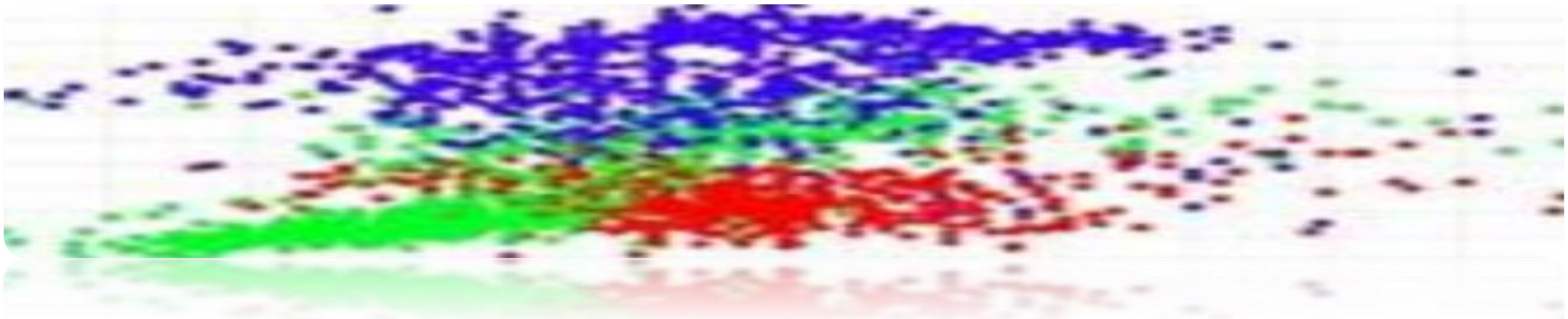
Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam



BASICS OF INFORMATION THEORY

Outline

- Information content
- Entropy, mutual information, relative entropy
- Compression, Shannon’s Noiseless Coding Theorem
- Minimum description length
- Noisy channels, Noisy-Channel Coding Theorem
- Modeling information content in natural language

Information theory: Overview

Claude Shannon
(1916 – 2001)

- Landmark publication in 1948: **"A Mathematical Theory of Communication"**
- Fundamental problem: Exact or approximate reproduction of a received message

- Main assumptions
  - Communication happens over a channel
  - Information is measurable and always ≥ 0, and is never lost
  - The bit is the most fundamental unit of information

- **Coding theory** is a direct application area of information theory
  - **Data compression** (Source Coding)
  - **Error correcting codes** (Channel Coding)

Information content

- Formalization of **information content**

  - 1. Intuition: The more surprising a piece of information (i.e. event), the higher its information content

$$h(x) \uparrow \Leftrightarrow P(x) \downarrow$$

  - 2. Intuition: The content of two independent events x and event y should simply add up (**additivity**)

$$h(x \cap y) = h(x) + h(y)$$

Define $h(x) := -\log_2 P(x)$

Entropy

- Let X be a discrete random variable
- Definition: The **entropy of X** is defined as

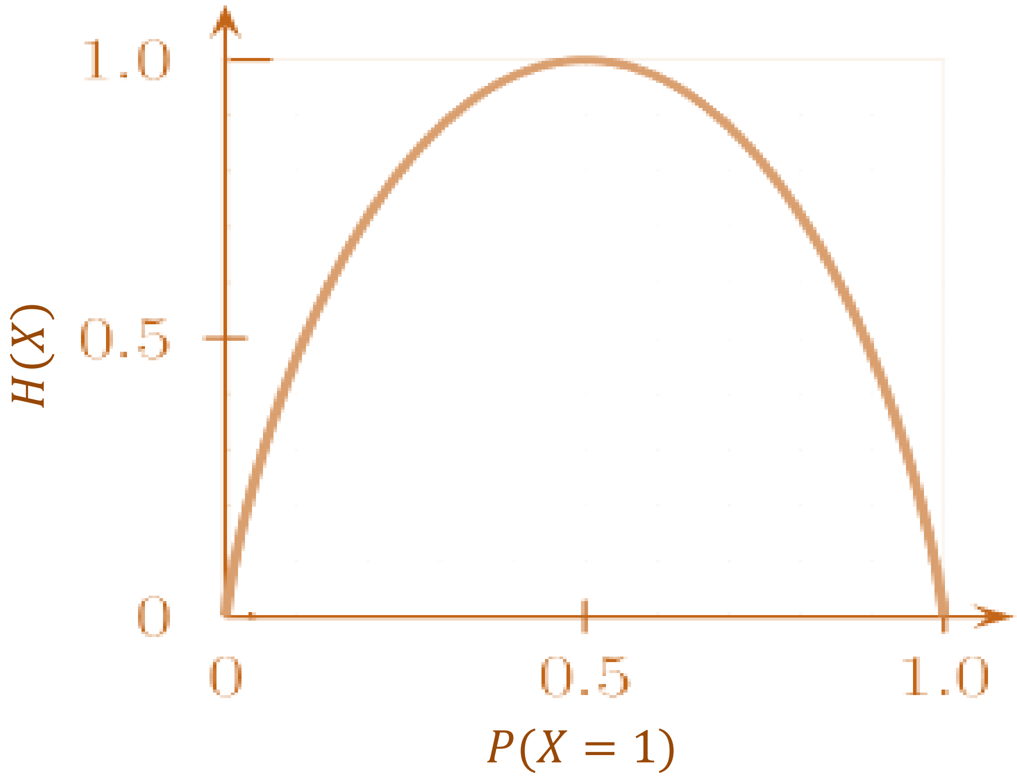
$$H(X) = -\sum_{x \in dom(X)} P(x) \log P(x)$$

