Chapter 7: Channel capacity

Chapter 7 outline

- Definition and examples of channel capacity
- Symmetric channels
- Channel capacity properties
- Definitions and jointly typical sequences
- Channel coding theorem: achievability and converse
- Hamming codes
- Channels with feedback
- Source-channel separation theorem
Generic communication block diagram

Communication system
Capacity: key ideas

- choose input set of codewords so they are “non-confusable” at the output
- number of these that we can choose will determine the channel's capacity!
- number that we can choose will depend on the distribution \( p(y|x) \) which characterizes the channel!
- for now we deal with discrete channels

Discrete channel capacity

Definition: A discrete channel is the (physical or abstract) link connecting input \( X \in \mathcal{X} \) and the output \( Y \in \mathcal{Y} \), described by the conditional probability \( p(y|x) \) that the output is \( y \) when the input \( x \).

- Time-Invariant Transition-Probability Matrix: \( P_{i,j} = p(Y = y_j | X = x_i) \).
  \( P \) : each row sum=1.
- Memoryless: \( p(y_n | x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_{n-1}) = p(y_n | x_n) \)
- DMC = Discrete Memoryless Channel
Mathematical description of capacity

• Information channel capacity:

\[ C = \max_{p(x)} I(X; Y) \]

• Operational channel capacity:

Highest rate (bits/channel use) that can communicate at reliably

• Channel coding theorem says: information capacity = operational capacity

Channel capacity

\[ C = \max_{p(x)} I(X; Y) \text{ bits/channel use} \]

Capacity achieving input distribution.

\[ I(X; Y) = \sum_{x, y} p(x, y) \log \left( \frac{p(x, y)}{p(x)p(y)} \right) \]
### Noiseless channel

**Capacity?**
- 0 → 0
- 1 → 1

Channel capacity

\[
C = \max_{p(x)} I(X; Y)
\]
\[
= \max_{p(x)} H(X) - H(X|Y)
\]
\[
= \max_{p(x)} H(X) - 0
\]
\[
= 1
\]
Channel capacity

Capacity?

\( \log_2(9) \) bits/channel use

\[
C = \max_{p(x)} I(X; Y) \\
= \max_{p(x)} H(X) - H(X|Y) \\
= \max_{p(x)} H(X) - \log_2(3) \\
= \log_2(27) - \log_2(3) = \log_2(9)
\]

Binary erasure channel

Capacity?

1-f bits/channel use

\[
C = \max_{p(X)} I(X; Y) \\
= \max_{p(x)} H(X) - H(X|Y) \\
= \max_{p(x)} H(X) - \sum p(y) H(X|Y = y) \\
= \max_{p(x)} H(X) - [p(1-f)H(X|Y = 0) + (1-p)(1-f)H(X|Y = 1) + pf(1-p)f)H(X|Y = e)] \\
= \max_{p(x)} H(X) - [p(1-f)0 + (1-p)(1-f)0 + fH(X)] \\
= \max_{p(x)} H(X)(1 - f) \\
= 1 - f
\]
Channel capacity

Binary symmetric channel

Capacity?
1 - H(f) bits/channel use

\[ C = \max_{p(x)} I(X;Y) \]
\[ = \max_{p(x)} H(Y) - H(Y|X) \]
\[ = \max_{p(x)} H(Y) - \sum p(x)H(Y|X = x) \]
\[ = \max_{p(x)} H(Y) - [\sum p(x)H(f)] \]
\[ = \max_{p(x)} H(Y) - [H(f)] \]
\[ \leq 1 - H(f) \]

[Cover+Thomas pg.187]

Transition probability matrix

- Binary Symmetric Channel: \( \mathcal{X} = \{0, 1\} \) and \( \mathcal{Y} = \{0, 1\} \)

\[
\begin{pmatrix}
1-f & f \\
f & 1-f
\end{pmatrix}
\]

- Binary Erasure Channel: \( \mathcal{X} = \{0, 1\} \) and \( \mathcal{Y} = \{0, e, 1\} \)

\[
\begin{pmatrix}
1-f & f & 0 \\
f & 0 & f \\
0 & f & 1-f
\end{pmatrix}
\]

