Shannon-inspired research tales
on Duality, Encryption, Sampling and Learning

Kannan Ramchandran
University of California, Berkeley
Shannon’s incredible legacy

- A mathematical theory of communication
- Channel capacity
- Source coding
- Channel coding
- Cryptography
- Sampling theory
- ...

(1916-2001)
And many more...

- Boolean logic for switching circuits (MS thesis 1937)

- Juggling theorem:
  \[ H(F+D) = N(V+D) \]

  - **F**: the time a ball spends in the air,
  - **D**: the time a ball spends in a hand,
  - **V**: the time a hand is vacant,
  - **N**: the number of balls juggled,
  - **H**: the number of hands.

- ...
Story: Shannon meets Einstein

As narrated by Arthur Lewbel (2001)

“The story is that Claude was in the middle of giving a lecture to mathematicians in Princeton, when the door in the back of the room opens, and in walks Albert Einstein.

Einstein stands listening for a few minutes, whispers something in the ear of someone in the back of the room, and leaves. At the end of the lecture, Claude hurries to the back of the room to find the person that Einstein had whispered too, to find out what the great man had to say about his work.

The answer: Einstein had asked directions to the men’s room.”
Outline
Three “personal” Shannon-inspired research stories:

Chapter 1
  Duality between source coding and channel coding – with side-information (2003)

Chapter 2
  Encryption and Compression – swapping the order (2003)

Chapter 3
  Sampling and Learning – Sampling below Nyquist rate and efficient learning (2014)
Chapter 1

Duality
• source & channel coding
• with side-information

Sandeep Pradhan  Jim Chou
Shannon’s celebrated 1948 paper

**The Bell System Technical Journal**

**Vol. XXVII  July, 1948  No. 3**

**A Mathematical Theory of Communication**

**By C. E. SHANNON**

**INTRODUCTION**

The recent development of various methods of modulation such as PM and PPM which exchange bandwidth for signal-to-noise ratios has intensified the interest in a general theory of communication. A basis for such a theory is contained in the important papers of Nyquist 1 and Hartley 2 on this subject. In the present paper we will extend the theory to include a number of new factors, in particular the effect of noise in the channel, and the savings possible due to the statistical structure of the original message and due to the nature of the final destination of the information.

The fundamental problem of communication is that of reproducing at one point either exactly or approximately a message selected at another point. Frequently the messages have meaning; that is, they refer to or are correlated according to some system with certain physical or conceptual entities. These semantic aspects of communication are irrelevant to the engineering problem. This significant aspect is that the actual message is one selected from a set of possible messages. The system must be designed to operate for each possible selection, not just the one which will actually be chosen since this is unknown at the time of design.

If the number of messages in the set is finite then this number or any monotonic function of this number can be regarded as a measure of the information produced when one message is chosen from the set, all choices being equally likely. As was pointed out by Hartley the most natural choice is the logarithmic function. Although this definition must be generalized considerably when we consider the influence of the statistics of the message and when we have a continuous range of messages, we will in all cases use an essentially logarithmic measure.

The logarithmic measure is more convenient for various reasons:

1. It is practically more useful. Parameters of engineering importance

---


**Fig. 1—Schematic diagram of a general communication system.**
Source coding

Entropy of a random variable
= minimum number of bits required to represent the source

\[ H(X) = \mathbb{E}_X \left[ \log \left( \frac{1}{p(X)} \right) \right] \]
Rate-distortion theory - 1948

- Trade-off between *compression rate* and the *distortion*

**PART V: THE RATE FOR A CONTINUOUS SOURCE**

**27. Fidelity Evaluation Functions**

In the case of a discrete source of information we were able to determine a definite rate of generating information, namely the entropy of the underlying stochastic process. With a continuous source the situation is considerably more involved. In the first place a continuously variable quantity can assume an infinite number of values and requires, therefore, an infinite number of binary digits for exact specification. This means that to transmit the output of a continuous source with *exact recovery* at the receiving point requires, in general, a channel of infinite capacity (in bits per second). Since, ordinarily, channels have a certain amount of noise, and therefore a finite capacity, exact transmission is impossible.

