A Brief Review of Coding Theory

Pascal O. Vontobel
Information Theory Research Group
Hewlett-Packard Laboratories

USC, Los Angeles, CA, November 10, 2006

Reliable Communication

One of the main motivation for studying coding theory is because one would like to reliably transmit information over noisy channels.

Discrete Memoryless Channels (Part 1)

A simple class of channel models is the class of discrete memoryless channels (DMCs). A DMC is a statistical channel model that is characterized by

- a discrete (or at most countably infinite) input alphabet $\mathcal{X}$
- a discrete (or at most countably infinite) output alphabet $\mathcal{Y}$

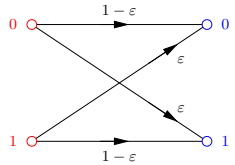
Discrete Memoryless Channels (Part 2)

(list continued)

- a conditional probability mass function (pmf) $P_{Y_i|X_i}(y_i|x_i)$ that tells us the probability of observing the output symbol y_i given that the input symbol x_i was sent
- the fact that the transmission at different time indices is statistically independent, i.e., using $\mathbf{x} \triangleq (x_1, \ldots, x_n)$ and $\mathbf{y} \triangleq (y_1, \ldots, y_n)$ we have

$$P_{\mathbf{Y}|\mathbf{X}}(\mathbf{y}|\mathbf{x}) = \prod_{i=1}^{n} P_{Y_i|X_i}(y_i|x_i)$$

The Binary Symmetric Channel

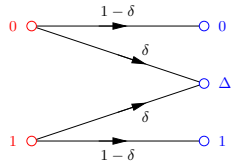


Let $\varepsilon \in [0,1]$. A simple model is e.g. the binary symmetric channel (BSC) with cross-over probability ε. It is a DMC

- with input alphabet $\mathcal{X} = \{0,\ 1\}$,
- with output alphabet $\mathcal{Y} = \{0,\ 1\}$,
- and with conditional probability mass function

$$P_{Y_i|X_i}(y_i|x_i) = \begin{cases} 1-\varepsilon & (y_i = x_i) \\ \varepsilon & (y_i \neq x_i) \end{cases}.$$

The Binary Erasure Channel



Let $\delta \in [0,1]$. Yet another popular model is the binary erasure channel (BEC). It is a DMC

- with input alphabet $\mathcal{X} = \{0,\ 1\}$,
- with output alphabet $\mathcal{Y} = \{0,\ \Delta,\ 1\}$,
- and with conditional probability mass function

$$P_{Y_i|X_i}(y_i|x_i) = \begin{cases} 1-\delta & (y_i = x_i) \\ \delta & (y_i = \Delta) \end{cases}.$$

The Binary-Input AWGNC

Let σ^2 be a non-negative real number. Another popular model (which is strictly speaking not a DMC, though) is the binary-input additive white Gaussian noise channel (AWGNC). It is a memoryless channel model

- with discrete input alphabet $\mathcal{X} = \{0,\ 1\}$,
- with continuous output alphabet $\mathcal{Y} = \mathbb{R}$,
- and with conditional probability density function

$$p_{Y_i|X_i}(y_i|x_i) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{(y_i - \overline{x}_i)^2}{2\sigma^2}\right),$$

where

$$\overline{x}_i \triangleq 1 - 2x_i \triangleq \begin{cases} +1 & (x_i = 0) \\ -1 & (x_i = 1) \end{cases}.$$

Uncoded Transmission

Consider a BSC with cross-over probability $\varepsilon \in [0, 1/2]$.
Assume that we use uncoded transmission, i.e. we directly send the information bits over the BSC.