- Z channel: \( \mathcal{X} = \{0, 1\} \) and \( \mathcal{Y} = \{0, 1\} \)

\[
\begin{pmatrix}
1 & 0 \\
f & 1-f
\end{pmatrix}
\]

Time-Invariant Transition-Probability Matrix: \( P_{i,j} = p(Y = y_j|X = x_i) \)
Symmetric channels

- Weakly Symmetric Channels:
  - All rows are permutations of each other:
    - Each row of $P$ have has the same entropy
  - All columns have the same sum:
    - If $X$ is uniform then $Y$ is uniform

\[
p(y) = \sum_{x \in \mathcal{X}} p(y|x)p(x) = \frac{1}{|\mathcal{X}|} \sum_{x \in \mathcal{X}} p(y|x) = \frac{1}{|\mathcal{Y}|}
\]

- Symmetric channels:
  - All rows are permutations of each other
  - All columns are permutation of each other

- Symmetric Channels $\Rightarrow$ Weakly Symmetric Channels

Capacity of weakly symmetric channels

Theorem: For a weakly symmetric channel,

\[
C = \log |\mathcal{Y}| - H(\text{ row of transition matrix}),
\]

and this is achieved by a uniform distribution on the input alphabet.
Properties of the channel capacity

\[ C = \max_{p(x)} I(X; Y) \]

- \( C \geq 0 \) since \( I(X; Y) \geq 0 \)
- \( C \leq \log |\mathcal{X}| \) since \( C = \max I(X; Y) \leq \max H(X) = \log |\mathcal{X}| \)
- \( C \leq \log |\mathcal{Y}| \) for the same reason
- \( I(X; Y) \) is a continuous function of \( p(x) \)
- \( I(X; Y) \) is a concave function of \( p(x) \). Since \( I(X; Y) \) is a concave function over a closed convex set, local max = global max.

Preview of the channel coding theorem

- What happens when we use the channel \( n \) times?
- The capacity of \( n \) uses of the channel:
  \[ C^{(n)} = \frac{1}{n} \max_{p(x_1, x_2, \ldots, x_n)} I(X_1, X_2, \ldots, X_n; Y_1, Y_2, \ldots, Y_n) \]
- For discrete memoryless channel we have:
  \[ I(X^n; Y^n) = H(Y_1, Y_2, \ldots, Y_n) - H(Y_1, Y_2, \ldots, Y_n | X_1, X_2, \ldots, X_n) \]
  \[ = H(Y_1, Y_2, \ldots, Y_n) - \sum_{i=1}^{n} H(Y_i | Y_{i-1}, \ldots, Y_1, X_1, X_2, \ldots, X_n) \]
  \[ = \sum_{i=1}^{n} H(Y_i | X_{i-1}, \ldots, Y_1) - \sum_{i=1}^{n} H(Y_i | X_i) \]
  \[ \leq \sum_{i=1}^{n} H(Y_i) - \sum_{i=1}^{n} H(Y_i | X_i) = \sum_{i=1}^{n} I(X_i; Y_i) \leq nC \]
- \( C^{(n)} \) equals \( C \) and the optimal \( n \) symbol input distribution is i.i.d.
Preview of the channel coding theorem

- For large $n$, subsets of inputs to channel produce essentially disjoint subsets of outputs.

- For each typical input sequence (how many are there?) there are about $2^{nH(Y|X)}$ possible $Y$ sequences, all equally likely.

- Want to ensure that no two typical $X$ sequences produce the same $Y$ sequence.

- There are $2^{nH(Y)}$ typical $Y$ sequences. Dividing, we get $2^{nH(Y)}/2^{nH(Y|X)} = 2^{nI(X;Y)}$ distinguishable input sequences.

Let’s make this rigorous!
Definitions

**Definition:** Discrete channel. A discrete channel is the (physical or abstract) link connecting input $X \in \mathcal{X}$ and the output $Y \in \mathcal{Y}$, described by the conditional probability $p(y|x)$ that the output is $y$ when the input $x$.