- Example
  - Let X be a random variable with 8 equally possible states
  - What is the average number of bits needed to encode a state of X?

$$H(X) = -\sum_{x \in dom(X)} P(x) \log P(x)$$

$$= -8\,\frac{1}{8}\log\frac{1}{8} = 3$$

Entropy function for a binary random variable X



Properties of entropy

- $\log n \geq H[X] \geq 0$, where $n = |dom(X)|$
- For discrete random variable X the entropy $H[X]$ is maximized for $f_X(x) = \frac{1}{|dom(X)|}$
- For continuous random variable X, $H[X]$ is maximized for $f_X(x) = N\big(E(X), Var(X)\big)$
- $H[Y|X] = -\sum_X \sum_Y P(X,Y) \log P(Y|X)$ (conditional entropy, i.e., the average additional information to encode Y when X is known)
- $H[X,Y] = H[Y|X] + H[X]$ (joint entropy of X and Y)
- $H[X,Y] = H[X] + H[Y]$ for independent random variables X, Y (additivity)

Lossless compression: Huffman compression

- Let X be a random variable with 8 possible states

$$\{x_1, x_2, x_3, x_4, x_5, x_6, x_7, x_8\}$$

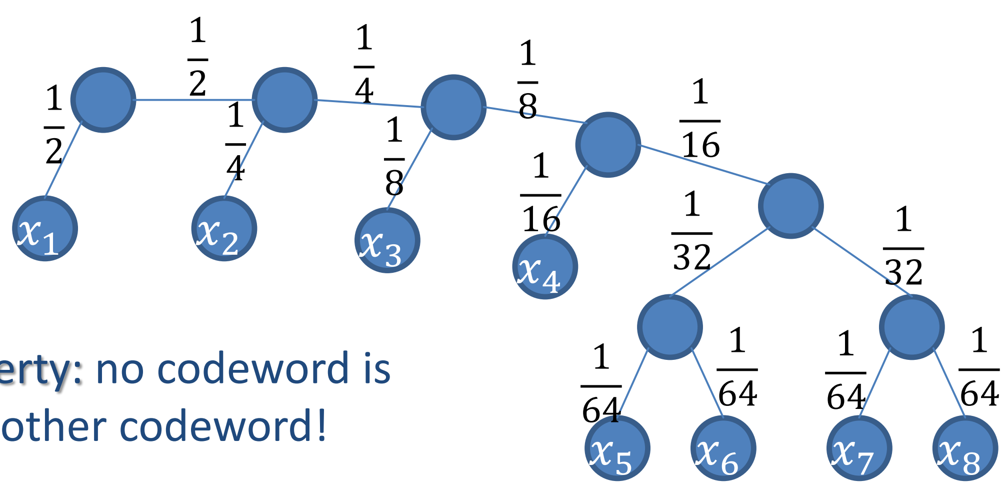
with occurrence probabilities

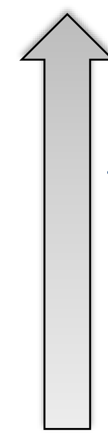
$$(\frac{1}{2}, \frac{1}{4}, \frac{1}{8}, \frac{1}{16}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64}, \frac{1}{64})$$

In any case: 3 bits would be sufficient to encode any of the 8 states.

Can we do better?

$$encoding: 0,10,110,1110,111100,111101,111110,111111$$





Bottom-up tree construction by combining lowest-frequency subtrees

Prefix property: no codeword is prefix of another codeword!

Noiseless Coding Theorem

- Let X be a random variable with n possible states.

- Theorem: For any noiseless encoding of the states of X, $H(X)$ is a lower bound on the average code length of a state of X.

- Corollary: The Huffman compression achieves the above lower bound and is thus optimal for noiseless compression

(Note that there are more sophisticated algorithms that encode several states at the same time, e.g. Lempel-Ziv-Welch algorithm. With appropriately generalized notions of variables and states, Shannon's theorem still holds.)

Example: Ziv-Lempel compression (e.g., LZ77)

- Use lookahead window and backward window to scan text

- Identify in lookahead window the longest string that occurs in backward window and replace the string by a pointer to its previous occurrence

- Text is encoded in triples $(previous, length, new)$

  $previous$: distance to previous occurrence

  $length$: length of the string

  new: symbol following the string

More advanced variants use adaptive dictionaries with statistical occurrence analysis!

Example: Ziv-Lempel compression (e.g., LZ77)

Text: $A\ A\ B\ A\ B\ B\ B\ A\ B\ A\ A\ B\ A\ B\ B\ B\ A\ B\ B\ A\ B\ B$

Code: $(\emptyset, 0, A)(-1{,}1, B)(-2{,}2, B)(-4, 3, A)(-9, 8, B)(-3{,}3, \emptyset)$

- Note that LZ77 and other sophisticated lossless compression algorithms (e.g. LZ78, Lempel-Ziv-Welch,…) encode several states at the same time.

- With appropriately generalized notions of variables and states, Shannon’s noiseless coding theorem still holds!

