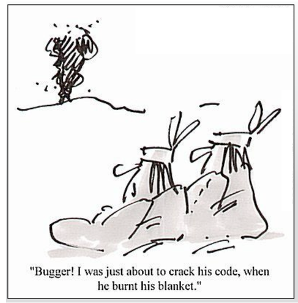
A quick introduction to information theory

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

History of (wireless) communications

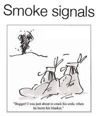


"Bugger! I was just about to crack his code, when he burnt his blanket."

Smoke signals

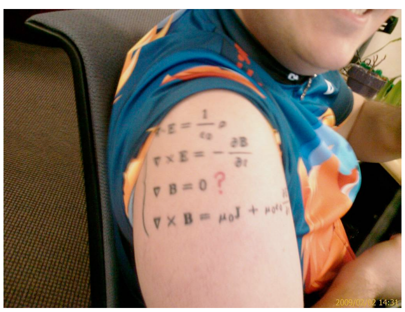
History of (wireless) communications

Smoke signals



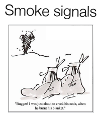
→

Maxwell's equations



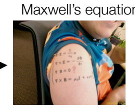
History of (wireless) communications

Smoke signals



→

Maxwell's equations



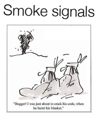
→

Marconi demonstrates wireless telegraph

Marconi

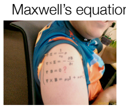
History of (wireless) communications

Smoke signals



→

Maxwell's equations

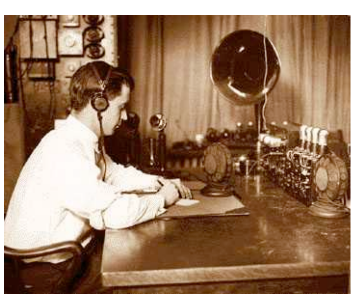


→

Marconi

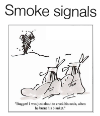
→

Detroit police cars radio dispatch in 1925



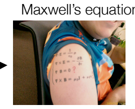
History of (wireless) communications

Smoke signals



→

Maxwell's equations

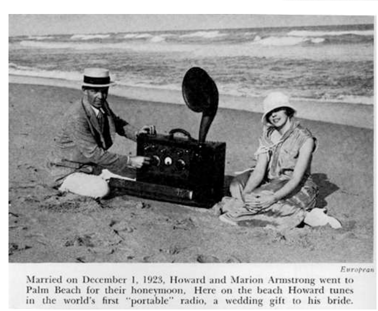


→

Marconi

→

Detroit police

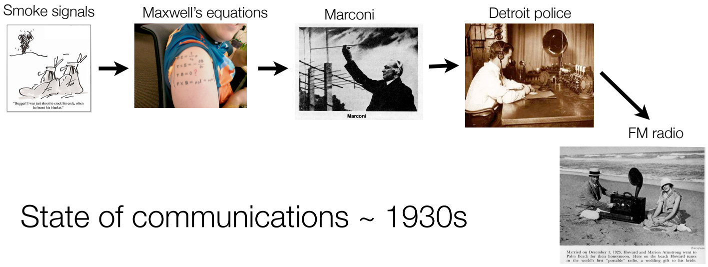


Married on December 1, 1923, Howard and Marion Armstrong went to Palm Beach for their honeymoon. Here on the beach Howard tunes in the world's first "portable" radio, a wedding gift to his bride.

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*

A Mathematical Theory of Communication. **Bell System Technical Journal, *27*, 379-423 & 623-656, 1948.**

BITS

How fast can we communicate?

What is information?

RANDOMNESS

How much can we compress information?

What is communication?

Information theory's claims to fame

Source coding

zip jpg mp3

- Source = random variable
- Ultimate data compression limit is the source's entropy H

Channel coding

Turbo-codes 00100000000 FADING CHANNEL

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

Reliable communication possible ↔ H<C

Technology independent limits!

