Power-laws and other heavy-tailed distributions and associated codes which are related to Zipf's law

Flemming Topsøe
University of Copenhagen
Department of Mathematical Sciences
Presentation at the 5th Trier Symposium on Quantitative Linguistics, December 2007

Work reported (first part) is joint work with Peter Harremoës, published in "Entropy", 2001 and in "Lecture Notes in Computer Science", 2006.
The presentation is an expansion of a presentation earlier this year in Erice, Italy.
For more information, consult my homepage at http://www.math.ku.dk/$\sim$topsoe/.

some comments

On the pages following please find the slides exactly as presented at the symposium. Some comments may be in order: As usual, the oral presentation gave room for further comments, and also made it possible to quickly gloss over material of a more technical nature, not all that suitable for presentation in a short talk. The latter applies especially to the page with theorems I, II and III. Readers interested in understanding this material should consult the sources pointed to at the front page and also consult the manuscript “Between Truth and Description” (available at my homepage but also appeared in a proceedings volume) as this manuscript contains a technical correction of an inaccuracy in the “Maximum Entropy Fundamentals” paper.

To assist the reader: Note that the conclusions with results appear right at the beginning, hence the ms. ends somewhat abruptly. After stating the results, the ms. falls into two parts: The first and more difficult part concerns theoretical considerations which lead to the isolation of the distributions one should ideally work with. I reckon that some readers will do well in turning quickly to the second part where speculations in idealized form on the possible emergence of basic semantic elements and associated means for communication are presented.

I shall “soon” work out one or two manuscripts where the ideas are presented at a more leisurely pace, one manuscript being addressed at a more general readership.

Overview of aim, results and limitations

Aim: To understand the basic structure of the "idealized communicator", a person with an infinite vocabulary acquired over time, the "Zipfean".

Limitations: We consider only the primitive semantic structure, that of words. The words are ranked, starting with the most frequent word. Assigning probabilities to the words, we obtain the associated probability distribution $P = (p_1, p_2, \cdots)$ or, equivalently, the associated coding pattern κ (to be explained later).

Any acceptable distribution $P = (p_1, p_2, \cdots)$ for a Zipfean is referred to as a Zipfean distribution. Thus:

Key tasks:

- to identify the Zipfean distributions
- to develop their basic properties
- to explain how they emerge over time.

Results:

- identification: there are infinitely many Zipfean distributions; they are identified in precise mathematical terms, either via their point probabilities or via the associated code lengths
- properties:
  - they can be realized with finite effort per word
  - they imply stability: the Zipfean does not have to change the basic structure over time
  - they ensure that the language is flexible, allowing the Zipfean to increase the expressive power as required for any conceivable specialized purpose
- emergence: this is suggested to be related to a learning hierarchy of distributions or codes.

Limitation: testing of theory is difficult and still lacking.

Distributions and codes

A reminder: (binary) codes without probabilities:

alphabet $\mathbb{A}$	code-word	code-word length (κ)
a	11	2
e	00	2
i	01	2
o	100	3
u	1010	4
y	1011	4

Given possible lengths κ_i, there exists a (prefix-free) code with these lengths iff Kraft's inequality holds:

$$\sum_{i \in \mathbb{A}} 2^{-\kappa_i} \leq 1 .$$

Equality: most natural (compression!). Then there is a duality $probability \leftrightarrow code$ ($P \leftrightarrow \kappa$):

$\kappa_i = \log \frac{1}{p_i}$ (the code length function κ adapted to P)

$p_i = 2^{-\kappa_i}$ (the distribution P matching κ).

One more example: Coding letters in Dickens: *"A tale of two cities"*

Alphabet	**frequency**	**probability**	**optimal (Huffman) code word**	**length**	**ideal length**
e	72883	12.49 %	000	3	3.00
t	52396	8.98 %	010	3	3.48
a	47064	8.07 %	1110	4	3.63
o	45118	7.73 %	1100	4	3.69
n	41310	7.08 %	1101	4	3.82
i	39786	6.82 %	1010	4	3.87
h	38360	6.57 %	1000	4	3.93
s	36772	6.30 %	1001	4	3.99
r	35956	6.16 %	0010	4	4.02
d	27485	4.71 %	0110	4	4.41
l	21523	3.69 %	10110	5	4.76
u	16218	2.78 %	00110	5	5.17
m	14923	2.56 %	00111	5	5.29
w	13835	2.37 %	01110	5	5.40
c	13224	2.27 %	01111	5	5.46
f	13155	2.25 %	111100	6	5.47
g	12120	2.08 %	111101	6	5.59
y	11849	2.03 %	111110	6	5.62
p	9453	1.62 %	101110	6	5.95
b	8140	1.40 %	101111	6	6.16
v	5065	0.87 %	1111110	7	6.85
k	4635	0.79 %	11111110	8	6.98
x	666	0.11 %	1111111101	10	9.77
q	655	0.11 %	1111111100	10	9.80
j	622	0.11 %	1111111110	10	9.87
z	213	0.04 %	1111111111	10	11.42
total = 583.426		**100 %**	**mean length = 4.19**		**H=4.16**

Huffman ≈ *combinatorial entropy* (4.19 bits).

