\{i(X^n(w); Y^n) \le T\}}_{\mathcal{B}(w)}$$

- If both $\mathcal{A}$ and $\mathcal{B}$ occur, then the threshold decoder must return the correct estimate $\hat{W} = 1$,

$$\mathcal{A} \cap \mathcal{B} \implies \hat{W} = W$$

This means that the error probability can be upper bounded as follows:

$$\begin{aligned} \mathbb{P}\left[\hat{W} \neq W\right] &\leq \mathbb{P}[\mathcal{A}^c \cup \mathcal{B}^c] && \text{(De Morgan's Law)} \\ &\leq \mathbb{P}[\mathcal{A}^c] + \mathbb{P}[\mathcal{B}^c] && \text{(Union Bound)} \end{aligned}$$

where the superscript c denotes the complement of the event. The event $\mathcal{A}^c$ means that the true message has failed to reach the threshold. The event $\mathcal{B}^c$ means that an imposter message has risen above the threshold.

- We first consider the probability $\mathbb{P}[\mathcal{A}^c]$.
  - The event $\mathcal{A}$ corresponds to the input $X^n(1)$ and output Y^n associated with the correct message. By construction, the entries of the vectors are jointly iid with

$$(X_k(1), Y_k) \overset{\text{iid}}{\sim} p_X(x) p_{Y|X}(y|x), \quad k = 1, \ldots, n$$

Recall that, $p_X(x)$ is the prior distribution on the codebook and $p_{Y|X}(y|x)$ is the channel. Therefore, the information density corresponds to the sum of iid random variables:

$$i(X^n(1); Y^n) = \sum_{k=1}^{n} \underbrace{i(X_k(1); Y_k)}_{\text{iid random variables}}$$

The expected valued of each term in the sum is equal to the mutual information in a single use of the channel

$$\mathbb{E}[i(X_k(1); Y_k)] = I(X; Y) \quad \text{where} \quad (X, Y) \sim p_X(x) p_{Y|X}(y|x)$$

  - By the law of large numbers, the normalized information density converges to its expectation:

$$\frac{1}{n} i(X^n(1); Y^n) = \frac{1}{n} \sum_{k=1}^{n} i(X_k(1); Y_k) \to I(X; Y) \qquad \text{as } n \to \infty$$

  - If we chose any $\epsilon > 0$ and consider the threshold

$$T = n(I(X; Y) - \epsilon)$$

then there exists N_ϵ such that for all $n \geq N_\epsilon$,

$$\begin{aligned} \mathbb{P}[\mathcal{A}^c] &= \mathbb{P}[i(X^n(1); Y^n) \leq n(I(X; Y) - \epsilon)] \\ &= \mathbb{P}\left[\frac{1}{n} \sum_{k=1}^{n} i(X_k(1); Y_k) \leq I(X; Y) - \epsilon\right] \leq \epsilon \end{aligned}$$

- Next we consider the probability $\mathbb{P}[\mathcal{B}^c]$.
  - Applying the union bound yields

$$\mathbb{P}[B^c] = \mathbb{P}\left[\bigcup_{w=2}^{2^{nR}} \{i(X^n(w); Y^n) > T\}\right] \leq \sum_{w=2}^{2^{nR}} \mathbb{P}[i(X^n(w); Y^n) > T]$$

- For $w \neq 1$, the vector $X^n(w)$ and the output Y^n are independent. The joint distribution on these vectors satisfies:
$$(X_k(1), Y_k) \overset{\text{iid}}{\sim} p_X(x)p_Y(y)$$
- To bound the probability $\mathbb{P}[i(X^n(w); Y^n) > T]$ we will use a *Chernoff bound*. Recall that for a random variable Z, we have, for any $\lambda > 0$,
$$\mathbb{P}[Z \geq t] = \mathbb{P}[\lambda Z \geq \lambda t] = \mathbb{P}\Big[2^{\lambda Z} \geq 2^{\lambda t}\Big] \leq \frac{\mathbb{E}\big[2^{\lambda Z}\big]}{2^{\lambda t}}.$$
- Applying the Chernoff bound with $Z = i(X^n; Y^n)$ and $\lambda = 1$ we have
$$\begin{aligned}
\mathbb{P}[i(X^n(w); Y^n) > T] &\leq 2^{-T}\mathbb{E}\Big[2^{i(X^n(w);Y^n)}\Big] \\
&= 2^{-T} \sum_{x^n, y^n} p_{X^n}(x^n)p_{Y^n}(y^n)2^{i(x^n;y^n)} \\
&= 2^{-T} \sum_{x^n, y^n} p_{X^n}(x^n)p_{Y^n}(y^n)\left(\frac{p_{X^n,Y^n}(x^n, y^n)}{p_{X^n}(x^n)p_{Y^n}(y^n)}\right) \\
&= 2^{-T} \sum_{x^n, y^n} p_{X^n,Y^n}(x^n, y^n) \\
&= 2^{-T}
\end{aligned}$$
- Putting it all together,
$$\mathbb{P}[\mathcal{B}^c] \leq \sum_{w=2}^{2^{nR}} 2^{-T} \leq 2^{nR-T}$$