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)

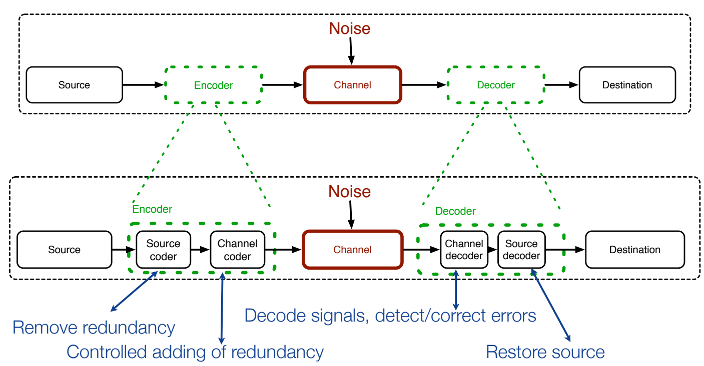
Mutual information I(X;Y)

- 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

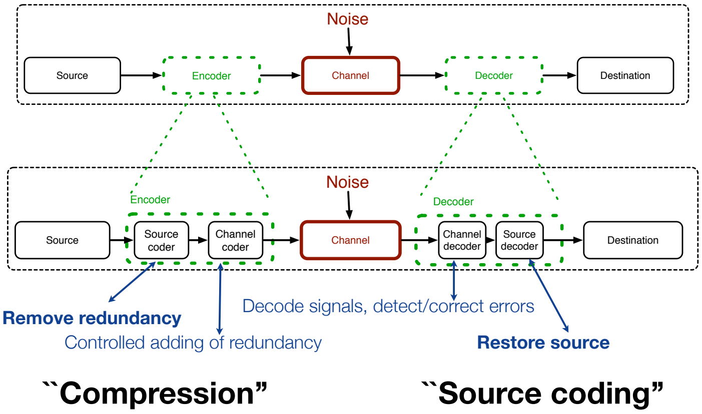
Source = random variable X, p(x)

X ➔ Channel p(y|x) ➔Y

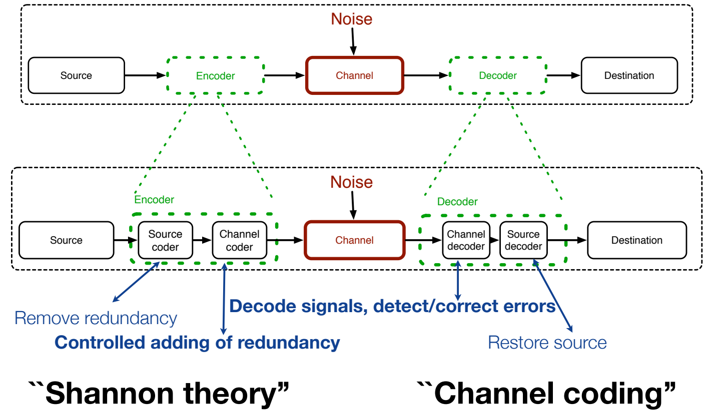
Source vs. channel coding



Source vs. channel coding



Source vs. channel coding



Source coding

Compression

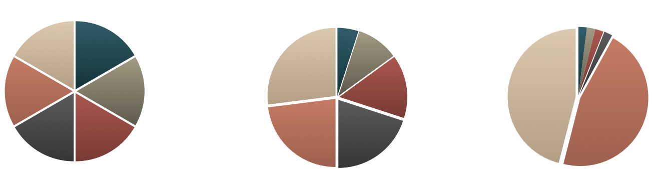
Main result in source-coding/compression

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

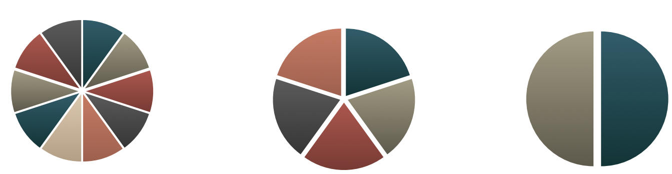
Definition: The entropy $H(X)$ of a discrete random variable X with pmf $p_X(x)$ is given by

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

Order these in terms of entropy



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?

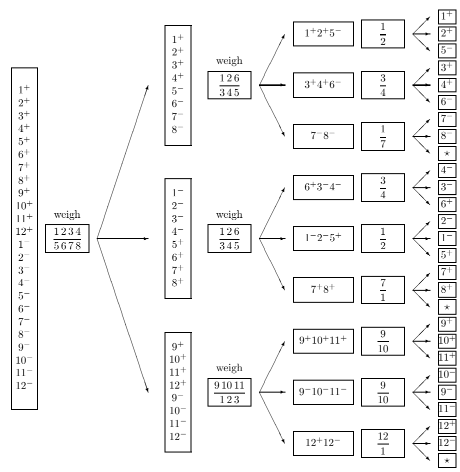
(D) $H(X) = -\sum_x p(x)\log_2(p(x))$

(E) entropy of a deterministic value is 0

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.*

12 balls weighing: 1 lighter or heavier

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



[Mackay textbook pg. 69]