Minimum description length (MDL)

- Given a model M,
- $l(M)$: Number of bits needed to encode M,
- $l(D|M)$: Number of bits needed to encode sample data D based on M (typically a function capturing information loss, e.g. $-\log P(D|M)$, for the data)

- **MDL principle**: Choose the model for D is one that minimizes $l(M) + l(D|M)$

- MDL from a probability perspective: Let $P(M|D)$ be the posterior probability of model M given D.

$$\underset{M}{\operatorname{argmax}} P(M|D) = \underset{M}{\operatorname{argmax}} \left(\frac{P(D|M)P(M)}{P(D)}\right)$$
$$= \underset{M}{\operatorname{argmin}}(-\log P(D|M) - \log P(M) + \log P(D))$$

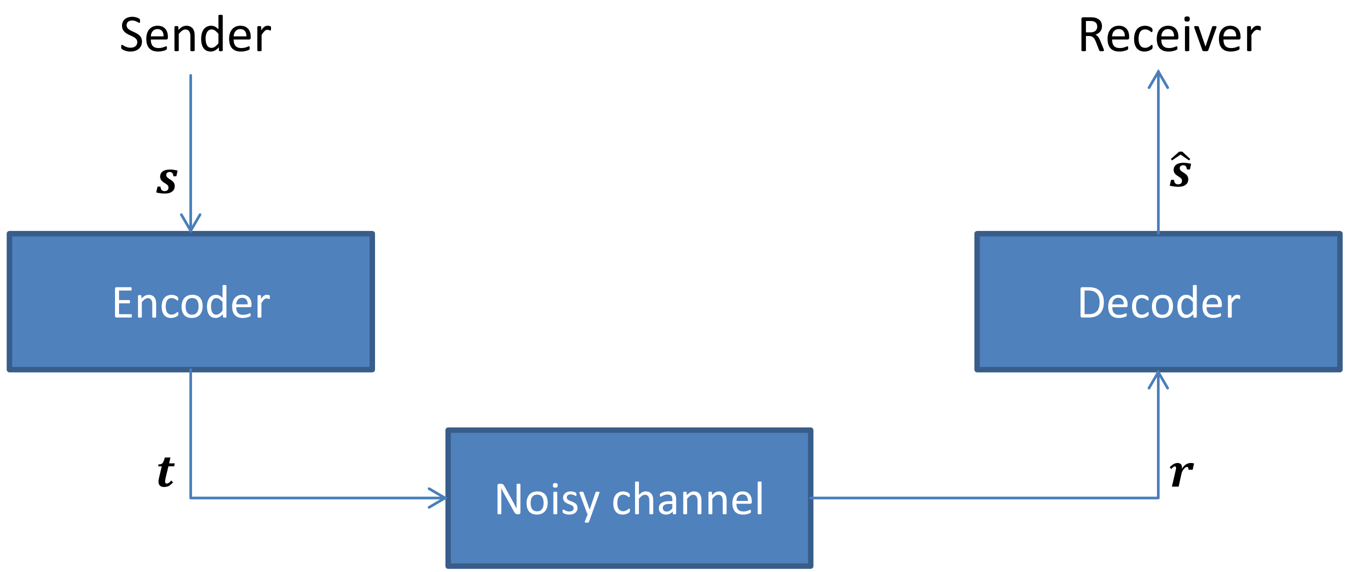
- **Occam's razor (by William of Ockham, 1285 – 1349)**

  Given multiple hypotheses that explain the same data (other things being equal), choose the most plausible one (≈ the simplest hypothesis).

Inductive inference and Kolmogorov Complexity

- Goal of inductive inference
  - Accurate predictions based on previous observations
- Simple examples
  - What is the continuation of 010011000111 … ?
  - What is the continuation of 2, 4, 6, 8, … ?
- For the last example sequence, $2k$ and $2k^4 - 20k^3 + 70k^2 - 98k + 48$ are both valid generating functions with different continuations
- The Kolmogorov Complexity (aka: algorithmic entropy) of a sequence $\boldsymbol{x}$
  - The size of the smallest program $\boldsymbol{m}$ that generates $\boldsymbol{x}$ as output (i.e., the amount of information contained in $\boldsymbol{m}$)
  - Is equivalent to the Minimum Description Length of $\boldsymbol{x}$

Noisy channels



- Encoder introduces systematic redundancy to $\boldsymbol{s}$, resulting in vector $\boldsymbol{t}$
- $\boldsymbol{r}$ is a noisy version of $\boldsymbol{t}$
- Decoder uses knowledge about systematic redundancy to reconstruct $\boldsymbol{s}$ as well as possible

- Main question: How to achieve perfect communication over an imperfect, noisy channel?

Example: Repetition Codes

s	0	0	1	0	1	1	0
t	000	000	111	000	111	111	000
n	000	001	000	000	101	000	000
r	000	001	111	000	010	111	000
$\hat{s}$	0	0	1	0	0	1	0
corrected errors		*					
undetected errors					*		

Example from D. J. McKay: Information Theory, Inference and Learning Algorithms

- Simple probabilistic approach
  - Suppose that $P(0) = P(1) = 0.5$ and that the probability of bit conversion due to noise (i.e., from 0 to 1 or vice versa) is q
  - Compute for each $\boldsymbol{r}_i$ the s_i that maximizes the likelihood ratio

$$\frac{P(\boldsymbol{r}_i \mid s_i)}{P(\boldsymbol{r}_i \mid \overline{s_i})} = \prod_{1 \leq j \leq 3} \frac{P(r_{ij} \mid t_{ij})}{P(r_{ij} \mid \overline{t_{ij}})} \text{, where } \frac{P(r_{ij} \mid t_{ij})}{P(r_{ij} \mid \overline{t_{ij}})} = \begin{cases} \frac{1-q}{q}, r_{ij} = t_{ij} \\ \frac{q}{1-q}, r_{ij} \neq t_{ij} \end{cases}$$

Channel capacity

Sender $\xrightarrow{X}$ Noisy channel $\xrightarrow{Y}$ Receiver

- X and Y are discrete random variables

$$P(X = x, Y = y) = \underbrace{P(Y = y | X = x)}_{\substack{\text{Inherent fixed property of} \\ \text{the communication channel}}} P(X = x)$$