**Definition:** $n$-th extension. The $n$-th extension of the DMC is the channel $(\mathcal{X}^n, p(y^n|x^n), \mathcal{Y}^n)$, where $p(y_k|x^k, y^{k-1}) = p(y_k|x_k)$, $k = 1, 2, \ldots, n$.

Definitions

**Definition:** Channel code. An $(M, n)$ code for the channel $(\mathcal{X}, p(y|x), \mathcal{Y})$ consists of the following:

1. An index set $\{1, 2, \ldots, M\}$ over messages $W$.

2. An encoding function $X^n : \{1, 2, \ldots, M\} \rightarrow \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots$. (This set is called the **codebook $C$**.)

   $x^n(W)$ passes through the channel and is received as a random sequence $Y^n \sim p(y^n|x^n)$.

3. A (deterministic) decoding function $g : \mathcal{Y}^n \rightarrow \{1, 2, \ldots, M\}$,

   which is an estimator $\hat{W} = g(Y^n)$ of $W \in \{1, 2, \ldots, M\}$. It declares an error if $\hat{W} \neq W$. 

**Definitions**

**Definition: Conditional probability of error.**

Let \( \lambda_i = \Pr \{ g(Y^n) \neq i | X^n = x^n(i) \} = \sum_{y^n} p(y^n | x^n(i)) I_{\{g(y^n) \neq i\}}(y^n) \)

be the **conditional probability of error** given that index \( i \) was sent.

**Definition: Maximal probability of error.** The maximal probability of error \( \lambda^{(n)} \) for an \((M, n)\) code is defined as

\[
\lambda^{(n)} = \max_{i \in \{1, 2, \ldots, M\}} \lambda_i.
\]

**Definition: Average probability of error.** The (arithmetic) average probability of error \( P_\varepsilon^{(n)} \) for an \((M, n)\) code is:

\[
P_\varepsilon^{(n)} = \frac{1}{M} \sum_{i=1}^{M} \lambda_i.
\]

- Note that \( P_\varepsilon^{(n)} = \Pr \{ W \neq g(Y^n) \} \) if \( W \) is chosen uniformly.
- Also, \( P_\varepsilon^{(n)} \leq \lambda^{(n)} \); i.e., the average probability of error is less than the maximal probability of error.

**Definition: Rate.** The rate \( R \) of an \((M, n)\) code is \( R = \frac{\log M}{n} \) bits per transmission.

**Definition: Achievability.** A rate \( R \) is called **achievable** if there exists a sequence of \( ([2^{nR}], n) \) codes such that \( \lambda^{(n)} \) (i.e., maximal \( \Pr \{ \text{Error} \} \)) tends to 0 as \( n \to \infty \). Note \( (2^{nR}, n) \) codes mean \( ([2^{nR}], n) \) codes.

**Definition: Capacity.** The capacity of a channel is the supremum of all achievable rates.
What's our goal?

Mathematical description of capacity

- Information channel capacity:

\[ C = \max_p I(X; Y) \]

- Operational channel capacity:

Highest rate (bits/channel use) that can communicate at reliably

- Channel coding theorem says: information capacity = operational capacity
Recall the definition of typical sequences....

- **Definition:** The typical set $A_\epsilon^{(n)}$ with respect to $p(x)$ is the set of sequences $(x_1, x_2, \ldots, x_n) \in \mathcal{X}^n$ with the property
  \[
  2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}.
  \]

- All sequence in the typical set have roughly equal probabilities. Clearly, a new notion of approximation is used in such a statement. We call that the “exponential approximation.”