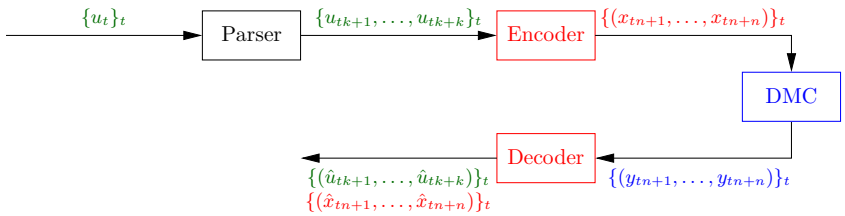
Our best decision about x_i will be

$$\hat{x}_i \triangleq y_i.$$

It is easily seen that the error probability is

$$\Pr\left(\hat{X}_i \neq X_i\right) = \varepsilon.$$

A Better Approach (Part 1)



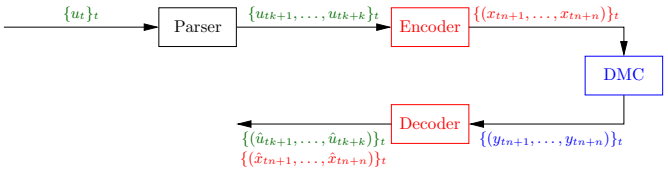
- Firstly, we parse the string of information symbols into blocks of length k.
- Secondly, instead of sending the components of the

  information word $(u_{tk+1},\ldots,u_{tk+k})$

  over the channel, we map (encode) the information word to a

  codeword $(x_{tn+1},\ldots,x_{tn+n})$,

A Better Approach (Part 2)



Based on the

observed channel output $(y_{tn+1},\ldots,y_{tn+n})$

we make a decision

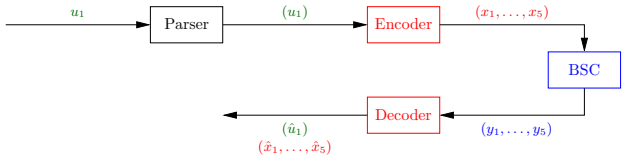
$(\hat{u}_{tk+1},\ldots,\hat{u}_{tk+k})$ about the information vector $(u_{tk+1},\ldots,u_{tk+k})$,

or a decision

$(\hat{x}_{tn+1},\ldots,\hat{x}_{tn+n})$ about the codeword $(x_{tn+1},\ldots,x_{tn+n})$.

A Better Approach (Part 3)

Consider the following en-/de-coding scheme with $\mathcal{U}=\mathcal{X}=\{0,1\}$, $k=1$, and $n=5$ that is used for data transmission over a BSC with cross-over probability $\varepsilon\in[0,\ 1/2]$. (Without loss of generality, we can focus on $t=0$.)



- If $(u_1)=(0)$ then we send the codeword $\mathbf{x}=(0,\ 0,\ 0,\ 0,\ 0)$.
- If $(u_1)=(1)$ then we send the codeword $\mathbf{x}=(1,\ 1,\ 1,\ 1,\ 1)$.
- We use the decoder

$$(\hat{u}_1)=\begin{cases}(0) & \text{if } \mathbf{y} \text{ contains more zeros than ones}\\ (1) & \text{if } \mathbf{y} \text{ contains more ones than zeros}\end{cases}.$$

A Better Approach (Part 4)

- For obvious reasons, the above coding scheme is called a repetition code.
- The rate of the code is $R=k/n=1/5$.
- The error probability is

$$\Pr\left(\hat{U}_1\neq U_1\right)=\binom{5}{3}(1-\varepsilon)^2\varepsilon^3+\binom{5}{4}(1-\varepsilon)^1\varepsilon^4+\binom{5}{5}(1-\varepsilon)^0\varepsilon^5,$$

  which for small ε is clearly smaller than in the uncoded case, but we have to pay for this improvement by sending more symbols over the channel.
- Despite this initial success, one has the feeling that one could construct much better rate-$1/5$ codes by taking k and n larger with $n=5k$.