- Definition: The channel capacity is given by

$$C = \max_{P(X)} \underbrace{\sum_x \sum_y P(X = x, Y = y) \log\left(\frac{P(X = x, Y = y)}{P(X = x)P(Y = y)}\right)}_{\text{Mutual information between } X \text{ and } Y}$$

- The **mutual information** measures the information that X and Y share in bits

Source information rate

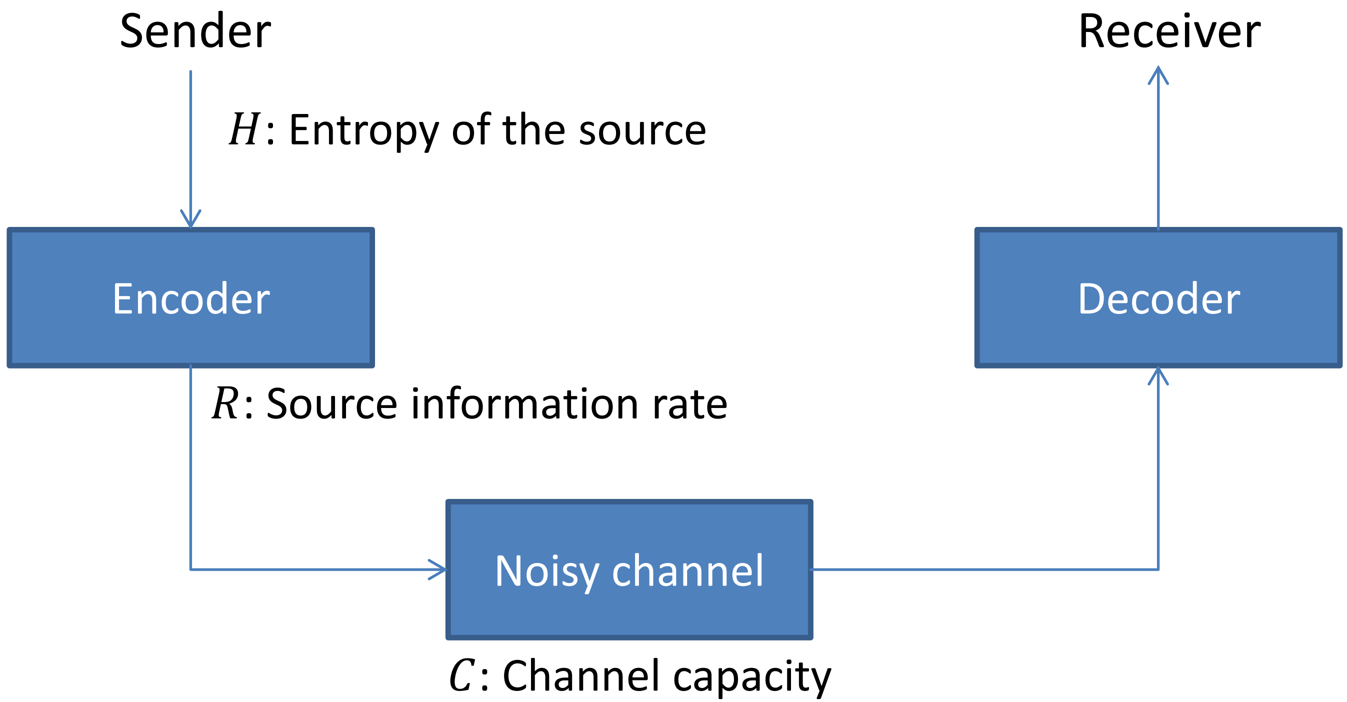
- Definition: The source information rate (or average block entropy) is the average number of channel symbols (i.e., bits) used to represent a source symbol in any block of source symbols

$$R = \lim_{n\to\infty} \frac{1}{n} H(X_n, X_{n-1}, \dots, X_1)$$

- A source is called memory-less iff

$$R = \lim_{n\to\infty} \frac{1}{n} H(X_n, X_{n-1}, \dots, X_1) = \lim_{n\to\infty} \frac{1}{n} \sum_{1\le i\le n} H(X_i) = \frac{n}{n} H(X) = H(X)$$

Noisy-Channel Coding Theorem (simplified)



- Theorem: Error-free transmission is possible if $H \leq R \leq C$.

Other important concepts of information theory (1)

- **Relative entropy (Kullback-Leibler Divergence)**
  Let f and g be two probability density functions over random variable X. Assuming that g is an approximation of f, the additional average number of bits to encode a state of X through g is given by

$$KL(f \parallel g) = \int_x f(x) \log \frac{f(x)}{g(x)} \, dx$$

- Properties of relative entropy
  - $KL(f \parallel g) \geq 0$ (Gibbs' inequality)
  - $KL(f \parallel g) \neq KL(g \parallel f)$ (asymmetric)

- **Jensen-Shannon Divergence** (symmetric measure based on KL divergence):
  $JS(f, g) = \alpha \, KL(f \parallel g) + \beta \, KL(g \parallel f)$ with $\alpha + \beta = 1$

Other important concepts of information theory (2)