- If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then
  \[
  H(X) - \epsilon \leq -\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \leq H(X) + \epsilon.
  \]

**Let’s make this 2-D!**

---

**Jointly typical sequences**

**Definition:** The set $A_\epsilon^{(n)}$ of jointly typical sequences $\{(x^n, y^n)\}$ with respect to the distribution $p(x, y)$ is the set of $n$-sequences with empirical entropies $\epsilon$-close to the true entropies:

\[
A_\epsilon^{(n)} = \{(x^n, y^n) \in \mathcal{X}^n \times \mathcal{Y}^n : \left| -\frac{1}{n} \log p(x^n) - H(X) \right| < \epsilon, \\
\left| -\frac{1}{n} \log p(y^n) - H(Y) \right| < \epsilon, \\
\left| -\frac{1}{n} \log p(x^n, y^n) - H(X, Y) \right| < \epsilon \},
\]

where

\[
p(x^n, y^n) = \prod_{i=1}^n p(x_i, y_i).
\]
Joint Asymptotic Equipartition Theorem (AEP)

Theorem: Joint AEP Let \((X^n, Y^n)\) be \(n\)-sequences iid \(p(x^n, y^n) = \prod_{i=1}^{n} p(x_i, y_i)\). Then:

1. \(\Pr\{ (X^n, Y^n) \in A_{\epsilon}^{(n)} \} \to 1 \) as \(n \to \infty\).

2. \(|A_{\epsilon}^{(n)}| \leq 2^{n(H(X,Y)+\epsilon)}\).

3. If \((\tilde{X}^n, \tilde{Y}^n) \sim p(x^n)p(y^n)\) so that \(\tilde{X}^n\) and \(\tilde{Y}^n\) are independent with the same marginals as \(p(x^n, y^n)\), then we have that:
   - \(\Pr\{ (\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon}^{(n)} \} \leq 2^{-n(I(X;Y)-3\epsilon)}\)
   - \(\Pr\{ (\tilde{X}^n, \tilde{Y}^n) \in A_{\epsilon}^{(n)} \} \geq (1-\epsilon)2^{-n(I(X;Y)+3\epsilon)}\) for sufficiently large \(n\).

Joint typicality images

- There are about \(2^{nH(X)}\) typical \(X\) in all
- Each typical \(Y\) is jointly typical with about \(2^{nH(X|Y)}\) of those typical \(X\)'s
- The jointly typical pairs are a fraction \(2^{-nI(X;Y)}\) of the inner rectangular
Channel coding theorem

Theorem: Channel coding theorem For a DMC, all rates below capacity $C$ are achievable.

- Specifically, for every rate $R < C$, there exists a sequence of $\left[2^{nR}\right], n)$ codes with maximum probability of error $\lambda^{(n)} \rightarrow 0$.
- Conversely, any sequence of $\left[2^{nR}\right], n)$ codes with $\lambda^{(n)} \rightarrow 0$ must have $R \leq C$.

A very counterintuitive result! Despite channel errors you can get arbitrarily low bit error rates provided that $R < C$.

Key ideas behind channel coding theorem

- Allow for arbitrarily small but nonzero probability of error
- Use channel many times in succession: law of large numbers!
- Probability of error calculated over a random choice of codebooks
- Joint typicality decoders used for simplicity of proof
- NOT constructive! Does NOT tell us how to code to achieve capacity!
Key ideas behind the channel coding theorem

- **Code construction**: Fix a distribution $p(x)$, generate a codebook $C$ as a $2^{nR} \times n$ matrix, with i.i.d. entries from $p(x)$. The codebook is fixed once generated, revealed to the decoder.

- **Transmission**: When transmitting a message $w \in \{1, \ldots, 2^{nR}\}$, we pick the $w$-th row of the codebook, $x^n(w) \in X^n$, and send it over $n$ uses of the channel.

- **Decoding**: We use joint typicality decoding. Upon receiving the channel output $y^n$, the receiver checks for each possible value of $w$, whether $x^n(w)$ is jointly typical with $y^n$
  - if there exists a unique $w$, such that $(x^n(w), y^n)$ are jointly typical, it decodes by claiming message $w$ is transmitted.
  - if there does not exist such an $w$, or if there exists more than one codeword jointly typical with $y^n$, the receiver declares an error.