This, however, evades the real issue. Practically, we are not interested in exact transmission when we have a continuous source, but only in transmission to within a certain tolerance. The question is, can we assign a definite rate to a continuous source when we require only a certain *fidelity of recovery*, measured in a suitable way. Of course, as the fidelity require-

**Mutual information:**

\[ R(D) = \min_{P_{Y|X}(y|x)} I(X;Y) \]

subject to \( \mathbb{E}[d(X,Y)] \leq D \)

*distortion measure*
Channel coding

• For rates $R < C$, can achieve arbitrary small error probabilities
• Used to be thought one needs $R \to 0$

$C(W) = \max_{P_X(x)} I(X;Y)$
subject to $\mathbb{E}[w(X)] \leq W$
Shannon’s breakthrough

• Communication before Shannon:
  – *Linear filtering* (Wiener) at receiver to remove noise

• Communication after Shannon:
  – Designing codebooks
  – *Non-linear estimation* (MLE) at receiver

Reliable transmission at rates approaching channel capacity
"There is a curious and provocative duality between the properties of a source with a distortion measure and those of a channel. This duality is enhanced if we consider channels in which there is a cost associated with the different input letters, and it is desired to find the capacity subject to the constraint that the expected cost not exceed a certain quantity....."
Shannon (1959)

...This duality can be pursued further and is related to a duality between past and future and the notions of control and knowledge. Thus, we may have knowledge of the past but cannot control it; we may control the future but not have knowledge of it.”
Functional duality

When is the *optimal encoder* for one problem functionally identical to the *optimal decoder* for the dual problem?
Duality example: Channel coding

You want to send message $m$: how big can you make $R$?

**Shannon’s result:**

$C_{BEC} = (1-p)$ bits per channel use

$p = 0.2$

$Cost (0) = 1$ ; $Cost (1) = 1$

Total budget $\leq 10,000$
What is the Shannon capacity?

- Encoder
- Decoder

$m \rightarrow \text{Encoder} \rightarrow 0 \overset{0.8}{\rightarrow} 0 \overset{0.2}{\rightarrow} * \overset{0.2}{\rightarrow} 1 \overset{0.8}{\rightarrow} 1 \rightarrow \text{Decoder} \rightarrow \hat{m}$

The decoder knows which bits are erased (channel output)

Suppose the encoder also knows which bits are erased (genie)

Send information in non-erased locations

- Number of non-erased bits
  - $\approx 10,000 \times (1 - p)$
  - $= 10,000 \times 0.8$
  - $= 8,000$

Surprise: the encoder does not need to know which bits are erased!
1) **Encoder & Decoder agree on a random codebook**

Shannon’s random coding argument

2) **Encoder encodes message**

*Output the codeword corresponding to the index*

3) **Decoder decodes message**

*Output the index corresponding to the closest codeword*
**Why does it work?**

**Codebook for channel coding**

- **Successful decoding if the non-erased string is unique**
- **8,000 bits will induce unique match if (random) codebook size is \( \leq 2^{8,000} \) w.h.p.**

**Table:**

<table>
<thead>
<tr>
<th>Input</th>
<th>Erased Locations</th>
</tr>
</thead>
<tbody>
<tr>
<td>1001000010101000...</td>
<td>1001000010101000...</td>
</tr>
<tr>
<td>1111011111101110...</td>
<td>1111011111101110...</td>
</tr>
<tr>
<td>1110000111001110...</td>
<td>1110000111001110...</td>
</tr>
<tr>
<td>...</td>
<td>...</td>
</tr>
<tr>
<td>1101011001010010...</td>
<td>1101011001010010...</td>
</tr>
</tbody>
</table>

**Diagram:**

- Say sending \( m = 3 \)
- \( 2^{8,000} \)
- 10,000
- IID random B(1/2) entries
- Input to the channel
- Channel will erase 20% of bits
- 8,000
- 2,000
- \( \leq 2^{8,000} \) w.h.p.
Source Coding Dual to the BEC: BEQ

$X \in \{0, 1, *\}^{10,000}$

01*1*00110...