Examples of codes

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

Symbol	Probability	Codeword
1	Pr(1) =0.5	C(1) = 0
2	Pr(2) = 0.25	C(2) = 10
3	Pr(3)=0.125	C(3) = 110
4	Pr(4)=0.125	C(4) =111

Decode 0110111100110 134213

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)? 1.75 bits

$\frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{8}\times 3$

Main result 1: data compression

Theorem: Data Compression Let $X^n \overset{iid}{\sim} p(x)$ 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

$$\textbf{L(C)} = E\left[\frac{1}{n}l(X^n)\right] \le H(X) + \epsilon$$

for n sufficiently large.

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.

Strong versus Weak Typicality

- Intuition behind typicality?
- $\mathcal{X} = \{\clubsuit, \diamondsuit, \heartsuit, \spadesuit\}$ 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
  $\clubsuit\clubsuit\clubsuit\clubsuit\diamondsuit\diamondsuit\heartsuit\spadesuit \quad -\log p(x) = 14 = 8 \times 1.75$
- weakly typical $\Rightarrow \log p(x) = nH(X)$
  $\clubsuit\clubsuit\diamondsuit\diamondsuit\diamondsuit\diamondsuit\diamondsuit\diamondsuit \quad -\log p(x) = 14 = 8 \times 1.75$
- not typical at all $\Rightarrow \log p(x) \neq nH(X)$
  $\spadesuit\spadesuit\spadesuit\spadesuit\spadesuit\spadesuit\spadesuit\spadesuit \quad -\log p(x) = 24 \neq 8 \times 1.75$

Definition: weak typicality

- *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

$$2^{-n(H(X)+\epsilon)} \leq p(x_1, x_2, \ldots, x_n) \leq 2^{-n(H(X)-\epsilon)}.$$

- If $(x_1, x_2, \ldots, x_n) \in A_\epsilon^{(n)}$, then

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

The typical set visually

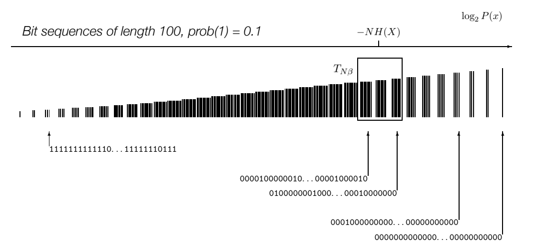


Figure 4.12. Schematic diagram showing all strings in the ensemble X^N ranked by their probability, and the typical set $T_{N\beta}$.

How to count the # in the typical set?

[Mackay pg. 81]

Most + least likely sequences NOT in typical set!!

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 μ and variance $\sigma^2 < \infty$. Let

$$S_n \triangleq \frac{1}{n}[X_1 + X_2 + \ldots + X_n]$$

- *Theorem: Weak Law of Large Numbers*

$$S_n \to \mu \qquad \text{in probability}$$

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

$$-\frac{1}{n}\log p(X_1, X_2, \ldots, X_n) \to H(X) \qquad \text{in probability.}$$

Properties of the typical set

1. If $(x_1, x_2, \cdots, x_n) \in A_\epsilon^{(n)}$ then $H(X) - \epsilon \leq -\frac{1}{n}\log p(x_1, x_2, \cdots, x_n) \leq H(X) + \epsilon$
2. $\Pr\{A_\epsilon^{(n)}\} > 1 - \epsilon$ for n sufficiently large.
3. $(1-\epsilon)2^{n(H(X)-\epsilon)} \leq |A_\epsilon^{(n)}| \leq 2^{n(H(X)+\epsilon)}$ for n sufficiently large.

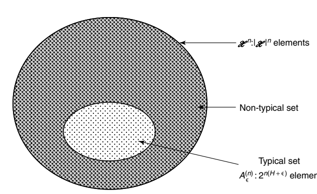


FIGURE 3.1. Typical sets and source coding.

[Cover+Thomas pg. 60]

Consequences of the AEP

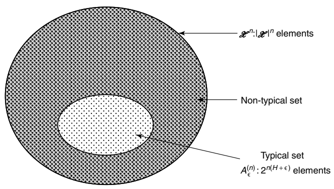


FIGURE 3.1. Typical sets and source coding.

Typical set contains almost all the probability!