Random codes

- We generate a $(\lfloor 2^{nR} \rfloor, n)$ code $C$ at random according to $p(x)$. Each codeword $x^n(w)$, $w = 1, \ldots, M = \lfloor 2^{nR} \rfloor$ is generated with probability
  \[
  p(x^n(w)) = \prod_{i=1}^{n} p(x_i(w)) \quad \Rightarrow \quad P(C) = \prod_{w=1}^{M} \prod_{i=1}^{n} p(x_i(w))
  \]

- The codewords are rows of the matrix:
  \[
  C = \begin{bmatrix}
  x_1(1) & x_2(1) & \cdots & x_n(1) \\
  \vdots & \vdots & \ddots & \vdots \\
  x_1(M) & x_2(M) & \cdots & x_n(M)
  \end{bmatrix}
  \]

- A message $W$ is chosen according to a uniform distribution $\Pr(W = w) = 2^{-nR}$, $w = 1, 2, \ldots, 2^{nR}$.
Transmission

- The \( w \)-th codeword \( X^n(w) \), the \( w \)-th row of \( C \) is sent over the channel.
- Receiver receives sequence \( Y^n \) according to the distribution
  \[
P(y^n|x^n(w)) = \prod_{i=1}^{n} p(y_i|x_i(w)).
\]
- Receiver guesses which message was sent: declares \( \hat{w} \) was sent if:
  1. \( (X^n(\hat{w}), Y^n) \) is jointly typical
  2. There is NO other index \( w' \neq \hat{w} \) such that \( (X^n(w'), Y^n) \in A^{(n)}_\varepsilon \).
   Otherwise it declares an error.
- Decoding error if \( \hat{w} \neq w \). Let \( \mathcal{E} \) be the error event \( \{\hat{W} \neq W\} \).

Probability of error

- We compute the average probability of error, averaged over all \( w \) and all \( C \),
  \[
P(E) = \sum_C P(E|C)P(C) = \sum_C P(C) \frac{1}{M} \sum_{w=1}^{M} \lambda_{w}(C)
  = \frac{1}{M} \sum_{w=1}^{M} \sum_C P(C)\lambda_{w}(C)
\]
- **Key observation:** the inner sum does not depend on \( w \) because of the symmetric generation process of the code. Thus,
  \[
P(E) = \sum_C P(C)\lambda_1(C) = P(E|W = 1)
\]
- Let \( E_i = \{(x^n(i), y^n) : (x^n(i), y^n) \in A^{(n)}_\varepsilon\} \). Then
  \[
P(E|W = 1) = P(\overline{E}_1 \cup E_2 \cup \ldots \cup E_M) \leq P(\overline{E}_1) + \sum_{i=2}^{M} P(E_i)
  \]
Probability of error

\[ P(E|W = 1) \leq P(E_1) + \sum_{i=2}^{M} P(E_i) \]

- Now remember the properties of typical sequences.
- Property 1 \( \Rightarrow \) \( P(E_1) = 1 - P(E) \leq \epsilon \)
- \( X^n(1) \) and \( X^n(w) \) are independent if \( w > 1 \). This implies that \( Y^n \) is also independent of \( X^n(w) \). Thus,
- Property 3 \( \Rightarrow \) \( P(E_i) \leq 2^{-n(I(X;Y) - 3\epsilon)} \)
- If \( R \leq I(X;Y) \), we conclude that
  \[ P(E|W = 1) \leq \epsilon + (M - 1)2^{-n(I(X;Y) - 3\epsilon)} \leq \epsilon + 2^n R 2^{-n(I(X;Y) - 3\epsilon)} \leq 2\epsilon \]
  where the last inequality holds provided that \( n \) is large enough.

Random codes?

- We have proved that averaged over the random codebooks, and all possible messages, the probability of error can be made to decay to 0, as long as \( R < I \).
- If we set \( p(x) \) to be the PMF which maximizes \( I(X;Y) \), the above condition \( R \leq I(X;Y) \) becomes \( R < C \)
- There must exist at least one code \( C^* \) for which the average probability of error goes to zero as \( n \) goes to infinity.
- Since above,
  \[ P(E|C^*) = \frac{1}{2^n R} \sum_w \lambda_w(C^*) \leq 2\epsilon \]
  at least half of the codewords of \( C^* \) must have probability of error less than \( 4\epsilon \). We keep such codewords to form a code which has \( 2^n R^{-1} \) codewords. This code has maximal probability of error less than \( 4\epsilon \) and a rate \( R - \frac{1}{n} \). Thus, when \( n \to \infty \) it achieves the rate \( R \).
10.3 Proof of the noisy-channel coding theorem