A Better Approach (Part 5)

- The code (or codebook) is the set of all codewords:

$$\mathcal{C} \triangleq \left\{ \mathbf{x} \in \mathcal{X}^n \;\middle|\; \text{there exists an } \mathbf{u} \in \mathcal{U}^k \text{ s.t. } \mathbf{x} = \text{Encoder}(\mathbf{u}) \right\}$$

- The dimensionless rate of the code is

$$R \triangleq \frac{k}{n}.$$

- The dimensioned rate of the code is

$$R \triangleq \frac{k \log_2 |\mathcal{U}|}{n} \quad \text{[bits per channel use]}.$$

Note that if $|\mathcal{U}| = 2$ then the dimensionless and the dimensioned rate are equal. In the following, we will mostly deal with the case $|\mathcal{U}| = |\mathcal{X}| = 2$ and so we will simply talk about the rate R.

A Better Approach (Part 6)

- An important quantity characterizing a code is the minimum Hamming distance

$$d_{\min}(\mathcal{C}) \triangleq \min_{\substack{\mathbf{x},\ \mathbf{x}' \in \mathcal{C} \\ \mathbf{x} \neq \mathbf{x}'}} d_{\mathrm{H}}(\mathbf{x}, \mathbf{x}'),$$

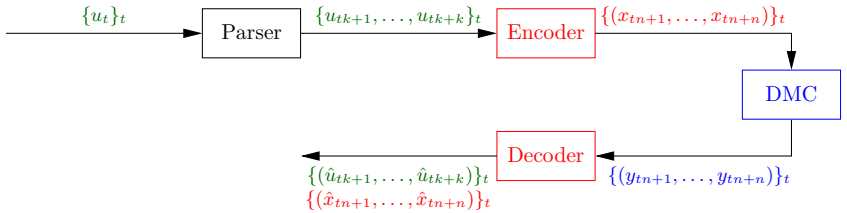
where $d_{\mathrm{H}}(\mathbf{x}, \mathbf{x}')$ is the Hamming distance between $\mathbf{x}$ and $\mathbf{x}'$.

- For a linear block code we have

$$d_{\min}(\mathcal{C}) = \min_{\substack{\mathbf{x} \in \mathcal{C} \\ \mathbf{x} \neq \mathbf{0}}} w_{\mathrm{H}}(\mathbf{x}),$$

where $w_{\mathrm{H}}(\mathbf{x})$ is the Hamming weight of $\mathbf{x}$.

Information Theory (Part 1)



What does information theory tell us about our setup?

$\Longrightarrow$ To every DMC we can associate a number called the capacity C [bits per channel use].

Information Theory (Part 2)

Channel Coding Theorem

- Let the (dimensioned) rate R be such that $R < C$.
- Fix an arbitrary $\epsilon > 0$.
- Then there exists a sequence of encoders/decoders with information word length k_ℓ and block length n_ℓ with

$$R = \frac{k_\ell \log_2(|\mathcal{U}|)}{n_\ell}$$

such that the block error probability fulfills

$$\Pr\left(\left(\hat{U}_1, \ldots, \hat{U}_{k_\ell} \right) \neq (U_1, \ldots, U_{k_\ell}) \right) \quad < \quad \epsilon$$

as $k_\ell \to \infty$ (and therefore as $n_\ell \to \infty$).

Information Theory (Part 3)

Converse to the Channel Coding Theorem

- Let the (dimensioned) rate R be such that $R > C$.
- Then for any sequence of encoders/decoders with information word length k_ℓ and block length n_ℓ with

$$R = \frac{k_\ell \log_2(|\mathcal{U}|)}{n_\ell}$$

the block error probability

$$\Pr\Big(\big(\hat{U}_1, \ldots, \hat{U}_{k_\ell}\big) \neq (U_1, \ldots, U_{k_\ell})\Big)$$

is strictly bounded away from zero for any k_ℓ (and therefore also for any n_ℓ). For more precise statements, see e.g. Cover and Thomas [1].