Compressed bit-stream 8,000 bits

Want the average distortion to be $\leq 0.2$

$p(0) = p(1) = 0.4$;
$p(*) = 0.2$

d$(x, \hat{x}) = \begin{cases} 
0 & \text{if } \hat{x} = x \text{ for } x \in \{0, 1\} \\
\infty & \text{if } \hat{x} \neq x \text{ for } x \in \{0, 1\} \\
1 & \text{if } x = * 
\end{cases}$

* is like a “don’t care” symbol (e.g., perceptually masked symbols). How can we exploit this for compression?

Martinian and Yedidia, 2004
Source Coding Dual to the BEC: BEQ

The decoder does not need to know which symbols are ‘*’!

Send the non-* bits:

01100110...

\[ R_{BEQ}(0.2) \geq 0.8 \text{ bits/symbol} \]

Number of non ‘*’ symbols to send

\[ \approx 10,000 \times (1 - p(\ast)) \]
\[ = 10,000 \times 0.8 = 8,000 \]

Suppose the decoder also knows which are the ‘*’ symbols (genie)

\[ p(0) = p(1) = 0.4 \quad p(\ast) = 0.2 \]

<table>
<thead>
<tr>
<th>Source Encoder</th>
<th>Source Decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>( \hat{X} )</td>
<td>( X )</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Source</th>
<th>Encoder</th>
<th>Decoder</th>
</tr>
</thead>
<tbody>
<tr>
<td>( X )</td>
<td>( 01\ast 1*00110... )</td>
<td>( \hat{X} )</td>
</tr>
</tbody>
</table>

\( p(\ast) = 0.2 \)
Source Coding Dual to the BEC: BEQ

String Length 10,000

Source Encoder

Compressed bitstream 8,000 bits

Source Decoder

Want the average distortion to be ≤ 0.2

How would you do it?

Use channel decoder as source encoder

Use channel encoder as source decoder

$p(0) = p(1) = 0.4$;
$p(*) = 0.2$
Shannon’s prescription: random coding

1) Encoder & Decoder agree on a random codebook
Shannon’s random coding argument

2) Encoder encodes message
Output the codeword corresponding to the index
Output the index corresponding to the closest codeword

3) Decoder decodes message
Output the index corresponding to the closest codeword
Output the codeword corresponding to the index
Why does it work?

- **Successful encoding if the “non-*” part of input string is present in the codebook.**
- **8,000 bits will induce an exact match if random codebook size is \( \geq 2^{8,000} \) w.h.p.**

IID random 
B(1/2) entries

---

Codebook for source coding

**Index of the codeword that exactly matches the non-* part of input string**

- Bitstream of length 10,000
  - \( p(0) = p(1) = 0.4 \)
  - \( p(*) = 0.2 \)

---

locations with *

**2,000**

---

10,000

---

28,000
Knowledge of the erasure pattern

Channel coding

The encoder does not need to know the don’t care locations.
The decoder knows the erasure pattern.

Source coding

The encoder knows the don’t care locations.
The decoder does not need to know the don’t care locations.
Duality between source and channel coding

Given a source coding problem with source distribution $q(x)$, optimal quantizer $p^*(\hat{x}|x)$, distortion measure $d(x, \hat{x})$ and distortion constraint $D$

There is a dual channel coding problem with channel $p^*(x|\hat{x})$ cost measure $w(\hat{x})$ and cost constraint $W$ such that

$$R(D) = C(W)$$

$$w(\hat{x}) = c_1 D(p^*(x|\hat{x}) || q(x)) + \theta$$

$$W = E_{p^*(\hat{x})} w(\hat{x}).$$

Pradhan, Chou and R, 2003
For any given source coding problem, there is a dual channel coding problem such that:

• both problems induce the same optimal joint distribution
• the optimal encoder for one is functionally identical to the optimal decoder for the other
• an appropriate channel-cost measure is associated

Key takeaway

Source coding  
*distortion measure* is as important as the *source distribution*

Channel coding  
*channel cost measure* is as important as the *channel conditional distribution*
Duality between

source coding with side information

and

channel coding with side information
Source coding with side information (SCSI):

• (Only) decoder has access to side-information $S$

• Studied by Slepian-Wolf ‘73, Wyner-Ziv ’76, Berger ’77

• Applications: sensor networks (IoT), digital upgrade, secure compression.