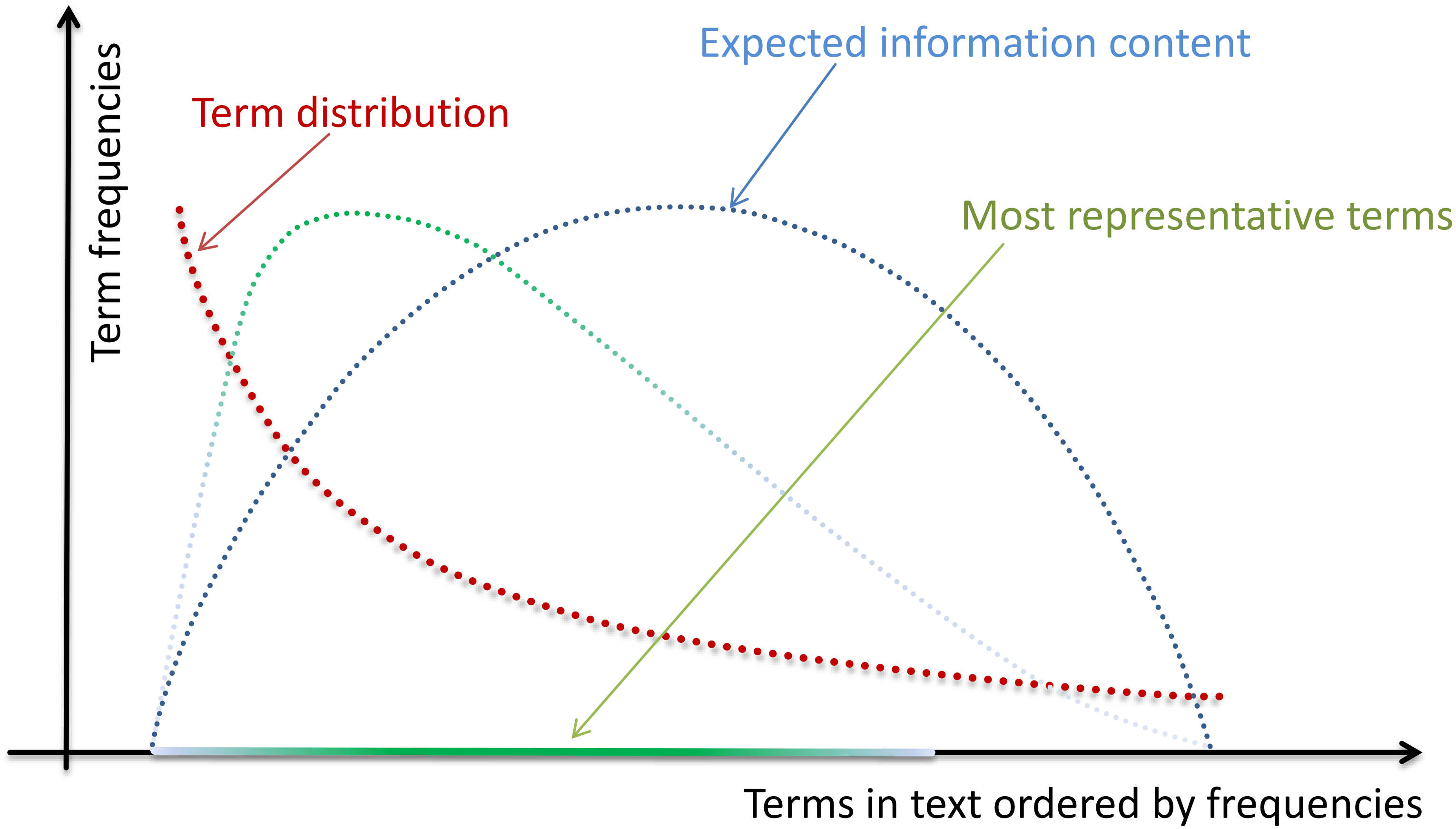
- Mutual information for continuous variables

  Let X and Y be two random variables with a joint distribution function P. The degree of their independence is given by

$$I[X,Y] = KL\big(P(X,Y) \parallel P(X)P(Y)\big) = \iint p(X,Y) \log \frac{P(X,Y)}{P(X)P(Y)} \, dX \, dY$$

- Properties of mutual information
  - $I[X,Y] \geq 0$
  - $I[X,Y] = 0$ if and only if X and Y are independent
  - $I[X,Y] = H[X] - H[X|Y] = H[Y] - H[Y|X] = H[X] + H[Y] - H[X,Y]$
    (i.e., the entropy reduction of X by being told the value of Y)