Analogy

Imagine that we wish to prove that there is a baby in a class of one hundred babies who weighs less than 10 kg. Individual babies are difficult to catch and weigh. Shannon’s method of solving the task is to scoop up all the babies and weigh them all at once on a big weighing machine. If we find that their average weight is smaller than 10 kg, there must exist at least one baby who weighs less than 10 kg — indeed there must be many! Shannon’s method isn’t guaranteed to reveal the existence of an underweight child, since it relies on there being a tiny number of elephants in the class. But if we use his method and get a total weight smaller than 1000 kg then our task is solved.

Figure 10.3. Shannon’s method for proving one baby weighs less than 10 kg.

Channel coding theorem

Theorem: Channel coding theorem For a DMC, all rates below capacity $C$ are achievable.

- Specifically, for every rate $R < C$, there exists a sequence of $\left[2^nR\right], n$ codes with maximum probability of error $\lambda(n) \rightarrow 0$.
- Conversely, any sequence of $\left[2^nR\right], n$ codes with $\lambda(n) \rightarrow 0$ must have $R \leq C$.

A very counterintuitive result! Despite channel errors you can get arbitrarily low bit error rates provided that $R < C$. 

[Mackay textbook, pg. 164]
Converse to the channel coding theorem

• Based of Fano’s inequality:

Theorem: Fano’s inequality
For any estimator $\hat{X} : X \rightarrow Y \rightarrow \hat{X}$, with $P_e = \Pr\{X \neq \hat{X}\}$, we have

$$H(P_e) + P_e \log |\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y).$$

This implies $1 + P_e \log |\mathcal{X}| \geq H(X|Y)$ or $P_e \geq \frac{H(X|Y) - 1}{\log |\mathcal{X}|}$.

• $W \rightarrow X^n(W) \rightarrow Y^n \rightarrow \hat{W}$ forms a Markov chain

Lemma: Fano’s inequality. For a DMC with a codebook $C$ and the input message $W$ uniformly distributed over $2^nR$ we have

$$H(W|\hat{W}) \leq 1 + P_e^{(n)} nR.$$
Converse to the channel coding theorem

- Using the previous results we have that

\[
    nR = H(W) = H(W|Y^n) + I(W; Y^n)
\]

\[
    \leq H(W|Y^n) + I(X^n(W); Y^n)
\]

\[
    \leq 1 + P_e^{(n)}nR + I(X^n(W); Y^n)
\]

\[
    \leq 1 + P_e^{(n)}nR + nC
\]

- which implies that

\[
    R \leq \left( C + \frac{1}{n} \right) (1 - P_e^{(n)})^{-1} \rightarrow C \quad \text{for} \quad n \rightarrow \infty
\]

- The above can be written as

\[
    P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}
\]

which proves that if \( R > C \), we have the average probability of error bounded away from 0.

Weak versus strong converses

- Weak converse:

\[
    P_e^{(n)} \geq 1 - \frac{C}{R} - \frac{1}{nR}
\]

- Strong converse:

\[
    P_e^{(n)} \geq 1 - \frac{4A}{n(R-C)^2} - e^{-n(R-C)}
\]

for some finite positive constant \( A \).

- Channel capacity: sharp dividing point below which \( P_e^{(n)} \rightarrow 0 \) exponentially fast, and above which \( P_e^{(n)} \rightarrow 1 \) exponentially fast.
Practical coding schemes

From a constructive point of view, the object of coding is to introduce redundancy in such a way that errors can be detected and corrected.

- Repetition code: rate of the code goes to zero with block length.
- Error-detecting code: adding one parity bit enables the detection of an odd number of errors.
- Linear error-correcting code: vector space structure allows for a parity-check matrix to detect, locate, and correct multiple errors.