How many are in this set useful for source coding (compression)!

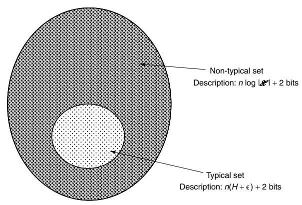


FIGURE 3.2. Source code using the typical set.

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:

By enumeration!

- if $x^n \in A_\epsilon^{(n)}$: '0' + at most $\boxed{1 + n(H(X) + \epsilon)}$
- if $x^n \notin A_\epsilon^{(n)}$: '1' + at most $1 + n \log |\mathfrak{X}|$

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

$$\begin{aligned} E[l(X^n)] &= \sum_{x^n} p(x^n) l(x^n) \\ &\leq n(H+\epsilon) + \epsilon n (\log |\mathcal{X}|) + 2 \\ &= n(H + \epsilon') \end{aligned}$$

AEP and data compression

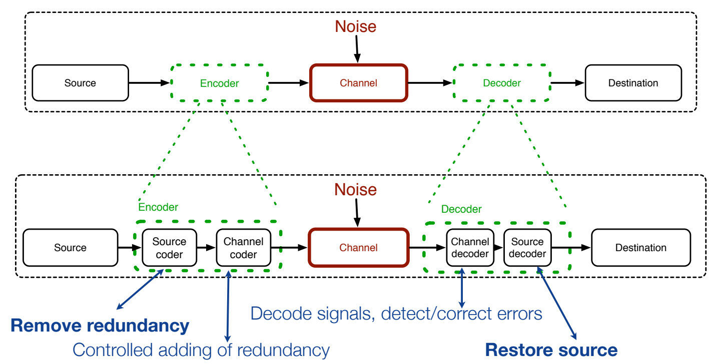
Theorem: Data Compression Let $X^n \overset{iid}{\sim} p(x)$ and let $\epsilon > 0$. Then there exists a code that maps sequences x^n of length n into binary strings such that the mapping is one-to-one (and therefore invertible) and

$$E\left[\frac{1}{n} l(X^n)\right] \leq H(X) + \epsilon$$

for n sufficiently large.

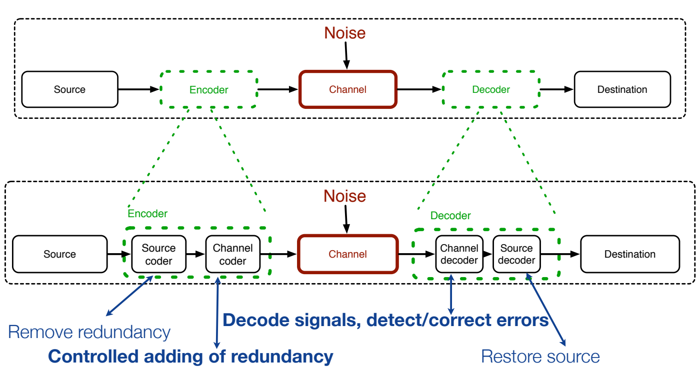
Surely $\log |\mathcal{X}|$ is enough, but $H(X) \leq \log |\mathcal{X}|$.

Source vs. channel coding



``Compression”** **``Source coding”

Source vs. channel coding

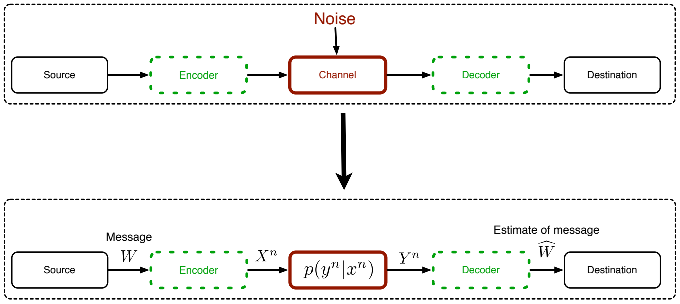


``Shannon theory”** **``Channel coding”