• No performance loss in some important cases
(Only) encoder has access to "interfering" side-information $S$

- Studied by Gelfand-Pinsker ‘81, Costa ‘83, Heegard-El Gamal ‘85

- Applications: data hiding, watermarking, precoding for known interference, writing on dirty paper, MIMO broadcast.

- No performance loss in some important cases
Channel coding with side information (CCSI):

- Encoder (only) has access to "interfering" side-information $S$
- Studied by Gelfand-Pinsker '81, Costa '83, Heegard-El Gamal '85
- Applications: data hiding, watermarking, precoding for known interference, writing on dirty paper, MIMO broadcast.
- No performance loss in some important cases
Duality between *source coding* & *channel coding with side information*

**Source coding with side information (SCSI)**

- **Source** → **Encoder** → bits → **Decoder** → bits → **Quantized Source**
- Internet of Things (IoT), video streaming, multiple description coding, secure compression

**Channel coding with side information (CCSI)**

- bits → **Encoder** → **Decoder** → bits
- Side-information → **Encoder** → Channel input → **Decoder** → Channel output
- Watermarking, data hiding, multi-antenna wireless broadcast

*Pradhan, Chou and R, 2003*
Chapter 2

Cryptography

- Compressing encrypted data

Mark Johnson  Prakash Ishwar  Vinod Prabhakaran
Cryptography – 1949

• Foundations of *modern cryptography*
• All theoretically unbreakable ciphers must have the properties of one-time pad

---

**Communication Theory of Secrecy Systems**

By C. E. SHANNON

1. Introduction and Summary

The problems of cryptography and secrecy systems furnish an interesting application of communication theory. In this paper a theory of secrecy systems is developed. The approach is on a theoretical level and is intended to complement the treatment found in standard works on cryptography. There, a detailed study is made of the many standard types of codes and ciphers, and of the ways of breaking them. We will be more concerned with the general mathematical structure and properties of secrecy systems.
Compressing Encrypted Data

“Correct” order

Source $X$ → Compress → $H(X)$ bits → Encrypt → $H(X)$ bits

Wrong order?

Source $X$ → Encrypt → $Y$ → Compress → $H(X)$ bits

Johnson & R, 2003
Example

Original Image → 10,000 bits → Encrypted Image

Encrypted Image → 5,000 bits → Compressed Encrypted Image

Decoding Compressed Image

Final Reconstructed Image
Key Insight!

- \( Y = X + K \) where \( X \) is independent of \( K \)
- **Slepian-Wolf theorem**: can send \( X \) at rate \( H(Y|K) = H(X) \)
X is uniformly chosen from \{[000], [001], [010], [100]\}

K is a length-3 random key (equally likely in \{0,1\}^3)

Correlation: Hamming distance between Y and K at most 1

Example: when \(K = [0 1 0]\), Y \(\Rightarrow\) [0 1 0], [0 1 1], [0 0 0], [1 1 0]

Y = X + K

Case 1

- Encoder computes \(X = Y + K \mod 2\)
- Encoder represents \(X\) using 2 bits
- Decoder outputs \(X \mod 2\)
Transmission at 2 bits/sample

Encoder => send index of the coset containing X.

Decoder => find a codeword in given coset closest to K

Example: Y=010 (K=110) => Encoder sends message 10
$Y$ (encrypted)

$Y = X + K$

$X$ (unencrypted & compressible)
Example: geometric illustration

Encoder \( X \)  \( m \)  \( m \)  \( \hat{X} \)  \( K \)

Decoder

\( K \)

Side information

\( X \)
Practical Code Constructions

• Use a linear transformation (hash/bin)
• Design cosets to have maximal spacing
  – State of the art linear codes (LDPC codes)
• Distributed Source Coding Using Syndromes (DISCUS)*

* Pradhan & R, ’03
Chapter 3

Sampling theory

• Sample and compute efficient sampling (and connections to learning)

Orhan Ocal

Xiao Li
Sampling theorem

Communication in the Presence of Noise

CLAUDE E. SHANNON, MEMBER, IRE

Theorem 1: If a function $f(t)$ contains no frequencies higher than $W$ cps, it is completely determined by giving its ordinates at a series of points spaced $1/2 W$ seconds apart. pointwise sampling!