- Combining the above bounds, we see that if $T = n(I(X;Y) - \epsilon)$ with
$$R \leq I(X;Y) - 2\epsilon$$
then for all n large enough,
$$\mathbb{P}\Big[\hat{W} \neq W\Big] \leq \epsilon + 2^{n(R-I(X;Y)+\epsilon)} \leq \epsilon + 2^{-n\epsilon}$$

- This means that, if the rate R is less than the mutual information $I(X;Y)$, then the probability of error converges to zero as n goes to infinity. For any $\epsilon > 0$, there exists N_ϵ such that for all $n \geq N_\epsilon$,
$$R < I(X;Y) - \epsilon \quad \Longrightarrow \quad \mathbb{P}\Big[\hat{W} \neq W\Big] \leq \epsilon$$
or equivalently,
$$R < I(X;Y) \quad \Longrightarrow \quad \mathbb{P}\Big[\hat{W} \neq W\Big] \to 0 \quad \text{as } n \to \infty$$

6.4.3 Strengthening the Proof

- We have not yet specified how to choose the distribution $p_X(x)$ for our codebook. If we use the distribution that maximizes the mutual information $I(X;Y)$, and thus achieves the information capacity, we have $I(X;Y) = C$ and so the condition becomes
$$R < C \quad \Longrightarrow \quad \mathbb{P}\Big[\hat{W} \neq W\Big] \to 0 \quad \text{as } n \to \infty$$

- To get rid of the average over codebooks recall that we have shown that for any $\epsilon > 0$, there exists N_ϵ such that for all $n \geq N_\epsilon$,

$$\sum_{\mathcal{C}'} \mathbb{P}[\mathcal{C} = \mathcal{C}'] \mathbb{P}\Big[\hat{W} \neq W \mid \mathcal{C} = \mathcal{C}'\Big] \leq \epsilon$$

  This means that there must exists at least one (nonrandom) codebook $\mathcal{C}^*$ with

$$\mathbb{P}\Big[\hat{W} \neq W \mid \mathcal{C} = \mathcal{C}^*\Big] \leq \epsilon,$$

  Otherwise the above statement would be violated.

- At this point, it is still possible that the codebook $\mathcal{C}^*$ contains some codewords with bad conditional error probabilities. But if the average error probability is small, how many bad codewords can there be?

  - Define the conditional error probability

$$\lambda(w) = \mathbb{P}\Big[\hat{W} \neq W \mid \mathcal{C} = \mathcal{C}^*, W = w\Big]$$

  - If $\mathbb{P}[W \neq W \mid \mathcal{C} = \mathcal{C}^*] \leq \epsilon$ then the number of "bad" codewords satisfies

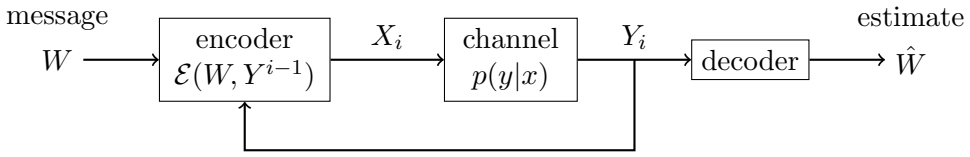
$$\begin{aligned}
\#\{\text{codewords with } \lambda(w) \geq 2\epsilon\} &= \sum_{w=1}^{2^{nR}} \mathbf{1}_{\{\lambda(w) \geq 2\epsilon\}} \\
&\leq \sum_{w=1}^{2^{nR}} \frac{\lambda(w)}{2\epsilon} \\
&= \frac{1}{2\epsilon} 2^{nR} \mathbb{P}\Big[\hat{W} \neq W \mid \mathcal{C} = \mathcal{C}^*\Big] \\
&\leq \frac{1}{2} 2^{nR} = 2^{n(R - \frac{1}{n})}
\end{aligned}$$