Channel coding

Error-correcting codes

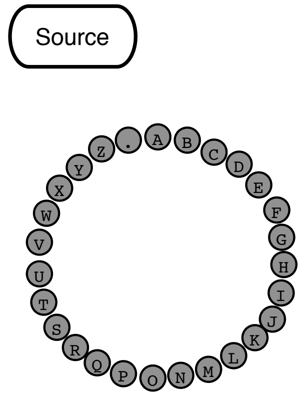
Communication system model



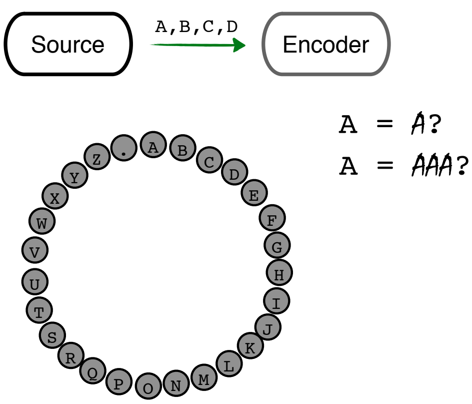
What is the **capacity** of this channel?

Intuitively **Formally**

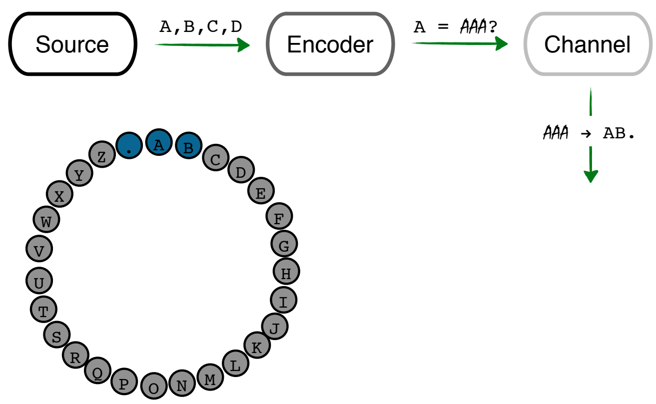
Channel capacity: a cute example



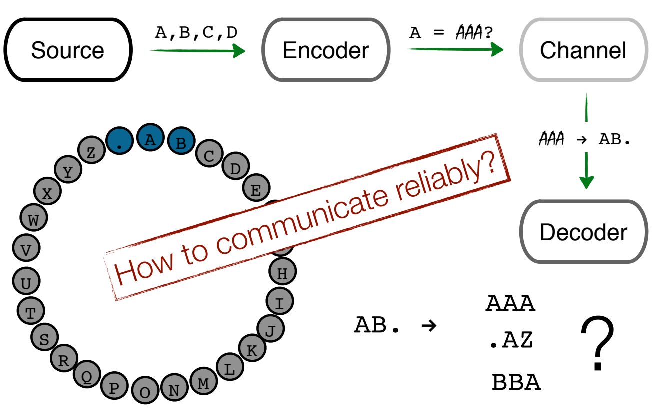
Channel capacity: a cute example



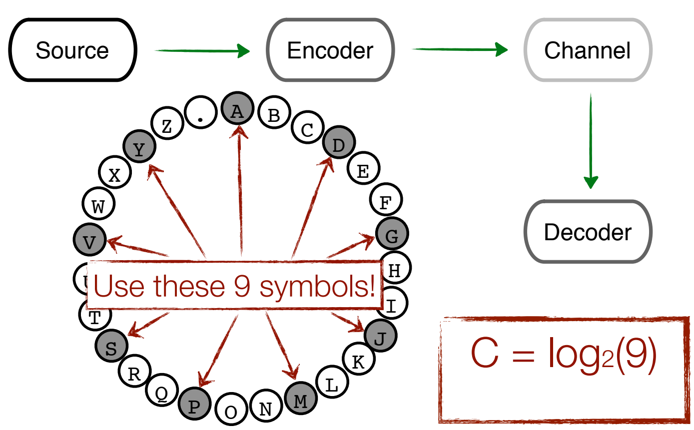
Channel capacity: a cute example



Channel capacity: a cute example

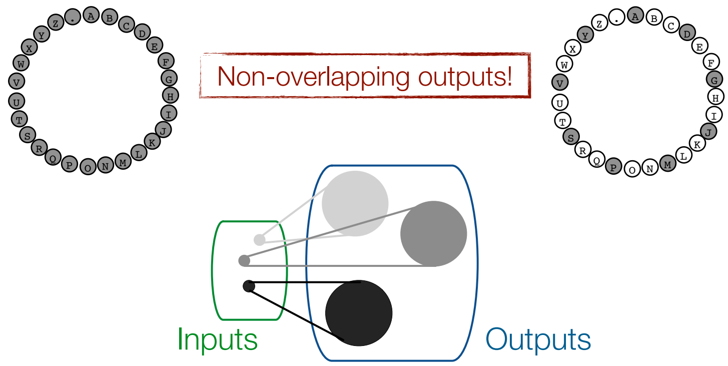


Channel capacity: a cute example



Capacity in general

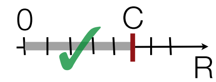
- Reduce the rate so as to produce



Mathematical description of capacity

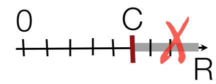
- Can achieve reliable communication for all transmission rates R:

R < C

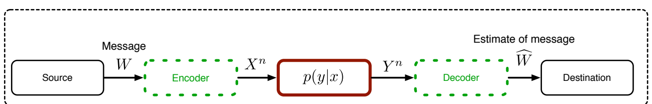


- BUT, probability of decoding error always bounded away from zero if

R > C

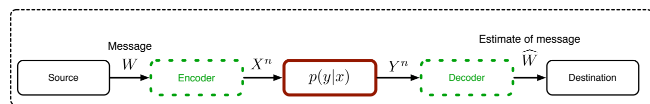


Capacity: key ideas



- “non-confusable” inputs
- # ``non-confusable” inputs = channel’s capacity
- channel capacity depends on p(y|x)

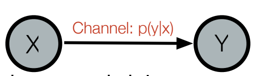
Point-to-point channel capacity



$$\boxed{C = \max_{p(x)} I(X;Y) \quad \text{bits/channel use}}$$

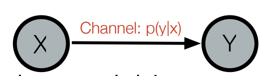
“mutual information” between X and Y

$$I(X;Y) = \sum_{x,y} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$



Mutual information between 2 random variables:

$$\begin{aligned} I(X;Y) &= \sum p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$



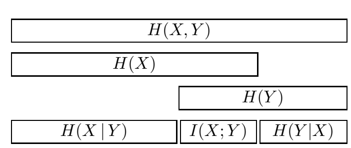
Mutual information between 2 random variables:

$$\begin{aligned} I(X;Y) &= \sum p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right) \\ &= H(X) - H(X|Y) \\ &= H(Y) - H(Y|X) \end{aligned}$$

(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



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

What do you **really** mean by

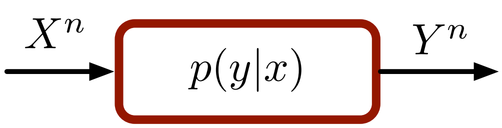
```
Highest rate (bits/channel use)
that can communicate at reliably
```

?

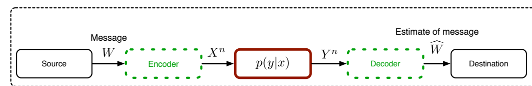
Definitions

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.

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



Definitions



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

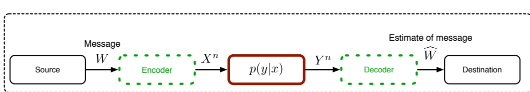
1. An index set $\{1, 2, \ldots, M\}$ over messages W.
2. An encoding function $X^n : \{1, 2, \ldots, M\} \to \mathcal{X}^n$, yielding codewords $x^n(1), x^n(2), \ldots$ (This set is called the *codebook* $\mathcal{C}$.)
   $x^n(W)$ passes through the channel and is received as a random sequence $Y^n \sim p(y^n \mid x^n)$.
3. A (deterministic) decoding function

$$g : \mathcal{Y}^n \to \{1, 2, ., ., .M\},$$

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

Send 1 of M messages over n channel uses

Definitions

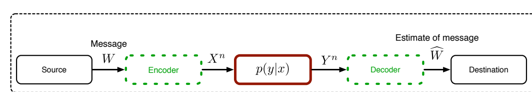


Definition: The *maximal probability of error* of an (M, n) code is defined as

$$\lambda^{(n)} = \max_{i \in \{1,2,\cdots m} \Pr\{g(Y^n) \neq i | X^n = x^n(i)\}$$

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

Definitions



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

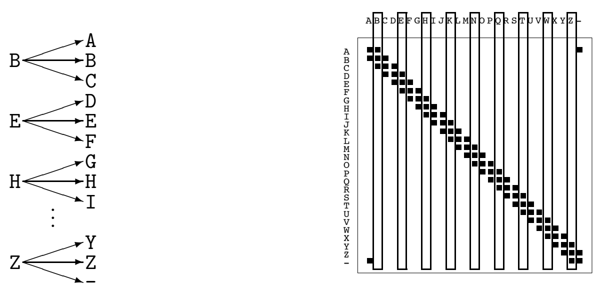
- Specifically, for every rate $R < C$, there exists a sequence of $(\lceil 2^{nR} \rceil, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.
- Conversely, any sequence of $(\lceil 2^{nR} \rceil, n)$ codes with $\lambda^{(n)} \to 0$ must have $R \le C$.

