Probability for linguists

John A Goldsmith

July 6, 2015

Overall strategy

1. probabilities and distributions
2. unigram probability
3. a word about *parametric* distributions
4. -1 $\times log_2$ probability (or *plog*: positive log probability)
5. *bigram* probability: conditional probability
6. *mutual information*: the log of the ratio of the observed to the "expected"
7. average plog $\rightarrow$ *entropy*
8. encoding events: compression, optimal compression, and cross-entropy
9. encoding grammars optimally

A distribution

Big point 1

A distribution is a list of numbers that are not negative and that sum to 1.

$$\sum_i p_i = 1$$

$$p_i \geq 0$$

A probabilistic grammar

- A probabilistic model, or grammar, is a universe of possibilities (“sample space”) + a distribution.
- A probabilistic grammar is a distribution over all strings of the IPA alphabet.
- It is not a formalism stating which strings are *in* and which are *out*.

The purpose of a probabilistic model

Big point 2

The purpose of a probabilistic model is to test the model against the data.

- Suppose we have some well-chosen data D. Then the best grammar is the one that assigns the highest probability to D, all other things being equal.
- The goal is not to test the data!
- Therefore: all grammars must be probabilistic, so they can be tested and evaluated.

Probability

- The *quantitative theory of evidence.*
- If we have *variable* data, then probability is the best model to use.
- If we have *categorical* (not variable) data, probability is still the best model to use.

Probabilities and frequencies

Probabilities and frequencies are not the same thing.

- Frequencies are *observed*.
- Probabilities are values in a system that a human being creates and *assigns*.
- We can choose to assign probabilities as the observed frequencies—buy that is not always a good idea.
- This is a good idea only so long as we don't need to handle yet-unseen (never before seen) data.
- In many cases, this choice maximizes the probability of the data.
- They both deal with *distributions* (i.e., the observed frequencies and the probability distributions of a model).

Probabilities and frequencies

Probabilities and frequencies are not the same thing.

- *Counts* are counts: the number of things or events that fall in some category.
- *Frequency* is ambiguous: it either means count (less often) or it means *relative frequency*: a ratio between a count of something and the total number of things that fall within the larger category.
- There are 63,147 occurrences of *the* in the Brown Corpus, out of 1,017,904; 6.2% of the words in the Brown Corpus are *the*.

English, French, Spanish

Let's take a look at some languages.
And for starters, let's just look at *unigram* frequencies: the frequencies at which items appear, not conditioned by the environment.

`people.cs.uchicago.edu/jagoldsm/course/class1`

Plogs

- We will assign probabilities to every outcome we consider.
- Each of these is typically quite small.
- We therefore use a slightly different way of talking about small numbers: plogs.

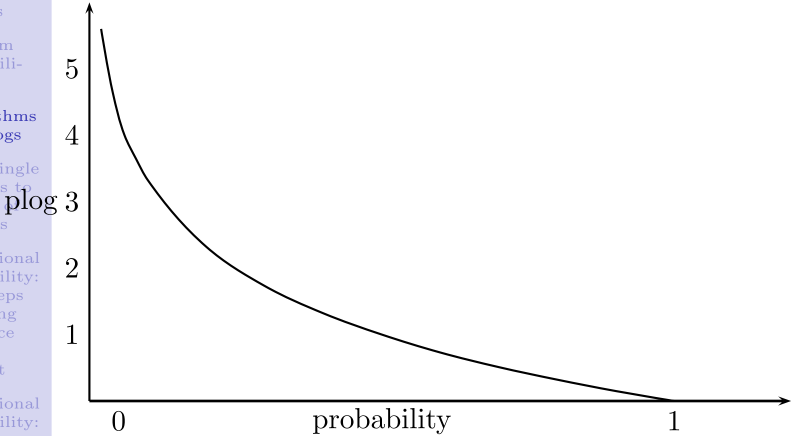
Inverse log probabilities, or *plogs*

A way to describe small numbers... upside down.