  - Thus, if we expunge the worst half of the codewords, the maximum conditional error of the remaining codewords is no greater than 2ϵ and the rate of the new codebook is $R - \frac{1}{n}$. Since this difference goes to zero as $n \to \infty$, we can conclude that

$$R < C \quad \Longrightarrow \quad P_{\text{e,max}}^{(n)} \to 0 \quad \text{as } n \to \infty$$

6.5 Channel Coding with Feedback

- What if the decoder could talk to the encoder? Can this improve the achievable rate?
- Illustration of feedback channel



- Encoder $\mathcal{E}(W, Y^{i-1})$ can use previous channel outputs
- **Theorem:** Feedback *cannot* increase capacity. For a discrete memoryless channel, the capacity with feedback, C_{FB}, is the same as the capacity without feedback:

$$C_{\text{FB}} = C.$$

- This surprising fact stems from the memorylessness of the channel. Of course feedback can help simplify our encoding and decoding schemes in terms of complexity.
- Proof:
  - Using the same steps as in the proof of the converse, we have

$$\begin{aligned} nR = H(W) &= H(W|\hat{W}) + I(W;\hat{W})) \\ &\le 1 + P_e^{(n)} nR + I(W;\hat{W}) && \text{(Fano's Inequality)} \\ &\le 1 + P_e^{(n)} nR + I(W;Y^n) && \text{(Data processing Inequality)} \end{aligned}$$

  - To bound the mutual information between W and Y^n, observe that

$$\begin{aligned} I(W;Y^n) &= H(Y^n) - H(Y^n|W) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \cdots, Y_{i-1}, W) && \text{(Chain Rule)} \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|Y_1, \cdots, Y_{i-1}, W, X_i) && X_i \text{ is a function of } (W, Y^{i-1}) \\ &= H(Y^n) - \sum_{i=1}^{n} H(Y_i|X_i) && \text{since } (W, Y^{i-1}) \to X_i \to Y_i \\ &\le \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i|X_i) && \text{(independence bound)} \\ &= \sum_{i=1}^{n} I(X_i;Y_i) \\ &\le nC && \text{(definition of capacity)} \end{aligned}$$

  - Thus, we have shown that

$$nR \le 1 + P_e^{(n)} nR + nC$$

  - Dividing both sides by n, we see that if $P_e^{(n)} \to 0$ as $n \to \infty$, it must be the case that $R \le C$.

6.6 Coding Theory

- Since Shannon's original paper, people have searched for ways to design efficient encoding and decoding schemes which reach capacity.
- The goal of a code is to introduce redundancy so that if some of the information is corrupted, it can still be recovered.
- **Example:** A *repetition code* is a $(2, n)$ coding scheme for a BSC were a single information bit is repeated over n channel uses.
  - For example, if $n = 5$, to send a 1 we send 11111 and to send 0 we send 00000.
  - The rate of this code is
$$R = \frac{1}{n}$$
  - The code can correct up to $(n-1)/2$ flips, but not any more.
- **Example:** A *parity check code* is a $(2^{n-1}, n)$ coding scheme for a BSC which sends $n-1$ information bits, and the n-th bit encodes the parity of the entire block, i.e. whether the number of 1's in the information bits is even or odd.
  - The rate is
$$R = \frac{n-1}{n}$$
  - Can detect an odd number of flips.
  - Cannot detect even number of flips.
  - Does not say how to correct flips.

