A quick introduction to information theory

Natasha Devroye
Assistant Professor
University of Illinois at Chicago
http://www.ece.uic.edu/~devroye

History of (wireless) communications

Smoke signals

Maxwell’s equations

Marconi demonstrates wireless telegraph

Detroit police cars radio dispatch in 1925

Armstrong demonstrates FM radio
History of (wireless) communications

- Smoke signals
- Maxwell's equations
- Marconi
- Detroit police
- FM radio

State of communications ~ 1930s
- mostly analog
- ad-hoc engineering, tailored to each application

Big Open Questions
- is there a general methodology for designing communication systems?
- can we communicate reliably in noise?
- how fast can we communicate?

Information theory - what, why, when


Bits vs. randomness

Information theory’s famous metrics

- Entropy $H(X)$
  - quantifies the amount of information, or randomness, in a source $X$
  - Ultimate data compression limit is the source’s entropy $H(X)$

- Mutual information $I(X;Y)$
  - quantifies how much knowledge of one of the random variables $X,Y$ can tell you about the other
  - Ultimate transmission rate is the maximal mutual information

Information theory’s claims to fame

- Source coding
  - Source = random variable
  - Ultimate data compression limit is the source’s entropy $H(X)$

- Channel coding
  - Channel = conditional distributions
  - Ultimate transmission rate is the channel capacity $C$

Reliable communication possible $\iff H(X) < C$

Technology independent limits!

Source vs. channel coding

- Source
  - Encoder
  - Channel
  - Source
  - Channel
  - Noise
  - Decoder
  - Destination

- Remove redundancy
- Controlled adding of redundancy
- Decode signals, detect/correct errors
- Restore source
Main result in source-coding/compression

- A source $X$ which outputs source symbols i.i.d. according to the probability mass function $p_X(x)$ may be compressed to $H(X)$ bits/source symbol.

**Definition:** The entropy $H(X)$ of a discrete random variable $X$ with pmf $p_X(x)$ is given by

$$H(X) = -\sum_x p_X(x) \log p_X(x) = -\mathbb{E}_{p_X(x)}[\log p_X(X)]$$

Order these in terms of entropy
Entropy of a random variable $H(X)$

(A) entropy is the measure of **average uncertainty** in the random variable

(B) entropy is the **average number of bits** needed to describe the random variable

(C) entropy is measured in bits?

(D) $H(X) = -\sum_x p(x) \log_2(p(x))$

(E) entropy of a deterministic value is 0

---

**Examples of codes**

*Example: (pg. 104) Let $X$ be a random variable with the following distribution and codeword assignment:

<table>
<thead>
<tr>
<th>Symbol</th>
<th>Probability $\Pr(x)$</th>
<th>Codeword</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>0.5</td>
<td>C(1) = 0</td>
</tr>
<tr>
<td>2</td>
<td>0.25</td>
<td>C(2) = 10</td>
</tr>
<tr>
<td>3</td>
<td>0.125</td>
<td>C(3) = 110</td>
</tr>
<tr>
<td>4</td>
<td>0.125</td>
<td>C(4) = 111</td>
</tr>
</tbody>
</table>

- **Decode 0110111100110**
- **What is $H(X)$?**
  - $\frac{1}{2}\log(2) + \frac{1}{4}\log(4) + \frac{1}{8}\log(8) + \frac{1}{8}\log(8) = 1.75$ bits
- **What is the expected codeword length $L(C)$?**
  - $\frac{1}{2} \times 1 + \frac{1}{4} \times 2 + \frac{1}{8} \times 3 + \frac{1}{8} \times 3 = 1.75$ bits

---

**12 balls weighing: 1 lighter or heavier**

- Total information contained?
- Each weighing gives you how much information (ideally)?
- Number of weighings needed?
- Strategy?

---

**Main result 1: data compression**

*Theorem: Data Compression* Let $X \sim f(p(x))$ and let $\epsilon > 0$. Then there exists a code that maps sequences $x^n$ of length $n$ into binary strings such that the mapping is one-to-one (and therefore invertible) and

$$L(C) = \mathbb{E} \left[ \frac{1}{n} I(X^n) \right] < H(X) + \epsilon$$

for $n$ sufficiently large.