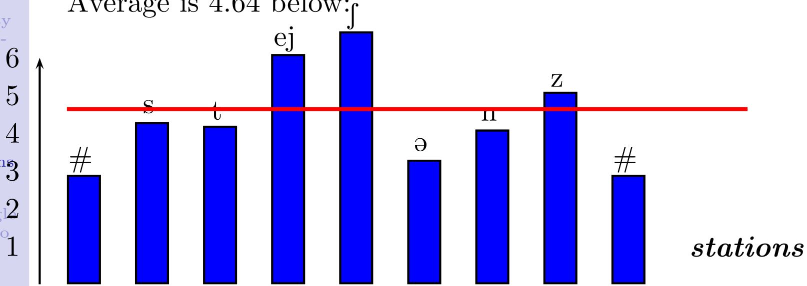
A probability	its plog
0.5	1
0.25	2
0.128	3
$\frac{1}{16}$	4
$\frac{1}{32}$	5
$\frac{1}{1024}$	10
...	...
$\frac{1}{1{,}000{,}000}$	almost 20

- The *bigger* the plog, the *smaller* the probability.
- It's a bit like a measure of markedness, if you think of more marked things as being less frequent.
- $plog(x) = -log_2(x) = log_2(\frac{1}{x})$

Plogs



Average is 4.64 below:



This diagram from a visually interactive program displaying phonological complexity at:

`http://hum.uchicago.edu/~jagoldsm/PhonologicalComplex`

Most and least frequent phonemes in English

rank	phoneme	frequency	plog
1	#	0.20	2.30
2	ə	0.066	3.92
3	n	0.058	4.10
4	t	0.056	4.17
5	s	0.041	4.61
6	r	0.040	4.76
7	d	0.037	4.85
8	l	0.035	4.94
9	k	0.026	5.27
10	ǽ	0.025	5.31
45	ɔ́y	0.000 78	10.32
46	æ̆	0.000 69	10.50
47	ž	0.000 54	10.84
48	ay̆	0.000 38	11.36
49	ă	0.000 36	11.42

average plogs

rank	orthography	phonemes	$av. plog_1$
1	a	ə	3.11
2	an	ən	3.44
3	to	tə	3.47
4	and	ənd	3.80
5	eh	έ	3.88
6	the	ə	3.88
7	can	kən	3.90
8	an	ǽn	3.91
9	Ann	ǽn	3.91
10	in	ín	3.91

Worst words in English

rank	orthography	phonemes	$av. plog_1$
63,195	bourgeois	bʌ̆ržwá	7.21
63,196	Ceausescu	čɔ̆čɛ́skŭ	7.21
63,197	Peugeot	p yŭžó	7.22
63,198	Giraud	žay̆ró	7.24
63,199	Godoy	gádoy̆	7.27
63,200	geoid	ǰíɔy̆d	7.40
63,201	Cesare	čĕzárĕ	7.40
63,202	Thurgood	θɚ́gʌ̆d	7.47
63,203	Chenoweth	čɛ́nɔ̆wɛ̆θ	7.49
63,204	Qureshey	kəréšĕ	7.54

Word counts and frequencies

	word	count	frequency	plog
1	the	69903	0.068271	3.87
2	of	36341	0.035493	4.81
3	and	28772	0.028100	5.15
4	to	26113	0.025503	5.29
5	a	23309	0.022765	5.46
6	in	21304	0.020807	5.59
7	that	10780	0.010528	6.57
8	is	10100	0.009864	6.66
9	was	9814	0.009585	6.70
10	he	9799	0.009570	6.70
11	for	9472	0.009251	6.77
12	it	9082	0.008870	6.82
13	with	7277	0.007107	7.14
14	as	7244	0.007075	7.14
15	his	6992	0.006829	7.19

Unigram model

- The probability of a string S, of length L, is $\lambda(L)$ times the probability of each of the symbols.
- $p_U(S) = \lambda(L) \times \prod_i S[i]$
- If we sum over *all* strings of a given length l, the sum of their probabilities is $\lambda(l)$. That's just math.
- This is the model that takes no information about ordering into account.
- Because plogs are additive, it makes sense to ask what the average plog of a word is. In the unigram model, they describe an extensive property.

Conditional probabilty

- p(A, given B)
- p($A|B$)
- $\frac{p(A\,and\,B)}{p(B)}$
- p(A's name is "John") < p(A's name is "John" given that A is male and American)
- p(A=Queen of hearts)
- p(A=Queen of hearts | A is a red card)

Conditional probability in a string

- p(S[i]=h given that S[i-1]=t)
- p(S[i]=h | S[i-1]=t)
- p (S[i]=book | S[i-1] = the) > p(S[i]=book)
- p (S[i]=the | S[i+1]=book) > p (S[i]=book)
- These are not statements of *causality*.