Mathematically, this process can be described as follows. Let $x_n$ be the $n$th sample. Then the function $f(t)$ is represented by

$$f(t) = \sum_{n=-\infty}^{\infty} x_n \frac{\sin \pi (2Wt - n)}{\pi (2Wt - n)}.$$  (7)

linear interpolation!
Aliasing phenomenon

### Time domain

**Input signal**

- **Sampling at rate 1**
- **Sampling at rate 1/2**

### Frequency domain

**Bandwidth of 1 Hz**

- **No aliasing** – can recovery by linear filtering
- **Spectrum is aliased!**
But what if the spectrum is sparsely occupied?

Henry Landau, 1967

- Know the frequency support
- Sample at rate “occupied bandwidth” $f_{\text{occ}}$ (Landau rate)

When you do not know the support?

- Feng and Bresler, 1996
- Lu and Do, 2008
- Mishali, Eldar, Dounaevsky and Shoshan, 2011
- Lim and Franceschetti, 2017
Filter bank approach

Input in frequency domain

Know the frequency support, filter and sample

Filtering

Sampling spectrum-blind?
Requires $2f_{oc}$. Can we design a constructive scheme?

Lu and Do, 2008
Puzzle: Gold thief

- One unknown thief
- Steals unknown but fixed amount from each coin
- What is min. no. of weighings needed?
  - 2 are enough!

100 grams each

Puzzle: Gold thief

\[
\begin{bmatrix}
1 & 1 & 1 & 1 & 1 & 1 \\
1 & 2 & 3 & 4 & 5
\end{bmatrix}
\begin{pmatrix}
-5 \\
-20
\end{pmatrix}
= 
\begin{pmatrix}
y_1 \\
y_2
\end{pmatrix}
\]

Ratio-test identifies the location
4-thieves among 12-treasurers

Key Ideas:
1. Randomly group the treasurers.
2. If there is a single thief problem
   ✓ Ratio test
   ✓ Iterate.

Questions:
1. How many groups needed?
2. How to form groups?
3. How to identify if a group has a single thief?
Main result

Any bandlimited signal $x(t) \in \mathbb{C}$ whose spectrum has occupancy $f_{occ}$ can be sampled asymptotically at rate $f_s = 2f_{occ}$ by a randomized “sparse-graph-coded filter bank” with probability 1 using $O(f_{occ})$ operations per unit time.

Remarks

- Computational cost $O(f_{occ})$ independent of bandwidth
- Requires mild assumptions (genericity)
- Can be made robust to sampling noise
Key insight for spectrum-blind sampling

- To reduce sampling rate, *subsample judiciously*
- *Filter bank* derived from *capacity-achieving codes for the Binary Erasure Channel (BEC)* (LDPC codes)
- Introduces aliasing (*structured noise*)
- *Non-linear recovery* instead of linear interpolation
Filter bank for sampling

- Sample the signal at rate $B$

- Filter and then sample at rate $B$
Filter bank for sampling

Aggregate sampling rate: \( N \frac{f_M}{N} = f_M = \text{Nyquist rate for } x(t) \)
‘Sparse-graph-coded’ filter bank

\[ Y(Bf) = \mathbf{X}(Bf) \mathbf{B} \]

where

\[ \mathbf{X}(f) = \begin{pmatrix} X_0(f) \\ \vdots \\ X_{N-1}(f) \end{pmatrix} \]

\[ \mathbf{B} = \begin{pmatrix} \mathbf{b}_0 & \mathbf{b}_1 & \cdots & \mathbf{b}_{N-1} \end{pmatrix} \]

\[ \mathbf{b}_m = \begin{pmatrix} 0 & 0 & 0 & \cdots & 0 & 1 & 0 & \cdots & 0 \end{pmatrix} \]
Example — sparse graph underlying the measurements

Sparse bipartite graph
Example — sparse graph underlying the measurements

visual cleaning for presentation:
remove edges that connect to non-active bands
Example — peeling

Measurement classification

- zero-ton: no signal
- single-ton: no aliasing
- multi-ton: aliasing

bands

channels

$X_0(f)$
$X_1(f)$
$X_2(f)$
$X_3(f)$
$X_4(f)$
$X_5(f)$
$X_6(f)$
$X_7(f)$
$X_8(f)$
$X_9(f)$
Example — peeling

Assume a mechanism:
identifies which channels have no aliasing (here B and F) and maps them to which bands they came from (here 1 and 4 resp.)

