Introduction

Shashank Vatedka

EE2340/5847

Agenda

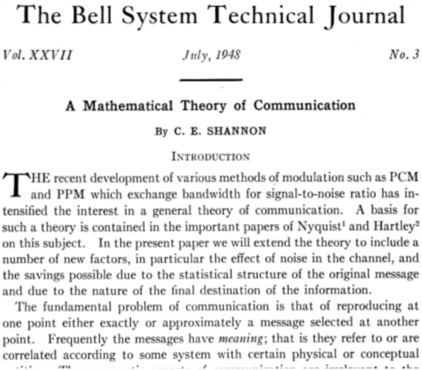
- Logistics
- Why should you study this course?
- Randomness and information
- Probability refresher

Why this course?

- Understand fundamental limits of processing information
- Basic for designing communication systems, compression algorithms, statistics

Where is information theory used?

Data compression and communication



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 PCM and PPM which exchange bandwidth for signal-to-noise ratio 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

Where is information theory used?

Cryptography

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.[1] 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.[2] 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.

Where is information theory used?

Machine learning

Deep Learning and the Information Bottleneck Principle

Naftali Tishby[1,2]   Noga Zaslavsky[1]

***Abstract*— Deep Neural Networks (DNNs) are analyzed via the theoretical framework of the information bottleneck (IB) principle. We first show that any DNN can be quantified by the mutual information between the layers and the input and output variables. Using this representation we can calculate the optimal information theoretic limits of the DNN and obtain finite sample generalization bounds. The advantage of getting closer to the theoretical limit is quantifiable both by the generalization bound and by the network's simplicity. We argue that both the optimal architecture, number of layers and features/connections at each layer, are related to the bifurcation points of the information bottleneck tradeoff, namely, relevant compression of the input layer with respect to the output layer. The hierarchical representations at the layered network naturally correspond to the structural phase transitions along the information curve. We believe that this new insight can lead to new optimality bounds and deep learning algorithms.**

I. INTRODUCTION

output. The information theoretic interpretation of minimal sufficient statistics [5] suggests a principled way of doing that: find a maximally compressed mapping of the input variable that preserves as much as possible the information on the output variable. This is precisely the goal of the Information Bottleneck (IB) method [6].

Several interesting issues arise when applying this principle to DNNs. First, the layered structure of the network generates a successive Markov chain of intermediate representations, which together form the (approximate) sufficient statistics. This is closely related to successive refinement of information in Rate Distortion Theory [7]. Each layer in the network can now be quantified by the amount of information it retains on the input variable, on the (desired) output variable, as well as on the predicted output of the network. The

Where is information theory used?

Physics

Statistical Physics and Information Theory

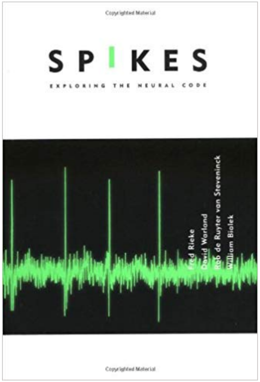
Neri Merhav

Department of Electrical Engineering
Technion — Israel Institute of Technology
Haifa 32000
Israel
merhav@ee.technion.ac.il

now
the essence of knowledge
Boston – Delft

Where is information theory used?

Biology



Where is information theory used?

Computer science

Theorem 1 (Brégman [3]) *Let $G = (A, B, E)$ be a bipartite graph with $|A|, |B| = n$. Then, the number of perfect matchings in G is at most*

$$\prod_{v \in A} (d(v)!)^{1/d(v)}.$$

Theorem 7.1 (Alon-Hoory-Linial [3])**.** *Let G be a graph on n vertices with average degree d and girth $g = 2r + 1$. Then*

$$n \geq 1 + d \sum_{i=0}^{r-1} (d-1)^i.$$

Compression

Given: English text file of size 1GB.

Compression

Given: English text file of size 1GB.

Can we compress this down to 1MB?

Compression

Given: English text file of size 1GB.

Can we compress this down to 1MB?

What is the minimum compression ratio we can achieve for a given file?

Compression

Given: English text file of size 1GB.

Can we compress this down to 1MB?

What is the minimum compression ratio we can achieve for a given file?

Fun experiment: Take any large english text file ($\gtrapprox$ 1MB. e.g., from gutenberg.com) and zip it. Find compression ratio.

Communication

Given: SNR = $0.1dB$, bandwidth = 10 MHz.