Addition is easier to understand than multiplication

- In the unigram model, the probabilility of the string = product of the probabilities of its symbols.[1]
- If we use plogs, the log probability of the string is the sum of the plogs of its symbols.

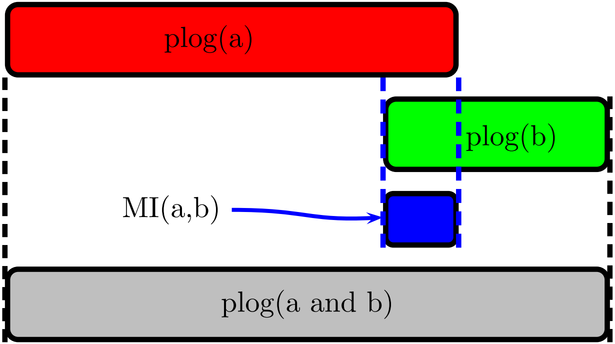
[1] ignoring length of string. . .

Using plogs with conditional probability

- The probability goes up when we use a better model (i.e., one that encodes more knowledge about the system) that takes into consideration the factors in the neighborhood that helped lead to the events we saw.
- The bigram conditional probability is usually greater than the unigram probability in real data.
- The difference between the bigram plog and the unigram plog is called the *mutual information* (MI).

$$log\frac{p(AandB)}{p(A)p(B)} = log\frac{p(AandB)}{p(A)}\frac{1}{p(B)} = log\, p(B|A) - log\, p(B)$$

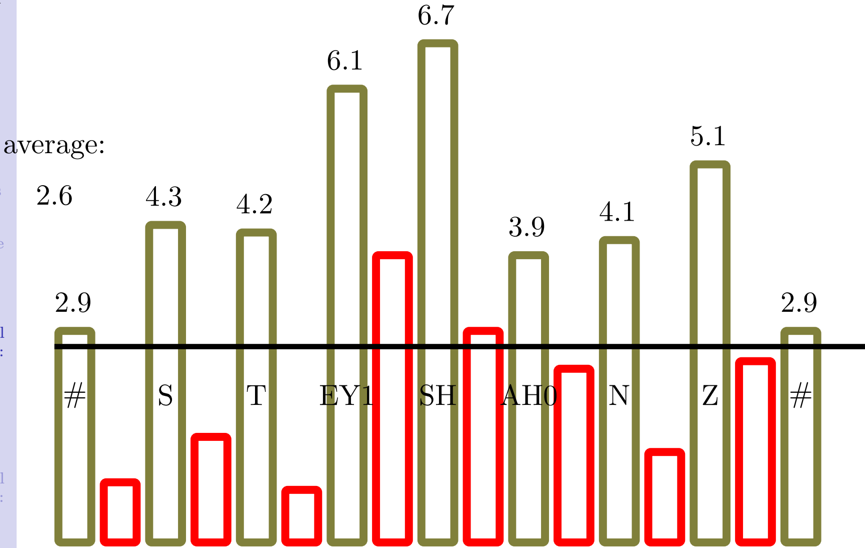
Pointwise mutual information (MI)



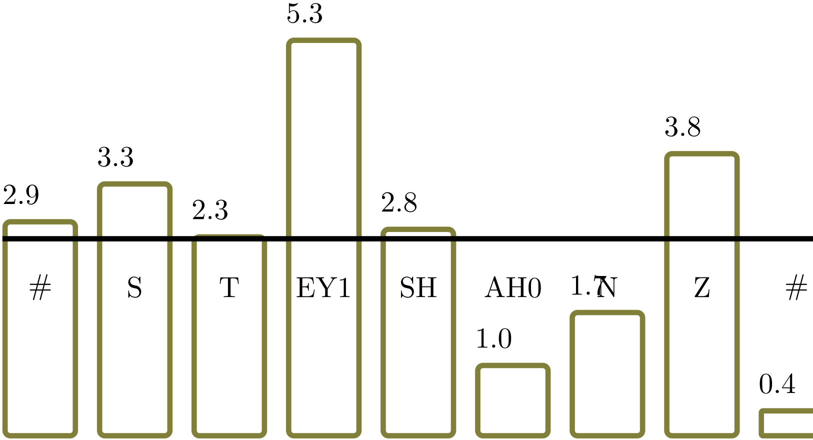
A reminder about events, and “a & b”