Idealizing ≈ *entropy* (4.16 bits).

Which distributions?

Definition: A distribution $P = (p_1, p_2, \cdots)$ is hyperbolic if it is not dominated by any power law.

Examples Consider a constant K and $P = (p_1, p_2, \cdots)$ of the form

$$p_n = \frac{1}{Z \cdot n(\log n)^K} \quad (\textit{not}\ \frac{1}{Z \cdot n^K}\ )$$

for $n \geq 2$ with Z a normalization constant (never mind about the value of p_1). Then this is a well defined hyperbolic distribution. One finds that this distribution has finite entropy if and only if $K > 2$. □

We shall argue that

the Zipfean distributions are precisely the hyperbolic distributions with finite entropy.

To realize the good sense in this, we shall – in consistency with Zipf's thinking – consider a certain game:

The game of least effort

- between Zipfean and the listener. Zipfean chooses P, listener chooses κ. They fight over average code length, $\Phi(P,\kappa) = \sum_{i\in\mathbb{A}} p_i\kappa_i$ with listener as minimizer, Zipfean as maximizer.

Values of the game satisfy

$$\sup_P \inf_\kappa \Phi(P,\kappa) \le \inf_\kappa \sup_P \Phi(P,\kappa)\,.$$

If equal and finite, the game is in equilibrium.

Clearly (!) $\inf_\kappa \Phi(P,\kappa) = \mathsf{H}(P)$, the entropy of P, hence the Zipfean's value is the MaxEnt-value:

$$\mathsf{H}_{\mathsf{max}}(\mathcal{P}) = \sup_{P\in\mathcal{P}} \mathsf{H}(P) = \sup_{P\in\mathcal{P}} \sum_{i\in\mathbb{A}} p_i \ln\frac{1}{p_i}\,.$$

The listeners value, $\mathsf{R}_{\mathsf{min}} = \mathsf{R}_{\mathsf{min}}(\mathcal{P})$, is the minimum of the specific risks $\mathsf{R}(\kappa|\mathcal{P}) = \sup_{P\in\mathcal{P}} \Phi(P,\kappa)$.

So, under equilibrium, MaxEnt=MinRisk.

Theorem I (equilibrium)

If $\mathcal{P}$ is convex and $\mathsf{H}_{\max}(\mathcal{P}) < \infty$, the game is in equilibrium and the listener has a unique optimal strategy κ^*. The matching distribution P^* defined by $p_i^* = 2^{-\kappa_i^*}$ is the MaxEnt-centre of attraction i.e., for any sequence $(P_n)_{n\geq 1}$ of distributions in $\mathcal{P}$ with $\mathsf{H}(P_n) \to \mathsf{H}_{\max}(\mathcal{P})$, it holds that $P_n \to P^*$.

Theorem II (entropy preservation)

Conditions as above. If P^* is power-dominated, then $\mathsf{H}(P_n) \to \mathsf{H}(P^*)$.

Theorem III (entropy loss)

If P^* is hyperbolic then, for every entropy level h with $\mathsf{H}(P^*) < h < \infty$, there exists a convex model $\mathcal{P}$ with P^* as centre of attraction and with $\mathsf{H}_{\max} = h$. The largest such model is the set of distributions P such that $\Phi(P, \kappa^*) \leq h$ with κ^* the code adapted to P^*, i.e., for all i, $\kappa_i^* = -\ln p_i^*$.

It is the possibility of entropy loss which is of prime interest. For the Zipfean, choosing such a distribution, stability and flexibility is possible at the same time!

Let us follow possible development of the Zipfean:

vocab.	repres.	length
$*$	1	1

primeval w. “universal word”

vocab.	repres.	length
1	1	1
2	11	2
3	111	3
.	.	.
n	$11\cdots1$	n
.	.	.

unary, primitive form

vocab.	repres.	length
1	0	1
2	10	2
3	110	3
.	.	.
n	$11\cdots0$	n
.	.	.

unary w. stop symbol

Criticism: Takes up *far* too much space!
(except for a *very* small vocabulary)

New idea needed:

vocab.	repres.	length
1	1	1
2	10	2
3	11	2
4	100	3
5	101	3
6	110	3
7	111	3
.	.	.
n	1xx$\cdots$ x	$\log n$
.	.	.

binary coding

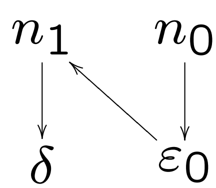
Criticism: Very efficient, can handle large (infinite!) vocabulary. - *But:* no stop symbol or equivalent. No good for conversation, sentences.

Principal difficulty: *Cannot* introduce stopping mechanism without introducing new representation symbols.

Conclusion: *Must* introduce new ideas and renounce on the efficiency expressed by the formula $\kappa_n = \log n$ or the like: $\kappa_n = c + \log n$ or similar.