Communication

Given: SNR = $0.1dB$, bandwidth = 10 MHz.

Can we design a communication system that can guarantee reliable communication at 4 Mbps?

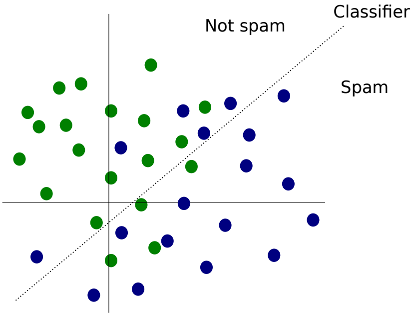
Communication

Given: SNR = 0.1dB, bandwidth = 10 MHz.

Can we design a communication system that can guarantee reliable communication at 4 Mbps?

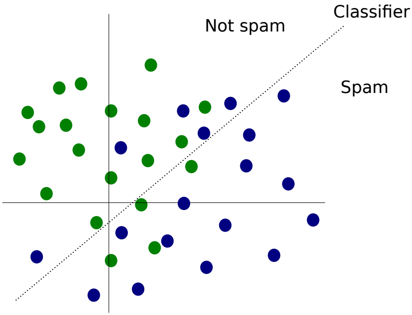
What is the maximum rate of reliable communication for a given SNR?

Classification/detection



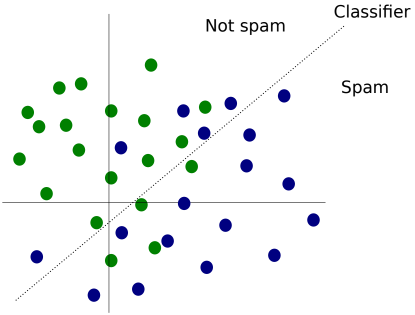
- **Problem:** Classify emails as spam/non-spam

Classification/detection



- **Problem:** Classify emails as spam/non-spam
- Types of error: (1) S given NS, (2) NS given S

Classification/detection



- **Problem:** Classify emails as spam/non-spam
- Types of error: (1) S given NS, (2) NS given S
- Minimize probability of (2) having fixed probability of (1)

The beginnings of information theory

The challenges of long-distance telecommunication:

- Attenuation: inverse square law
- Noise

The beginnings of information theory

The challenges of long-distance telecommunication:

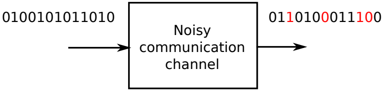
- Attenuation: inverse square law
- Noise

Proposed solutions:

- Increase signal power: not cost effective
- Repeaters: amplifies both signal and noise
- Sophisticated signal processing techniques: still limited

The communication problem

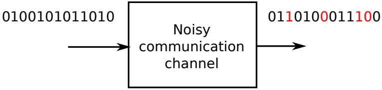
Transmit sequence of k bits, where each bit is corrupted (flipped) independently with probability $p \in (0, 1)$.



- Probability of bit error: $P_e^{\mathrm{bit}} = p$
- Solution: Coding

The communication problem

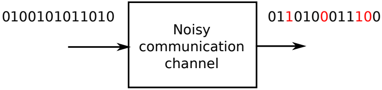
Transmit sequence of k bits, where each bit is corrupted (flipped) independently with probability $p \in (0, 1)$.



- Probability of bit error: $P_e^{\mathrm{bit}} = p$
- Solution: Coding
- Repetition code: $P_e^{\mathrm{bit}} = ?$

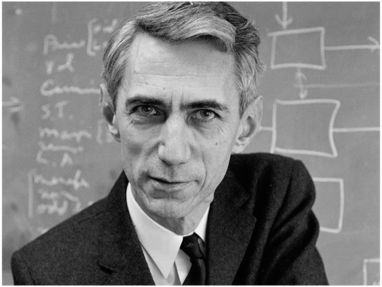
The communication problem

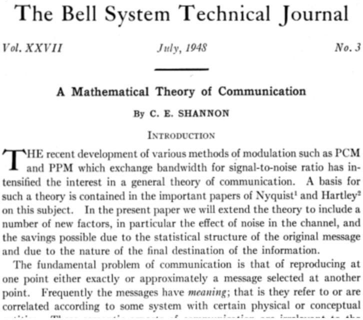
Transmit sequence of k bits, where each bit is corrupted (flipped) independently with probability $p \in (0, 1)$.