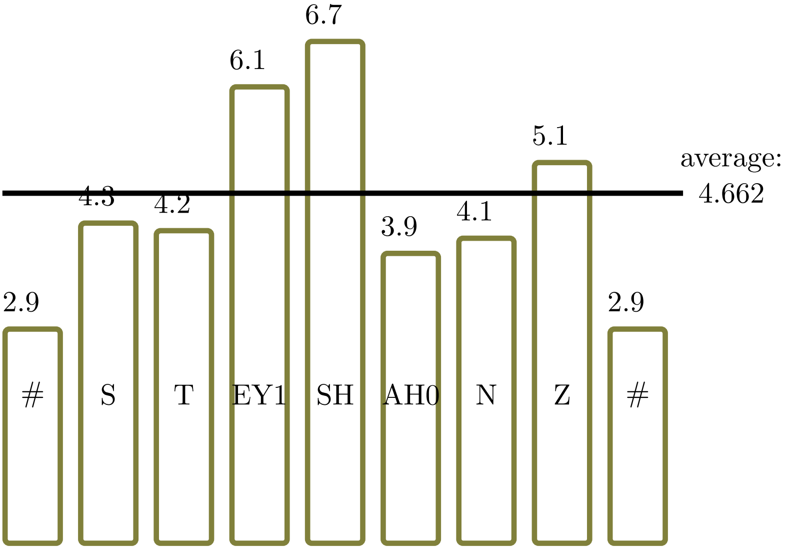
- There is no implicit statement about location of the events when we write “a & b”.
- p(W[i] = “of” & W[i+1]=“the”)
- p(W[i] = “of” & W[i+5] = “the”)
- If we look at the second, the MI will be very close to zero.

Unigram model with MI



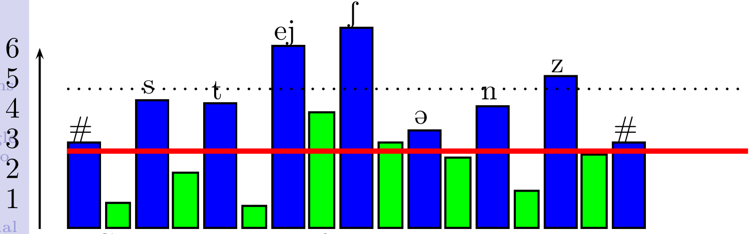
Bigram model



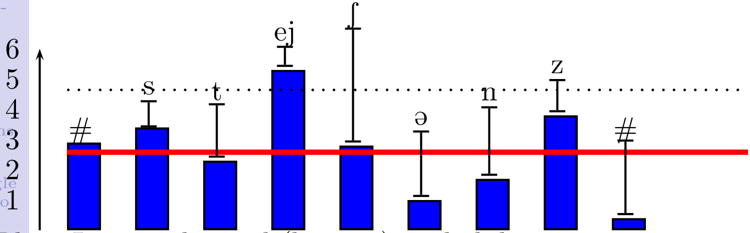

2.9
4.3
4.2
6.1
6.7
3.9
4.1
5.1
2.9
average:
4.662
#
S
T
EY1
SH
AH0
N
Z
#

- $p_U = \prod p(S[i])$
- $= p_U(\text{thecatisonthemat})$
- $= \text{p}_U(t) \times p_U(h) \times p_U(e) \times p_U(c) \ldots \times p_U(t)$
- $= p_U(a) \times p_U(a) \times p_U(c) \times p_U(e) \times p_U(e) \ldots \times p_U(t)$
- $=(\text{p}_U(a))^2 \times p_U(c) \times (p_U(e))^2 \times (p_U(e))^2 \ldots \times p_U(t)$
- $=\prod_{\text{l in alphabet A}} p(a)^{\text{count of l in string}}$

Average below is 2.58 (down from 4.64)



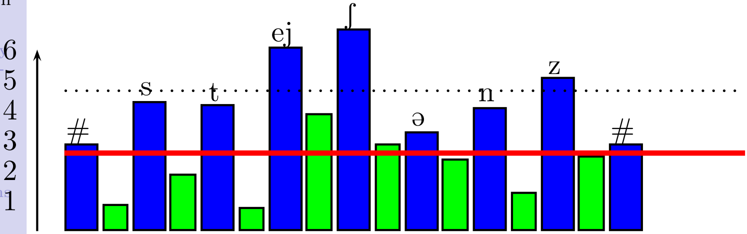
Green: Mutual information in *stations*
Blue: Unigram plot in *stations*