Information Theory (Part 4)

Note that the channel coding theorem is a purely existential result and is based on the use of so-called random codes, i.e. one can show that the "average random code" is good enough under maximum likelihood (ML) decoding.
A random code can be constructed as follows: the ?-entries in the encoding table below must be filled with randomly selected elements of $\mathcal{X}$. (Here shown for $|\mathcal{U}| = \{0, 1\}$, $k = 3$, and $n = 5$).

(u_1, u_2, u_3)	$(x_1, x_2, x_3, x_4, x_5)$
(0 , 0 , 0)	(? , ? , ? , ? , ?)
(0 , 0 , 1)	(? , ? , ? , ? , ?)
(0 , 1 , 0)	(? , ? , ? , ? , ?)
(0 , 1 , 1)	(? , ? , ? , ? , ?)
(1 , 0 , 0)	(? , ? , ? , ? , ?)
(1 , 0 , 1)	(? , ? , ? , ? , ?)
(1 , 1 , 0)	(? , ? , ? , ? , ?)
(1 , 1 , 1)	(? , ? , ? , ? , ?)

If one wants to generate a sequence of capacity-achieving (c.a.) codes then the ?-entries must be filled with randomly and independently selected elements from $\mathcal{X}$ according to the so-called c.a. input distribution. Moreover, k and n must go to ∞ whereby $R = k \log_2(|\mathcal{U}|)/n$.

Information Theory (Part 5)

For the binary-input AWGNC, the BSC, and the BEC this means that all entries should be randomly and independently chosen such that there are about the same number of zeros and ones.

- However, encoding has extremely high memory complexity because the whole encoding table has to be stored.
- Moreover, ML decoding (or even some sub-optimal decoding) of such a code has extremely high memory and computational complexity.

Encoding/decoding of such random codes of reasonable length and rate is highly impractical.
⇒ We need codes with more structure!
Luckily, the channel coding theorem imposes only small constraints on the codes, i.e. it leaves a lot of freedom in designing good codes.

Coding Theory (Part 1.1)

In order to obtain practical encoding and coding schemes, people have restricted themselves to certain classes of codes that have some structure that can be exploited for encoding/decoding. (Here we only discuss the case $\mathcal{U} = \mathcal{X} = \{0, 1\}$.) (Of course, by restricing oneself to certain classes of codes, it can happen that one loses in performance compared to the the best possible coding scheme where no restriction is imposed on the encoding and decoding complexity.)

- Restriction that the encoding map is linear over $\mathbb{F}_2$.
  - This allows one to use results from linear algebra.
  - Encoding can be characterized by a $k \times n$ matrix $\mathbf{G}$ over $\mathbb{F}_2$:

$$\mathcal{C} = \Big\{\mathbf{x} \in \mathbb{F}_2^n \;\Big|\; \text{there exists an } \mathbf{u} \in \mathbb{F}_2^k \text{ such that } \mathbf{x} = \mathbf{u} \cdot \mathbf{G}\Big\}.$$

    $\mathbf{G}$ is called the generator matrix.
  - The code $\mathcal{C}$ is a k-dimensional subspace of $\mathbb{F}_2^n$. The parameter k is therefore often called the dimension of the code.

Coding Theory (Part 1.2)

- Restriction that the encoding map is linear over $\mathbb{F}_2$ (continued).
  - A rank-$(n-k)$ matrix $\mathbf{H}$ of size $m \times n$ over $\mathbb{F}_2$ such that

$$\mathcal{C} = \left\{ \mathbf{x} \in \mathbb{F}_2^n \mid \mathbf{x} \cdot \mathbf{H}^{\mathsf{T}} = \mathbf{0} \right\}.$$

    is called a parity-check matrix. Note that $m \geq n-k$. (It is clear that for a given code $\mathcal{C}$ there are many possible parity-check matrices.)
  - Some simplifications can be done in the ML decoder.
  - The all-zeros word is always a codeword. For analysis purposes, we can always assume that the all-zeros codeword was sent. (For this statement we assumed that the channel is output-symmetric and that the decoder is symmetric.)