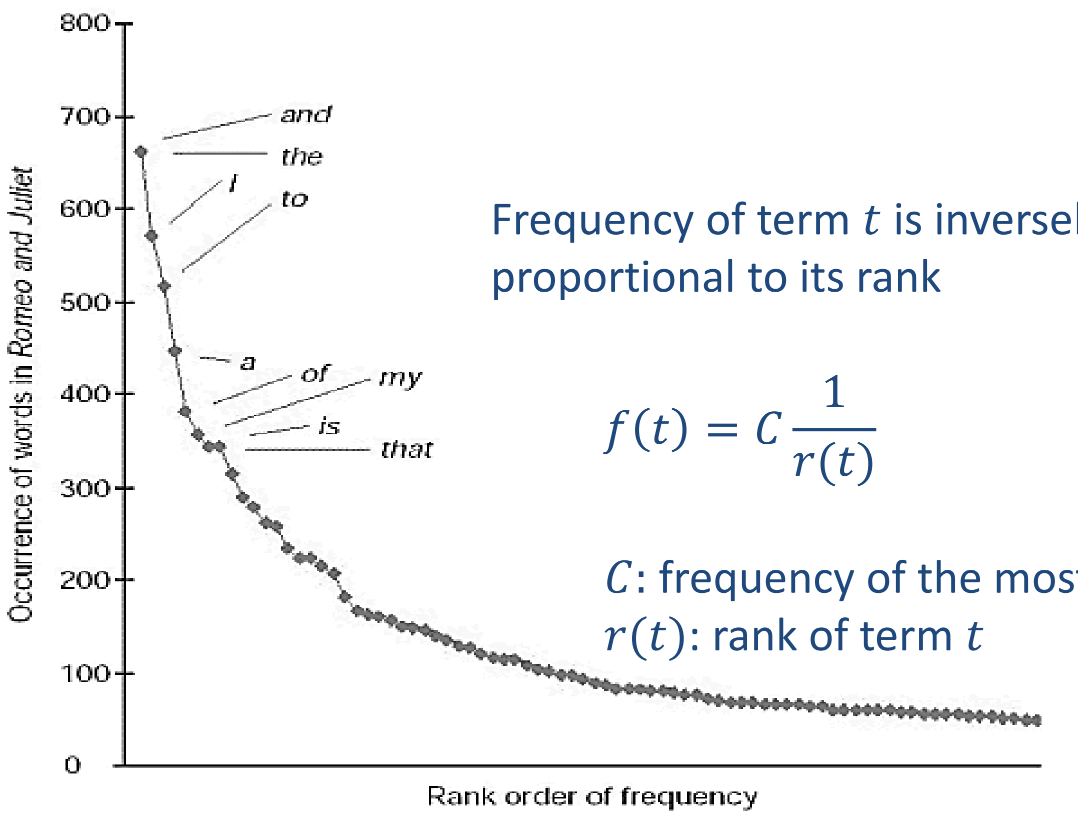
Modeling natural language



Is there a weighting scheme that gives higher weights to representative terms?

Zipf's law

- Linguistic observation in large text corpora
  - Few terms occur *very frequently*
  - Many terms occur *infrequently*