Example: channel coding

With permission from David J.C. Mackay
Example: channel coding

Rate?

\[ R = \frac{\text{# source bits}}{\text{# coded bits}} \]

Use repetition code of rate \( R = \frac{1}{n} \):  

- 0 → 000...0
- 1 → 111...1

Decoder?  Majority vote

Probability of error:

\[
P_e = \sum_{i=m+1}^{n} \binom{n}{i} f^i (1 - f)^{n-i}
\]

Need \( n \rightarrow \infty \) for reliable communication!

With permission from David J.C. Mackay
Channel capacity

- Is capacity $R = 0$?

- No! just need better coding!

- Now, we’re more interested in determining capacity than determining (finding codes) to achieve it

- Benchmarks

Practical coding schemes

From a constructive point of view, the object of coding is to introduce redundancy in such a way that errors can be detected and corrected.

- Repetition code: rate of the code goes to zero with block length.

- Error-detecting code: adding one parity bit enables the detection of an odd number of errors.

- Linear error-correcting code: vector space structure allows for a parity-check matrix to detect, locate, and correct multiple errors.
Linear block codes

- The encoder is mapping
  \[ f : \{0, 1\}^k \rightarrow \{0, 1\}^n \]
called \((n, k)\) codes

- A channel code maps \(2^k\) messages into a larger space of cardinality \(2^n\).

- A good channel code would have \(2^k\) messages all evenly and sparsely distributed in the space.

- A block code is called a linear block code iff its \(2^k\) codewords form a \(k\)-dimensional subspace of the \(n\)-dimensional vector space.

- Linear block codes:
  \[ \overline{x} = f(s) = s \cdot G \]
  where \(G\) is an \(k \times n\) matrix with binary entries.

Properties of linear block codes

- The codebook of a linear code is
  \[ C = \{ \overline{x} \in \{0, 1\}^n : \overline{x} = s \cdot G \text{ for some } s \in \{0, 1\}^k \} \]

- \(0^n = 0^k \cdot G\) is always a codeword

- if \(\overline{x}_1, \overline{x}_2 \in C\), then \(\overline{x}_1 + \overline{x}_2 = (s_1 + s_2) \cdot G\) is also a codeword

- For any \((n, k)\) linear code, there exists a parity check matrix \(H\) of dimension \((n - k) \times n\) such that
  \[ \overline{x} \cdot H^T = 0_{1 \times (n-k)} \text{ iff } \overline{x} \in C \]

- This implies that \(G \cdot H^T = 0_{k \times (n-k)}\).
Properties of linear block codes

- If \( C \) is a code with a generating matrix \( G \) in standard form,

\[
G = [I_k | A],
\]

where \( I_k \) denotes the identity matrix and \( A \) is some \( k \times (n - k) \) matrix, then

\[
H = [A^T | I_{n-k}]
\]

is a check matrix for \( C \).

- \( d_{\text{min}} = \min_{x \in C} |x| \) is closely related to the error correcting ability of the code.

- If a code has a minimum distance \( d_{\text{min}} \), it is guaranteed to correct \( [(d_{\text{min}} - 1)/2] \) errors.

Examples

- Repetition code: \( G = [1 \ 1 \ldots 1] \)

- Single parity check: \( G = [I_k | 1_{k \times 1}] \) and \( H = [1_{1 \times k} | I_{n-k}] \)

- Hamming Code: \( n = 2^m - 1, k = 2^m - m - 1, n - k = m, \) and \( d_{\text{min}} = 3 \)

\[
H = [I_m | Q]
\]

where \( Q \) consists of \( 2^m - m - 1 \) columns that are the \( m \)-tuples of weight 2 or more.

\[
G = [Q^T | I_{2^m-m-1}]
\]

- (7,4) Hamming code: \( m = 3 \)

\[
H = \begin{bmatrix}
1 & 0 & 0 & 1 & 0 & 1 & 1 \\
0 & 1 & 0 & 1 & 1 & 1 & 0 \\
0 & 0 & 1 & 0 & 1 & 1 & 1
\end{bmatrix}
\]
Hamming codes

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

- # of codewords?
- what are the codewords?

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

A curiosity: Venn diagrams + Hamming codes

Hamming codes can correct up to 1 error.
Achieving capacity

- Linear block codes: not good enough...