- Probability of bit error: $P_e^{\text{bit}} = p$
- Solution: Coding
- Repetition code: $P_e^{\text{bit}} = ?$
- $$R \stackrel{\text{def}}{=} \frac{k}{n}$$
- Want R to be as large as possible

The solution





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 PCM and PPM which exchange bandwidth for signal-to-noise ratio 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

"Consider a LONG message sequence"

$k \rightarrow \infty$

Models

- Source/noise is random
- Asymptotics: $k, n \to \infty$

How do we quantify information

Text prediction: What is the next letter in the sequence

- HELL-
- Q-
- A-

Randomness = Uncertainty

Information = reduction in uncertainty

Randomness and data compression

More random $\implies$ not easily compressible

`./code/generate_randomsequence.py`

What is this course really about?

Three quantities:

- Entropy
- Mutual information
- KL divergence/relative entropy

properties, consequences.

Probability

Refresher

- Random variable: discrete, continuous
- Probability mass function of a discrete rv
- Probability density function of a continuous rv
- Common rvs: Bernoulli, binomial, exponential, Gaussian
- Variance, standard deviation
- Higher order moments, moment generating function

Union bound

Lemma

If $\mathcal{E}_1$ and $\mathcal{E}_2$ are two events, then

$$\Pr[\mathcal{E}_1 \bigcup \mathcal{E}_2] \leqslant \Pr[\mathcal{E}_1] + \Pr[\mathcal{E}_2].$$

Markov inequality

Lemma

Suppose that X is a nonnegative random variable, and $\mathbb{E}X = \mu > 0$. Then, for all $a > 0$, we have

$$\Pr[X \geqslant a] \leqslant \frac{\mu}{a}.$$

Chebyshev's inequality

Lemma

Suppose that X is a random variable with mean μ and variance σ^2. Then, for any $a > 0$, we have

$$\Pr[|X - \mu| \geqslant a] \leqslant \frac{\sigma^2}{a^2}$$

Chernoff bound

Lemma

If X is a random variable with mean μ, then for every $a > 0$, we have

$$\Pr[X \geqslant \mu + a] \leqslant \min_{t>0} \frac{\mathbb{E}e^{t(X-\mu)}}{e^{ta}}$$

$$\Pr[X \leqslant \mu - a] \leqslant \min_{t>0} \frac{\mathbb{E}e^{-t(X-\mu)}}{e^{-ta}}$$

Random iid sequences

$X^n \stackrel{\text{def}}{=} (X_1, X_2, \ldots, X_n)$ is said to be $\sim$ iid p_X if

$$\Pr[X^n = x^n] = \prod_{i=1}^{n} p_X(x_i)$$

Sometimes called a memoryless source.

Chernoff bound for Bernoulli rvs

Lemma

If X^n is an iid random sequence with Bernoulli(p) components, then

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \geqslant p(1+\delta)\right] \leqslant \left(\frac{e^{\delta}}{(1+\delta)^{1+\delta}}\right)^{np} \leqslant e^{-\frac{\delta^2 np}{3}}$$

$$\Pr\left[\frac{1}{n}\sum_{i=1}^{n} X_i \leqslant p(1-\delta)\right] \leqslant \left(\frac{e^{-\delta}}{(1-\delta)^{1-\delta}}\right)^{np} \leqslant e^{-\frac{\delta^2 np}{3}}$$

for any $0 \leqslant \delta \leqslant 1$.

Markov chain

X^n is a first-order time-homogeneous Markov chain with transition probabilities $p_{X'|X}$ and initial distribution p_X if

$$\Pr[X^n = x^n] = p_X(x_1) \prod_{i=2}^{n} p_{X'|X}(x_i | x_{i-1})$$

Also called a first-order Markov source.

Properties of a Markov source

- Conditioned on the present, the future is independent of the past, i.e.,

$$\text{given } X_i, \qquad X_{i+k} \perp\!\!\!\perp (X_1, \ldots X_{i-1})$$

  for all i, k.

- If P denotes the transition probability matrix, then the stationary distribution is a pmf π such that

$$\pi P = \pi$$

k-th order Markov source

X^n is a k-th order Markov source with initial distribution p_{X^k} and transition probabilities $p_{X_{k+1}|X^k}$ if

$$p_{X^n}(x^n) = p_{X^k}(x_1, \ldots, x_k) \prod_{i=k+1}^{n} p_{X_{k+1}|X^k}(x_i | x_{i-1}, \ldots, x_{i-k})$$

No long-range dependencies in the source.