Blue: Log conditional (bigram) probability in *stations*

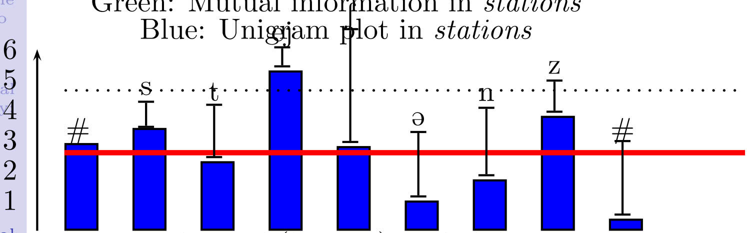
Decrease from unigram model is exactly the mutual information

Average below is 2.58 (down from 4.64)



Green: Mutual information in *stations*

Blue: Unigram plot in *stations*

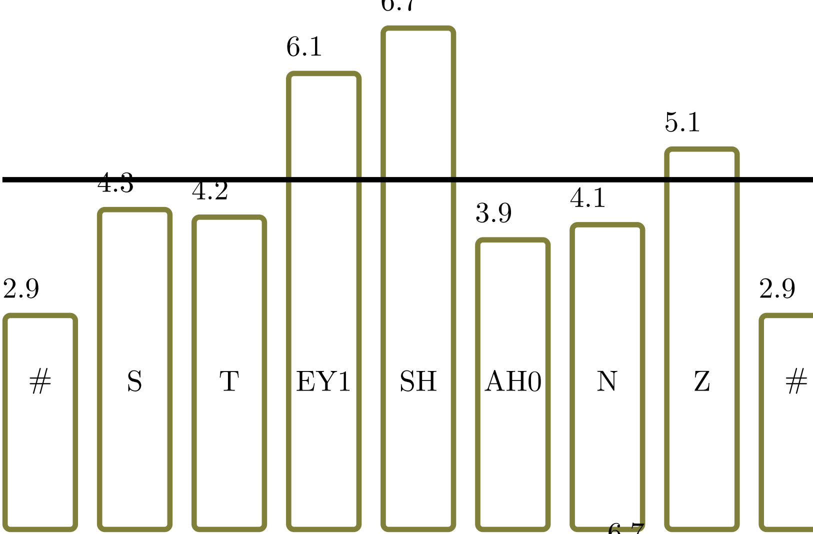


Blue: Log conditional (bigram) probability in *stations*

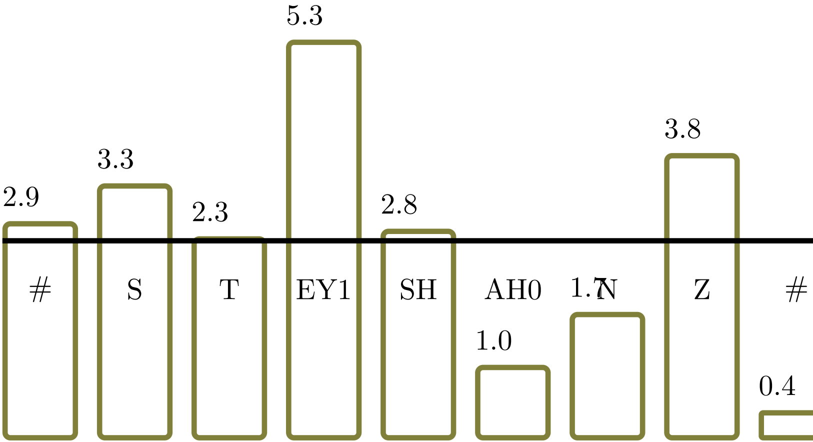
Decrease from unigram model is exactly the mutual information