---

You are given 12 balls, all equal in weight except for one that is either heavier or lighter. You are also given a two-pan balance to use. In each use of the balance you may put any number of the 12 balls on the left pan, and the same number on the right pan, and push a button to initiate the weighing; there are three possible outcomes: either the weights are equal, or the balls on the left are heavier, or the balls on the left are lighter. Your task is to design a strategy to determine which is the odd ball and whether it is heavier or lighter than the others in as few uses of the balance as possible.
Main idea

- Code over $n$ symbols (i.e., $X^n$) rather than symbol-by-symbol
- as $n \to \infty$ only certain “typical” sequences occur
- count the number of such “typical” sequences, each gets a codeword
- turns out there are about $2^{nH(x)}$ “typical” sequences, each about equally likely, so we need $nH(X)$ bits to encode $X^n$.

Definition: weak typicality

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

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

Counting the # in the typical set

**Weak Law of Large Numbers + the AEP**

- Let $X_1, X_2, \ldots$, be i.i.d distributed with mean $\mu$ and variance $\sigma^2 < \infty$. Let
  
  $S_n = \frac{1}{n} [X_1 + X_2 + \ldots + X_n]$

- **Theorem: Weak Law of Large Numbers**
  
  $S_n \to \mu$ in probability

- **Theorem: Asymptotic Equipartition Property (AEP):**
  
  If $X_1, X_2, \ldots \sim p(x)$, then

  $\frac{1}{n} \log p(x_1, x_2, \ldots, x_n) \to H(X)$ in probability.

Strong versus Weak Typicality

- Intuition behind typicality?

  - $X = (\bullet, \blacklozenge, \overline{\blacklozenge}, \blacklozenge)$ with pmf $p_X = [0.5; 0.25; 0.125; 0.125]$
  
  $\Rightarrow H(X) = 1.75$ bits.

- Sample sequences consisting of eight i.i.d samples

  - strongly typical $\Rightarrow$ correct proportions
    
    $\bullet\blacklozenge\overline{\blacklozenge}\blacklozenge \Rightarrow -\log p(x) = 14 = 8 \times 1.75$

  - weakly typical $\Rightarrow -\log p(x) = nH(X) = 14 = 8 \times 1.75$

  - not typical at all $\Rightarrow -\log p(x) \neq nH(X)$

The typical set visually

**How to count the # in the typical set?**

Most + least likely sequences

NOT in typical set!!

Properties of the typical set

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

2. $\Pr(A^{(n)}) > 1 - \epsilon$ for $n$ sufficiently large.

3. $(1 - \epsilon)2^{n(H(X) - \epsilon)} \leq |A^{(n)}| \leq 2^{n(H(X) + \epsilon)}$ for $n$ sufficiently large.
Consequences of the AEP

Let \( x^n \) denote \( \{x_1, x_2, \ldots, x_n\} \), and let \( l(x^n) \) be the length of the codeword corresponding to \( x^n \).

**Coding Scheme:**
- If \( x^n \in A^{(n)} \): '0' + at most \( 1 + n(H(X) + \epsilon) \)
- If \( x^n \notin A^{(n)} \): '1' + at most \( 1 + n \log |X| \)

If \( n \) is sufficiently large so that \( \Pr\{A^{(n)}\} \geq 1 - \epsilon \), the expected codeword length is

\[
E[l(x^n)] = \sum p(x^n) l(x^n) \\
\leq n(H + \epsilon) + cn(\log|X|) + 2 \\
= n(H + \epsilon)
\]

Surely \( \log |X| \) is enough, but \( H(X) \leq \log |X| \).

Consequences of the AEP

Source vs. channel coding

- **Source coding**
- **Channel coding**

Error-correcting codes

"Shannon theory"  "Channel coding"
What is the capacity of this channel?

Intuitively
Formally

How to communicate reliably?

Use these 9 symbols!

\[ C = \log_2(9) \]
Capacity in general

- Reduce the rate so as to produce non-overlapping outputs!

Source: A C K J I

\( \text{Inputs} \) \( \text{X}^{\infty} \) \( \text{p}(x) \) \( \text{Y}^{\infty} \) \( \text{Estimate of message} \) \( \text{W} \rightarrow \text{W} \rightarrow \text{Destination} \)

Mathematical description of capacity

- Can achieve reliable communication for all transmission rates \( R \):
  \[
  R < C \\
  \begin{array}{c}
  0 \\
  C
  \end{array}
  \]

- BUT, probability of decoding error always bounded away from zero if \( R > C \)
  \[
  \begin{array}{c}
  0 \\
  C
  \end{array}
  \]

Capacity: key ideas

- "non-confusable" inputs
- # "non-confusable" inputs = channel's capacity
- channel capacity depends on \( p(y|x) \)

Mutual information between 2 random variables:

\[
I(X; Y) = \sum_{x,y} p(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) \\
= H(X) - H(X|Y) \\
= H(Y) - H(Y|X)
\]

(A) \( I(X; Y) \) is the reduction in the uncertainty about \( X \) due to knowledge of \( Y \)

(B) if \( X, Y \) are independent \( I(X; Y) = 0 \)

(C) \( I(X; Y) \) is non-negative

\[
\begin{array}{c}
H(X,Y) \\
H(X) \\
H(Y) \\
H(X|Y) \\
I(X; Y) \\
H(Y|X)
\end{array}
\]
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

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 is \( x \).