Frequency of term t is inversely proportional to its rank

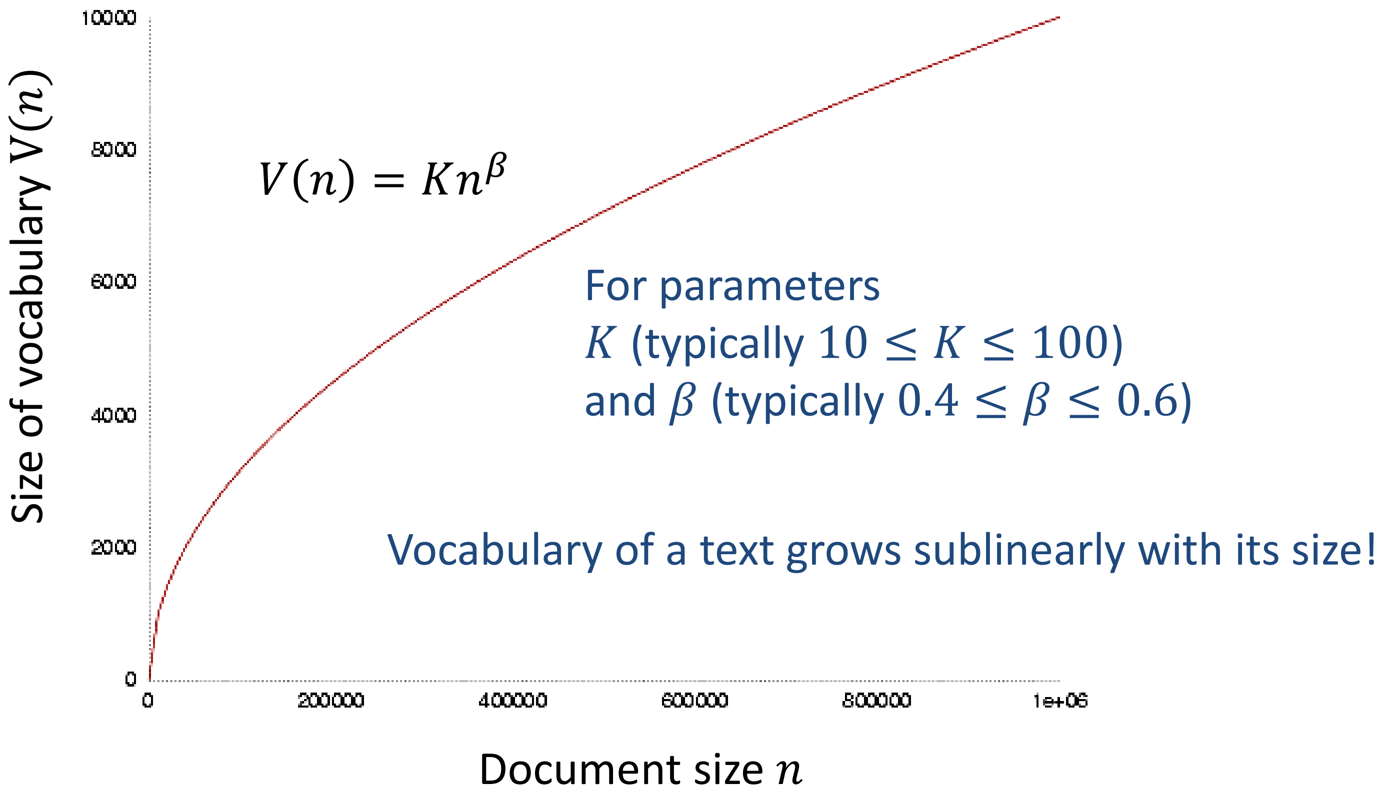
$$f(t) = C\frac{1}{r(t)}$$

C: frequency of the most frequent term
$r(t)$: rank of term t

Source: http://www.ucl.ac.uk/~ucbplrd/language_page.htm

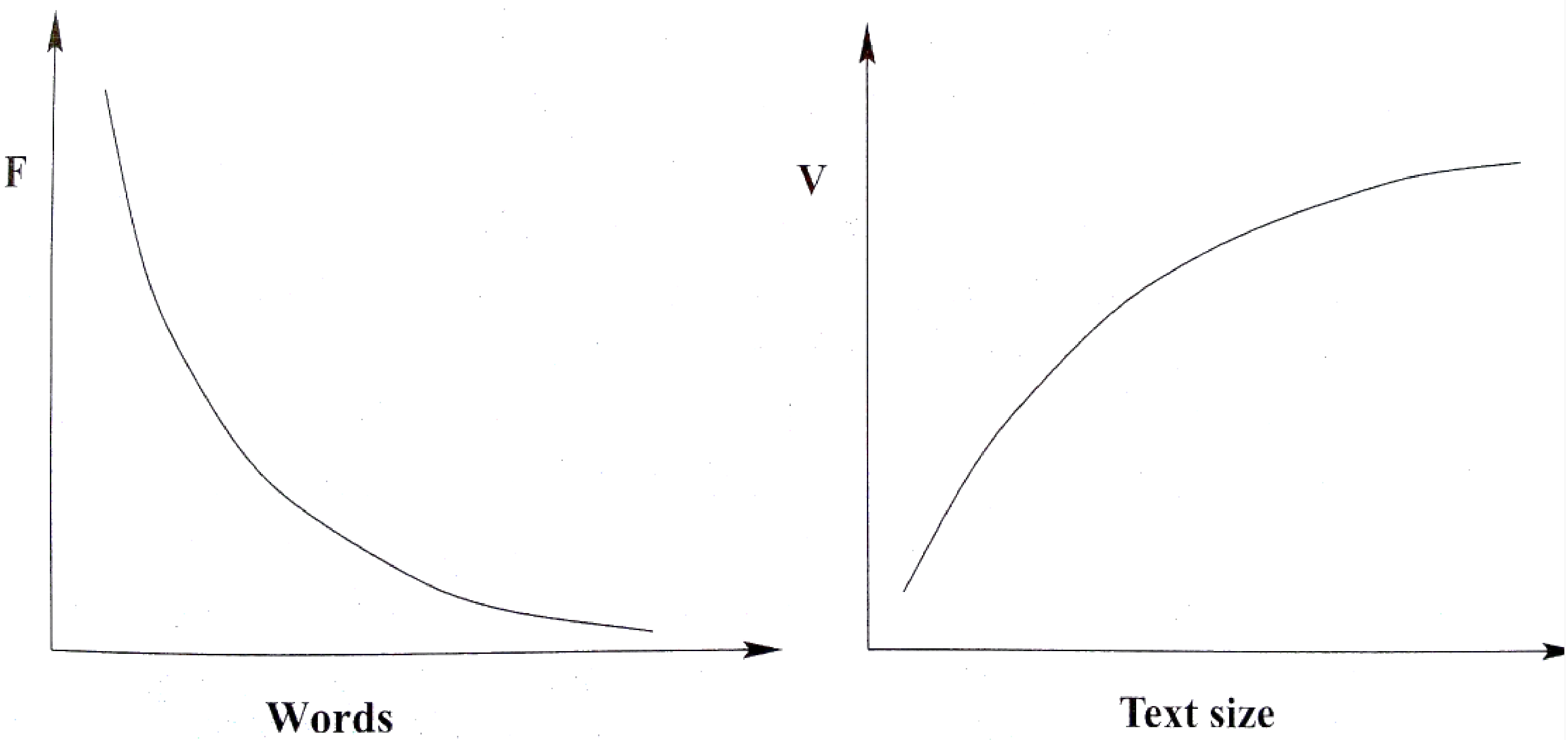
Heap's law

- Empirical law describing the portion of vocabulary captured by a document



See also: [Modern Information Retrieval](), 6.5.2

Zipf's law & Heaps' law

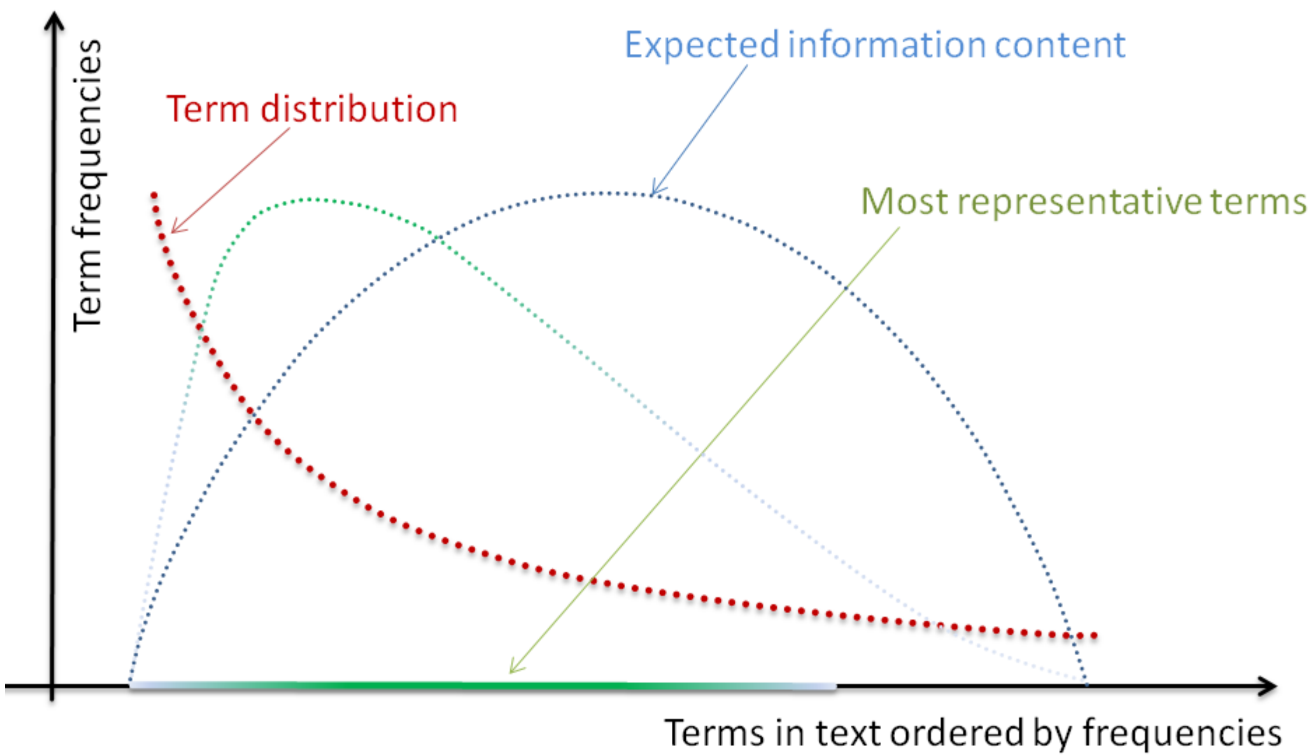


Source: [Modern Information Retrieval]

- Two sides of the same coin …
- Both laws suggest opportunities for compression

Tf-idf weighting scheme (1)

- Which are the terms that best represent a document?
- Is there a weighting scheme that gives higher weights to representative terms?



Tf-idf weighting scheme (2)

- Given a document, **by which terms is it best represented**?
  - Is there a weighting scheme that gives higher weights to representative terms?
  - Consider corpus with documents $D = \{d_1, \dots, d_n\}$ with terms from a vocabulary $V = \{t_1, \dots, t_m\}$.