Measurement classification

- **zero-ton:** no signal
- **single-ton:** no aliasing
- **multi-ton:** aliasing

bands | channels
--- | ---
$X_0(f)$ | A
$X_1(f)$ | C
$X_2(f)$ | D
$X_3(f)$ | E
$X_4(f)$ | B
$X_5(f)$ | 
$X_6(f)$ | 
$X_7(f)$ | 
$X_8(f)$ | F
$X_9(f)$ | 

bands | channels
--- | ---
A | $X_0(f)$
C | $X_1(f)$
D | $X_3(f)$
B | $X_4(f)$
E | $X_7(f)$
F | $X_8(f)$

bands | channels
--- | ---
1 | A
4 | F
2 | B
3 | 

Example — peeling

**mechanism:**
identifies which channels have no aliasing and maps them to which bands they came from

**output:**
- channel B: (red, index = 1)
- channel F: (blue, index = 4)
Example — peeling

**mechanism:**
identifies which channels have no aliasing and maps them to which bands they came from

**output:**
channel B: (red, index = 1)
channel F: (blue, index = 4)

*peel from channels they alias into!*
Example — peeling

**mechanism:**
identifies which channels have no aliasing and maps them to which bands they came from
Example — peeling

mechanism:
identifies which channels have no aliasing and maps them to which bands they came from

output:
channel D: (green, index = 8)
channel E: (cyan, index = 5)
Example — peeling

**mechanism:**
identifies which channels have no aliasing and maps them to which bands they came from

**output:**
- channel D: (green, index = 8)
- channel E: (cyan, index = 5)

*peel from channels they alias into!*
Example — peeling

**mechanism:**
identifies which channels have no aliasing and maps them to which bands they came from

*signal is completely recovered!*
Construction of the sparse-graph code

- Designed through capacity-approaching sparse-graph codes
- Connect each band to channels at random according to a carefully chosen degree distribution.
- Asymptotically, number of channels is \((1 + \epsilon)\) times the number of active bands

\[
P(\text{degree} = j) \propto \frac{1}{j-1} \quad \text{for } j=2,3,\ldots,D
\]

\[D > 1/\epsilon\]
Realizing the *mechanism*

Identify which channels have no aliasing and map them to bands

- **same magnitude response**
- **‘stairs’** phase response

H1(f) and H2(f) diagrams with magnitude and phase plots:

- **Identifies dark blue band as a singleton**
Numerical experiment

- Lebesgue measure $f_L = 0.1$
- Number of slices $N = 1000$
- Number of channels $M = 284$
- Sampling rate $f_S = 0.284$
Interesting connection

- **Minimum-rate spectrum-blind sampling**
- **Coding theory and sampling theory**
  - Capacity-approaching codes for erasure channels
  - Filter banks that approach Landau rate for sampling
"Peeling-based" turbo engine

Sparse-Graph Code

Divide

Concur

"Solve-if-trivial" sub-engine
Broad scope of applications

- Sparse-graph codes
  - Sub-Nyquist sampling theory
  - Sparse Spectrum (DFT/WHT)
  - Fast neighbor discovery for IoT (group testing)
  - Compressed sensing
  - Compressive phase retrieval
  - Sparse mixed linear regression

Authors and years:
- Pawar, R., 2013
- Li, Pawar, R., 2014
- Li, R., 2016
- Ocal, Li, R., 2016
- Lee, Pedarsani, R., 2015
- Yin, Pedarsani, Chen, R., 2016
Conclusion: Shannon’s incredible legacy

- A mathematical theory of communication
- Channel capacity
- Source coding
- Channel coding
- Cryptography
- Sampling theory
- …

*His legacy will last many more centuries!*

(1916-2001)
Thank you!