$\Rightarrow$ The resulting codes are called linear block codes.
$\Rightarrow$ A linear code of length n, dimension k, and minimum distance $d_{\min}$ is called an $[n,k]$ binary linear code or an $[n,k,d_{\min}]$ binary linear code.

Coding Theory (Part 2)

- Restriction that the encoding map is linear over $\mathbb{F}_2$ and that cyclic shifts of codewords are again codewords.
  - This allows one to use results from linear algebra and results about polynomials. ($\Rightarrow$ Fundamental theorem of algebra, discrete Fourier transform.)
  - Encoding can be characterized by a monic degree-$(n-k)$ polynomial $g(X) \in \mathbb{F}_2[X]$:

$$\mathcal{C} = \left\{ c(X) \in \mathbb{F}_2[X] \;\middle|\; \begin{array}{l} \text{there exists an } \; u(X) \in \mathbb{F}_2[X] \\ \text{s.t.} \quad \deg(u(X)) < k \;\; \text{and s.t.} \;\; c(X) = u(X) \cdot g(X) \end{array} \right\}.$$

    $g(X)$ is called the generator polynomial.
  - There is a monic degree-k polynomial $h(X) \in \mathbb{F}_2[X]$ such that

$$\mathcal{C} = \left\{ c(X) \in \mathbb{F}_2[X] \mid \deg(c(X)) < n \;\; \text{and s.t.} \;\; c(X) \cdot h(X) = 0 \pmod{X^n - 1} \right\}.$$

    $h(X)$ is called the parity-check polynomial.
  - Encoding can be done very efficiently (especially in hardware).

$\Rightarrow$ The resulting class of codes is called cyclic block codes.

Coding Theory (Part 3)

Some remarks:

- Cyclic block codes have traditionally been one of the most popular classes of codes.
  $\Rightarrow$ Reed-Solomon codes, BCH codes, Reed-Muller codes, etc.
- Within the class of linear block codes there are many special classes, e.g. the class of algebraic-geometry codes. (Here one can use the powerful Riemann-Roch Theorem.)
- etc.

See e.g. the book by MacWilliams and Sloane [2] that contains many results on "traditional" coding theory.

Coding Theory (Part 4)

- Modern coding theory is based on codes that have a sparse graphical representation with small state-space sizes.
  - For such codes, very efficient, although usually suboptimal, decoding algorithms are known (sum-product algorithm decoding, min-sum algorithm decoding, etc.).
  - Designing good codes is about finding graphical representations where these decoding algorithms work well.

"Traditional" vs. "Modern" Coding and Decoding

	Code design		Decoding
"Traditional"	Reed-Solomon codes etc.	→	?
"Modern"	?	→	Iterative decoding (Sum-product algorithm, etc.)

The Law of Large Numbers

The channel coding theorem and many other results in information theory rely on the law of large numbers. That is why coding/decoding works better the longer the codes are. However, in many practical applications one wants to limit delays. So, typically uses codes have block lengths of a few hundreds up to a few thousands (and sometimes a few ten thousands).

"Traditional" vs. "Modern" Coding and Decoding

	Code design		Decoding
"Traditional"	Reed-Solomon codes etc.	→	Berlekamp-Massey decoder etc.
"Modern"	Codes on Graphs (LDPC/Turbo codes, etc.)	→	Iterative decoding (Sum-product algorithm, etc.)

References

[1] T. M. Cover and J. A. Thomas, *Elements of Information Theory*. Wiley Series in Telecommunications, New York: John Wiley & Sons Inc., 1991. A Wiley-Interscience Publication.

[2] F. J. MacWilliams and N. J. A. Sloane, *The Theory of Error-Correcting Codes*. New York: North-Holland, 1977.

[3] J. L. Massey, *Applied Digital Information Theory I and II*. Lecture Notes, ETH Zurich, 1998. Available online under `http://www.isi.ee.ethz.ch/education/public/free_docs.en.html`.


invent