6.6.1 Hamming Codes

- *It is better to do the right problem the wrong way than to do the wrong problem the right way* — Richard Hamming
- The Hamming code words corresponding to a $(2^4, 7)$ coding scheme are given by:

0000000	0100101	1000011	1100110
0001111	0101010	1001100	1101001
0010110	0110011	1010101	1110000
0011001	0111100	1011010	1111111

- There are $2^4 = 16$ codewords of length $n = 7$. Thus the code carries $k = 4$ information bits and the rate is
$$R = \frac{4}{7}$$
- The minimum number d in which two codewords differ is class the **minimum distance** of the code.
- The Hamming code above has minimum distance $d = 3$. Thus is can detect up to two errors and correct one error.
- This example is a $(7, 4, 3)$ Hamming code since $n = 7, k = 4, d = 3$

- **Construction:** The Hamming codewords correspond to the null space of the *parity check matrix*

$$H = \begin{bmatrix} 0 & 0 & 0 & 1 & 1 & 1 & 1 \\ 0 & 1 & 1 & 0 & 0 & 1 & 1 \\ 1 & 0 & 1 & 0 & 1 & 0 & 1 \end{bmatrix}$$

  where all arithmetic is module 2. Since H has rank 3, the null space of H has dimension 4. Hence there are 2^4 such codewords.

- The null space of H is a linear subspace, and thus a linear combination of any two codewords is a codeword.

- **Encoding:** Note that we can use the first 4 bits of the codewords as the information bits (there are 2^4 possibilities) and the remaining 3 bits are determined by the code.

- **Decoding:** We want to identify the codeword $\boldsymbol{c}$ that is closest to the received vector $\boldsymbol{r}$. Can we do this without exhaustive search?

  - Note that each codeword $\boldsymbol{c}$ is annihilated by the parity check matrix H, i.e.

$$H\boldsymbol{c} = 0$$

  - Let $\boldsymbol{e}_i$ be a vector with a 1 in the ith position and 0's elsewhere.
  - If the received vector is given by $\boldsymbol{r} = \boldsymbol{c} + \boldsymbol{e}_i$, then

$$H\boldsymbol{r} = H(\boldsymbol{c} + \boldsymbol{e}_i) = H\boldsymbol{c} + H\boldsymbol{e}_i = H\boldsymbol{e}_i$$

    which is simply the ith column of H.
  - Thus, assuming that only one bit was flipped, the vector $H\boldsymbol{r}$ is the binary representation of index of the flipped bit. By flipping this bit in the received vector $\boldsymbol{r}$, we recover the original codeword $\boldsymbol{c}$.

- In general, if we use l rows in H, the code we obtain will have block length $n = 2^l - 1$, carry $k = n - l = 2^l - l - 1$ information bits and have minimum distance $d = 3$.

6.6.2 Beyond Hamming

- In early 1950s Reed and Solomon found a class of multiple error correcting codes for nonbinary channels.

- late 1950s BCH codes have precise control on number of errors that can be corrected.

- All compact disc players use two interleaved $(32, 28, 5)$ and $(28, 24, 5)$ Reed – Solomon codes that allows the decoder to corrects bursts of up to 4000 errors.

- In 1993, Turbo codes are shown to be near capacity achieving.

  - now used in 3G mobile communications
  - deep space satellite communications

- LDPC codes (introduced originally by Gallager in 1960) were shown to achieve rates near capacity using iterative message passage decoding in 1997

  - competes with Turbo Codes. Have lower decoding complexity.

- standard for digital television
- sometimes combined with Reed-Solomon and BCH codes for additional protection.

- In 2007, Arikan introduced Polar codes. Have explicit code construction. Provably achieve capacity and decoding is $O(n \log n)$.

6.6.3 The Hat Game

- Hats are assigned uniformly randomly to $N = 2^m - 1$ players. Each hat is colored 0 or 1.
- Rules of the game:
  - Players act a team – everyone wins or everyone loses.
  - Once hats have been placed, there no communication between team members.
  - When asked the color of their hats, all players must answer simultaneously
  - Each person is allowed to pass rather than guess a color
  - Team wins if at least one player guesses correctly and none guess incorrectly. Otherwise, the team loses.
- Consider $N = 2^2 - 1 = 3$ players.
  - **Strategy 1:** Everyone guesses independently
  $$\mathbb{P}[\text{win}] = 2^{-3} = 1/8 = 0.125$$
  - **Strategy 2:** Everyone guesses (with probability α) or passes independently
  $$\mathbb{P}[\text{win}] = (1 - \alpha/2)^3 - (1 - \alpha)^3$$
  - **Strategy 3:** Two people pass, third guesses $\mathbb{P}[\text{win}] = 0.5$
  - **Strategy 4:** Try this
    - If you see two hats that have the same color, guess the other color
    - If you see two hats of different colors, pass
    $$\mathbb{P}[\text{win}] = 0.75$$
- Consider $N = 2^3 - 1 = 7$ players [The rest is left as an exercise]