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

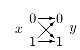


Count the # non-confusable subsets!

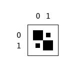
[Mackay textbook]

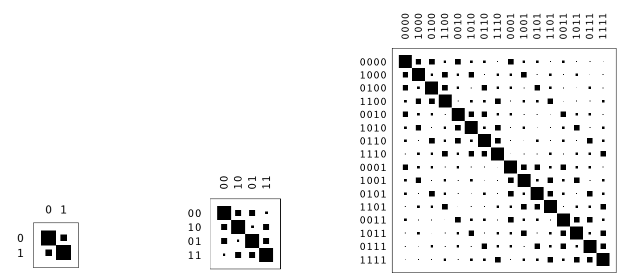
Intuition for the binary symmetric channel

Binary symmetric channel. $\mathcal{A}_X = \{0,1\}$. $\mathcal{A}_Y = \{0,1\}$.



$$\begin{aligned} P(y=0 \mid x=0) &= 1-f; & P(y=0 \mid x=1) &= f; \\ P(y=1 \mid x=0) &= f; & P(y=1 \mid x=1) &= 1-f. \end{aligned}$$



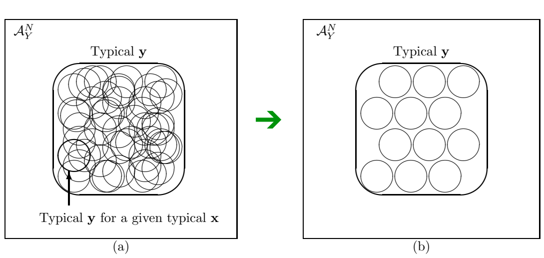


$N = 1$ $N = 2$ $N = 4$

[Mackay textbook]

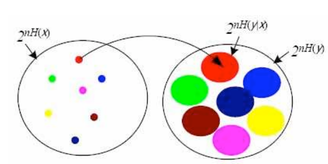
In general

Pick subset of typical X such that



[Mackay textbook]

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.

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 $(\lceil 2^{nR} \rceil, n)$ codes with maximum probability of error $\lambda^{(n)} \to 0$.
- Conversely, any sequence of $(\lceil 2^{nR} \rceil, n)$ 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 $R < C$.

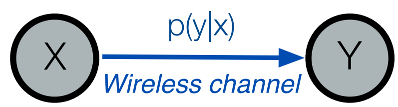
Use of information theory / channel capacity?

- 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)

?

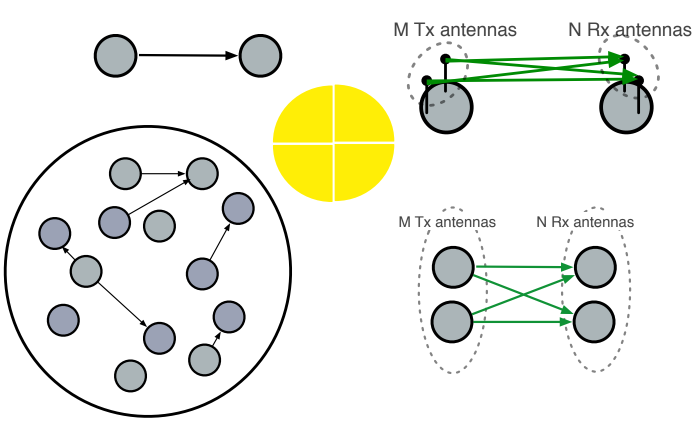
Point-to-point



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

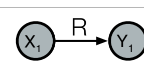
Is that the end of the story?

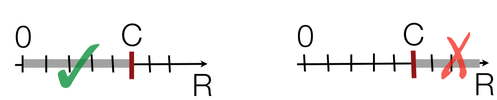
NO! what about networks *(multi-user information theory)*?