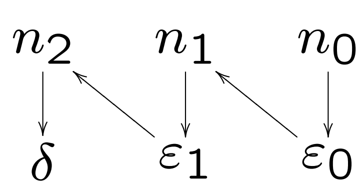
Again, new idea needed:

Mixing of unary and binary representation ...

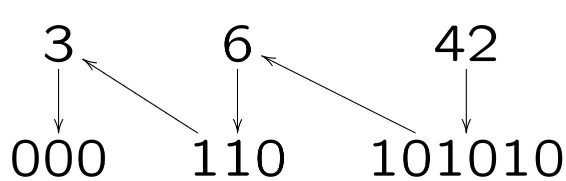


Basic unary/ binary scheme

e.g. 42 $\curvearrowright$ 101010 $\curvearrowright$ 6 $\curvearrowright$ 000000, hence basic order 0 representation of 42 is 000000101010, also denoted $\mathsf{LE}^0(42)$.
Standard Levenshtein-Elias representation corresponds to the scheme: Start with $n = n_0$ and continue:



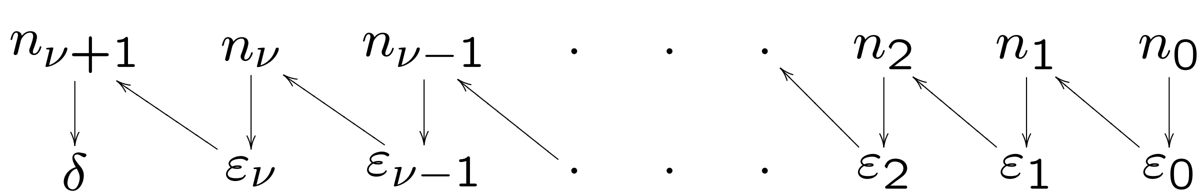
For example: calculation of $\mathsf{LE}(42) = \mathsf{LE}^1(42)$:



vocab.	repres.	length
1	011	3
2	001010	6
3	001011	6
4	0011100	7
5	0011101	7
6	0011110	7
.	. . .	.
n	0xx01x$\cdots$x	$\log n + 2 \cdot \log^{(2)} n$
.	.	.

$\mathsf{LE} = \mathsf{LE}^1$ representation

Finitely iterated LE representation



scheme for ν -fold LE iteration: LE^ν

Completely iterated LE representation

Start by representing 1 and 2. For general n, iterate until you reach ν with $\varepsilon_\nu = 11$ (corresponding to 3).

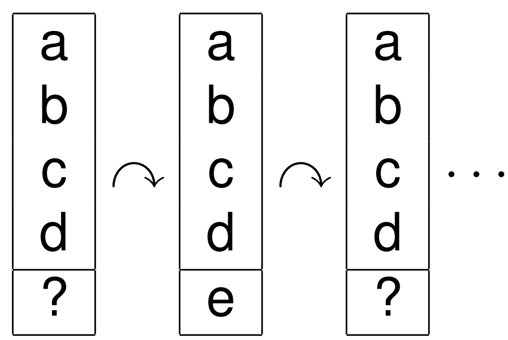
vocab.	repres.	length
1	0	1
2	10	2
3	110	3
4	111000	6
5	111010	6
6	111100	6
7	111110	6
8	1110010000	10
.	.	10
15	1110011110	10
16	11101100000	11
.	.	11
32	111101000000	12
.	.	.
n	1xxx$\cdots$x0	$\log n + \log^{(2)} n + \cdots$
.	.	.

complete iteration LE*. Compact: $\sum 2^{-\kappa_i} = 1$

Convenient decoding

Example: Decoding of $22 = \texttt{10110}$:

+	-	-	-	-	-	-	-	-	-	-	+
-	-	+	-	-	+	-	-	-	-	+	-
0	2	0	4	4	0	16	16	20	22	22	0
1	1	3	2	1	5	4	3	2	1	1	1
1	1	1	0	1	1	0	1	1	0	0	·

Decode column to column: 

a: start of word

b: start of ε-block

c: contribution in block to the “block sum”

d: number of binary digits to be read in block

e: actual binary digit

The learning hierarchi

Coding domain:

LE^0	$2\log n$
$\overline{LE}^0$	$c + K\log n$
LE^1	$\log n + 2\log^{(2)} n$
$\overline{LE}^1$	$c + \log n + K\log^{(2)} n$
LE^2	$\log n + \log^{(2)} n + 2\log^{(3)} n$
$\overline{LE}^2$	$c + \log n + \log^{(2)} n + K\log^{(3)} n$
$\cdots$	$\cdots$

Distribution domain:

special PL	$c\frac{1}{n^2}$
power laws	$c\frac{1}{n^K}$
HYP^1	$c\frac{1}{n(\log n)^2}$
$\overline{HYP}^1$	$c\frac{1}{n(\log n)^K}$
HYP^2	$c\frac{1}{n(\log n)(\log^{(2)} n)^2}$
$\overline{HYP}^2$	$c\frac{1}{n\log n(\log^{(2)} n)^K}$
$\cdots$	$\cdots$