Memoryless: \( p(y_1|x_1, x_2, \ldots, x_n, y_1, y_2, \ldots, y_n) = p(y_n|x_n) \)

\[ X^n \xrightarrow{p(y|x)} Y^n \]

**Definition:** The maximal probability of error of an \((M, n)\) code is defined as

\[ \lambda^{(n)} = \max_{i \in \{1, 2, \ldots, M\}} \Pr(g(Y^n) \neq i|X^n = x^n(i)) \]

**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.
Channel coding theorem

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

**Theorem:** Channel coding theorem For a DMC, all rates below capacity \( C \) are achievable.

- Specifically, for any rate \( R < C \), there exists a sequence of \( (2^nR, n) \) codes with maximum probability of error \( \lambda^n \to 0 \).
- Conversely, any sequence of \( (2^nR, n) \) codes with \( \lambda^n \to 0 \) 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
- NOT constructive! Does NOT tell us how to code to achieve capacity!

---

**Intuition for the noisy typewriter channel**

In general

\[ \text{Pick subset of typical } X \text{ such that} \]

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 \( 2^{nR(Y|X)} \) possible \( Y \) sequences, all equally likely.
- Want to ensure that no two typical \( X \) sequences produce the same \( Y \) sequence.
- There are \( 2^{nR(Y)} \) typical \( Y \) sequences. Dividing, we get \( \frac{2^{nR(Y)}}{2^{nR(Y|X)}} = 2^{nI(X;Y)} \) distinguishable input sequences.
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 $(\lfloor 2^{nR} \rfloor, n)$ codes with maximum probability of error $\lambda(n) \to 0$.
- Conversely, any sequence of $(\lfloor 2^{nR} \rfloor, n)$ codes with $\lambda(n) \to 0$ must have $R \leq C$.

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

---

Use of information theory / channel capacity?

- Benchmark for performance of practical systems
- Guideline in designing systems - what’s worth shooting for?
- Theoretical insights can lead to practical insights
- Pretty!

---

My research:
Multi-user Shannon theory
(determine capacity regions of networks)

---

Point-to-point

- Channel capacity ✓
- How to approach it for memoryless Gaussian noise channels ✓

*Is that the end of the story?*

---

NO! what about networks (multi-user information theory)?

---

Capacity and capacity regions

- Point to point capacity
- Multi-user capacity region
**Capacity regions**

- Outer bound
- Capacity region
- Achievable region

**Achievable rate region**

- Outer bound
- Capacity region
- Achievable region

- Propose a coding scheme (random codes!)
  \[ R_1 \leq I(X_1; Y | X_2) \]
  \[ R_2 \leq I(X_2; Y | X_1) \]
  \[ R_1 + R_2 \leq I(X_1, X_2; Y) \]

- Prove that as long as \[ \triangledown \] holds, reliable communication possible

**Outer bound**

- 
  \[ R_1 \leq I(X_1; Y | X_2) \]
  \[ R_2 \leq I(X_2; Y | X_1) \]
  \[ R_1 + R_2 \leq I(X_1, X_2; Y) \]

- Prove that error is bounded away from 0 when \[ \triangledown \] not satisfied
- Find a more capable channel whose capacity is known

**Capacity regions**

- Outer bound
- Capacity region
- Achievable region

- Limit of communication, NOT necessarily how to achieve it in practice!
- However, benchmark and guidance in practical designs

**Ultimate goal**

Capacity of arbitrary network where

\[ x_n(i) = f(w_i, y_n^{n-1}(i)) \]

**Very difficult -- start slow**

and arbitrarily correlated messages

**Key multi-user channels**

- Broadcast channel
- Relay channel
- Multiple-access channel
- Interference channel
Other areas of information theory

- Shannon theory
- Coding theory
- Coding techniques
- Complexity and cryptography
- Pattern recognition, Statistical learning and inference
- Source coding
- Detection and Estimation
- Communications
- Sequences
- At large

Questions?

Natasha Devroye
Assistant Professor
University of Illinois at Chicago
SEO 1039 -- come for a visit!
http://www.ece.uic.edu/Devroye