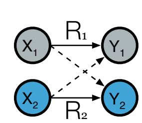
Capacity and capacity regions

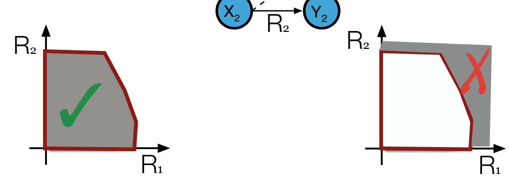
- Point to point **capacity**



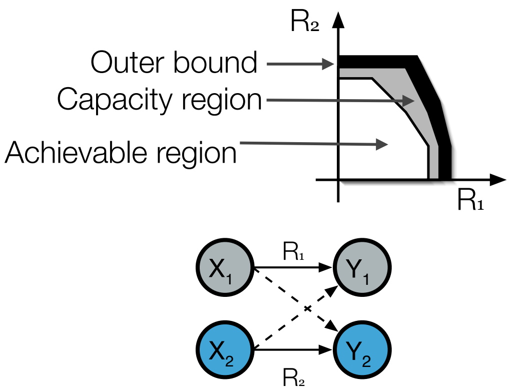


- Multi-user **capacity region**

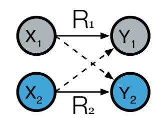


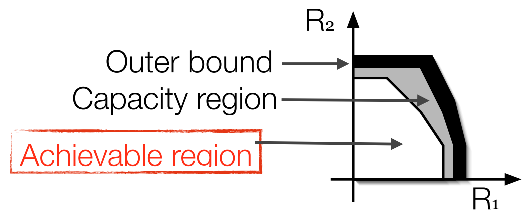


Capacity regions



Achievable rate region

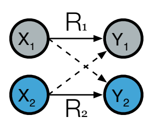


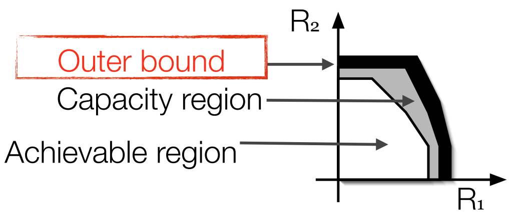


- Propose a coding scheme (random codes!)
- Prove that as long as ⇒ holds, reliable communication possible

$$R_1 \leq I(X_1;Y|X_2)$$
$$R_2 \leq I(X_2;Y|X_1)$$
$$R_1 + R_2 \leq I(X_1,X_2;Y)$$

Outer bound

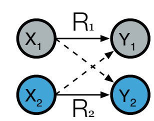


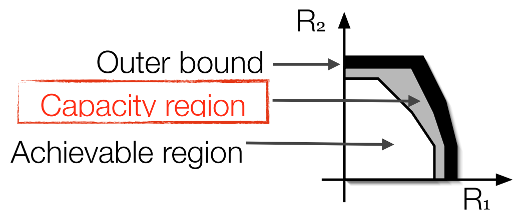


$$R_1 \leq I(X_1;Y|X_2)$$
$$R_2 \leq I(X_2;Y|X_1)$$
$$R_1 + R_2 \leq I(X_1,X_2;Y)$$

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

Capacity regions



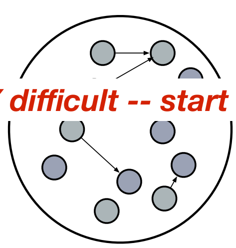


- Limit of communication, NOT necessarily how to achieve it in practice!
- However, benchmark and guidance in practical designs

Ultimate goal

Capacity of arbitrary network where

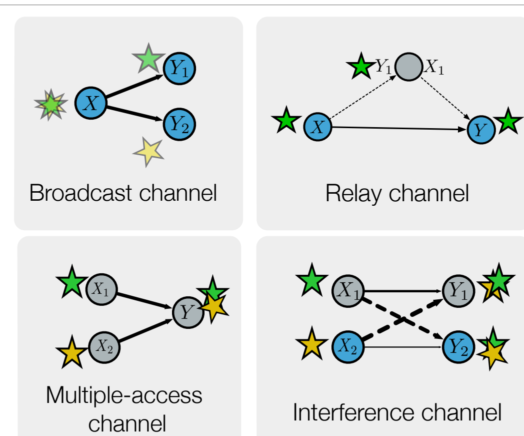
$$x_n(i) = f(w_i, y_1^{n-1}(i))$$



VERY difficult -- start slow

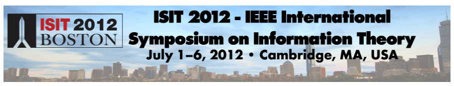
and arbitrarily correlated messages

Key multi-user channels



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