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

- Convolutional codes: not good enough......

- Turbo codes: 1993 Berrou et al. considered two interleaved convolutional codes with parallel cooperative decoders. Achieved close to capacity!

- LDPC codes: Low Density Parity Check Codes introduced by Gallager in his 1963 thesis, later kept alive by Michael Tanner (UIC’s provost!) in the 80s and then “re-discovered” in the 90s, where an iterative message passing algorithm used to decode was formulated. Also achieve close to capacity!

- Excellent survey is (linked to on course website)
Feedback capacity

Channel without feedback

Channel WITH feedback

Feedback capacity

- Assume error-free feedback: does it increase capacity?
- A \((2^nR, n)\) feedback code is
  - A sequence of mapping \(X_i = x_i(W, Y_1, Y_2, \ldots, Y_{i-1})\) for \(i = 1, 2, \ldots, n\)
  - A decoding function \(\hat{W} = g(Y^n)\)
- A rate is achievable if exist a sequence of \((2^nR, n)\) feedback codes such that the maximal \(\text{Pr}\{\text{Error}\}\) tends to 0 as \(n \to \infty\)
- By construction the feedback capacity \(C_{FB} \geq C\), and we may expect that \(C_{FB} > C\) as feedback allows for more reliable transmission

*Definition:* The capacity with feedback, \(C_{FB}\), of a discrete memoryless channel is the supremum of all rates achievable by feedback codes \(X_i(W, Y^{i-1})\).
Feedback capacity

Channel without feedback

Channel WITH feedback

*Theorem: Feedback capacity.* The capacity of a DMC with feedback, $C_{FB}$ is given by:

$$C_{FB} = C = \max_{p(x)} I(X; Y)$$

Source-channel separation

*When are we allowed to design the source and channel coder separately AND remain optimal from an end-to-end perspective?*
Source-channel separation

So far we have $R > H$ (compression) and $R < C$ (transmission). It is natural to ask whether the condition $H < C$ is necessary and sufficient for transmitting a source over a channel:

*Theorem*: Source-channel coding theorem If $V_1, V_2, \ldots, V_n$ is a finite-alphabet stochastic process satisfying the AEP and the condition $H(\mathcal{V}) < C$, then there exists a source-channel code with probability of error $\Pr(\hat{V}^n \neq V^n) \to 0$.

Conversely, if for any stationary stochastic process we have that $H(\mathcal{V}) > C$, then the probability of error is bounded away from zero.

- Important result: source coding and channel coding might as well be done separately since same capacity

Source-channel separation: achievability

- Errors arise from incorrect (i) encoding of $V^n$ or (ii) decoding of $Y^n$
- For $n > N_\epsilon$, there are only $2^{nH(\mathcal{V})+\epsilon} V^n$ sequences in the typical set.
- Encode using $nH(\mathcal{V}) + \epsilon$ bits, encoder error $< \epsilon$
- Transmit with error probability less than $\epsilon$ as long as $R = H(\mathcal{V}) + \epsilon < C$
- Total probability of error $< 2\epsilon$
Source-channel separation: converse

- \[ H(\mathcal{V}) \leq \frac{H(V_1, V_2, \ldots, V_n)}{n} \]
  \[ = \frac{H(V^n)}{n} \]
  \[ = \frac{1}{n} H(V^n | \hat{V}^n) + \frac{1}{n} I(V^n; \hat{V}^n) \]
  \[ \leq \frac{1}{n} (1 + P_{e}^{(n)} n \log |\mathcal{V}|) + \frac{1}{n} I(X^n; Y^n) \]
  \[ \leq \frac{1}{n} + P_{e}^{(n)} \log |\mathcal{V}| + C \]

- Let \( n \to \infty \), we have \( P_{e}^{(n)} \to 0 \) and hence
  \[ H(\mathcal{V}) \leq C \]

Source-channel separation

- For (time-varying) DMC we can design the source encoder and channel coder separately and still get optimum performance

- Not true for:
  - Correlated Channel and Source
  - Multiple access with correlated sources
  - Broadcast channel