  - The term frequency of term t_i in document d_j is measured by

$$tf(t_i, d_j) = \frac{freq(t_i, d_j)}{max_k\, freq(t_k, d_j)}$$

Normalisation makes estimation independent of document length.

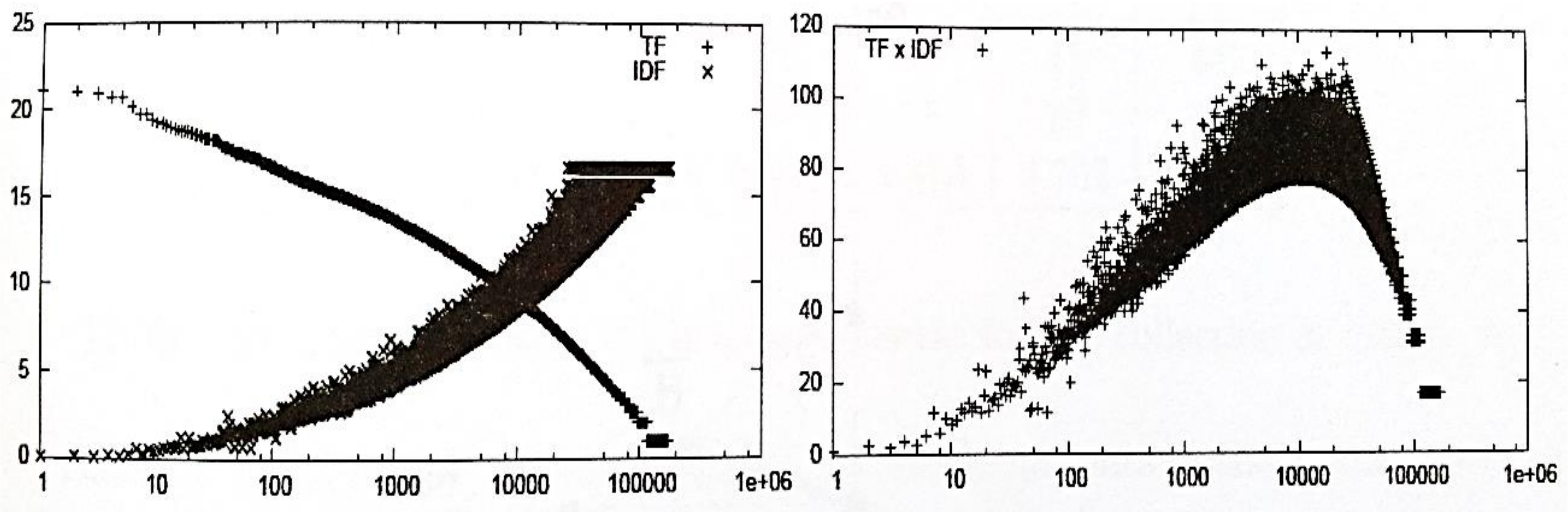
  - The inverse document frequency for a term t_i is measured by

$$idf(t_i, D) = \log \frac{|D|}{|\{d \in D;\, t_i\ occurs\ in\ d\}|}$$

Downweights terms that ocurr in many documents (i.e., **stop words:** the, to, from, if, …).

  - Central weighting scheme for scoring and ranking

Tf, idf, and tf-idf



Tf, idf, and tf-idf weights (plotted in log-scale) computed on a collection from Wall Street Journal (~99,000 articles published between 1987 and 1989)

Source: [Modern Information Retrieval]

Some more thoughts on information content

- We used a syntactic definition of information content
- But information content is not just a syntactic concept
- We also used a closed world assumption where the semantic context of sender and receiver was assumed to be the same and could thus be factored out
- What is the relation between semantics, information content, and communication context?