Using plogs with conditional probability

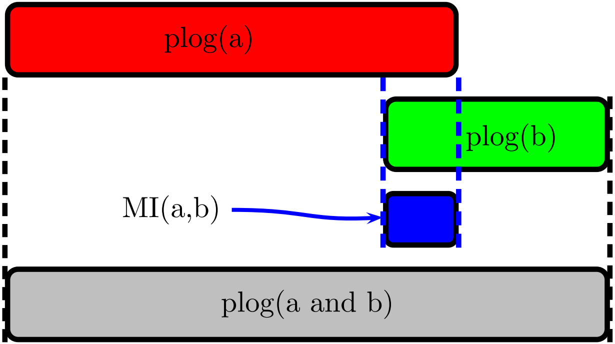
- We saw that the probability goes up when we use a better model that takes into consideration the factors in the neighborhood that helped lead to the events we saw.
- The bigram conditional probability is usually greater than the unigram probability in real data.
- The difference between the bigram plog and the unigram plog is called the *mutual information* (MI).



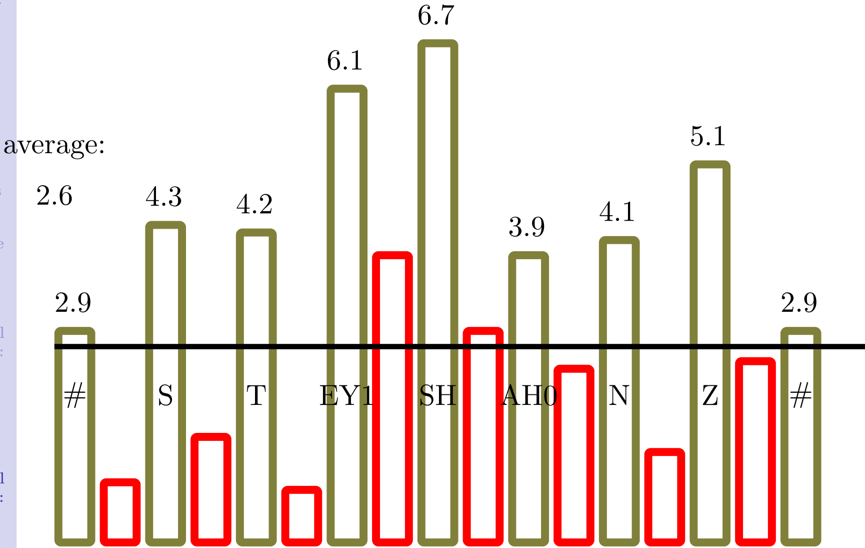
average:


2.9
3.3
2.3
5.3
2.8
1.0
1.7
3.8
0.4
#
S
T
EY1
SH
AH0
N
Z
#

Pointwise mutual information (MI)



Unigram model with MI



Word counts and frequencies: repeated

	word	count	frequency	plog
1	the	69903	0.068 271	3.87
2	of	36341	0.035 493	4.81
3	and	28772	0.028 100	5.15
4	to	26113	0.025 503	5.29
5	a	23309	0.022 765	5.46
6	in	21304	0.020 807	5.59
7	that	10780	0.010 528	6.57
8	is	10100	0.009 864	6.66
9	was	9814	0.009 585	6.70
10	he	9799	0.009 570	6.70
11	for	9472	0.009 251	6.77
12	it	9082	0.008 870	6.82
13	with	7277	0.007 107	7.14
14	as	7244	0.007 075	7.14
15	his	6992	0.006 829	7.19

Top of the Brown Corpus for words following *the*

	word	count	count / 69,936
0	first	664	0.009 49
1	same	629	0.008 99
2	other	419	0.005 99
3	most	419	0.005 99
4	new	398	0.005 69
5	world	393	0.005 62
6	united	385	0.005 51
7	state	271	0.004 18
8	two	267	0.003 82
9	only	260	0.003 72
10	time	250	0.003 57
11	way	239	0.003 42
12	old	234	0.003 35
13	last	223	0.003 19
14	house	216	0.003 09

Top of the Brown Corpus for words following *of*.

	word	count	count / 36,388
1	the	9724	0.267
2	a	1473	0.040 5
3	his	810	0.022 3
4	this	553	0.015 20
5	their	342	0.009 40
6	course	324	0.008 90
7	these	306	0.008 41
8	them	292	0.008 02
9	an	276	0.007 58
10	all	256	0.007 04
11	her	252	0.006 93
12	our	251	0.006 90
13	its	229	0.006 29
14	it	205	0.005 63
15	that	156	0.004 29

Cross entropy: where we keep the empirical frequencies, but vary the distribution whose plog we use to compute the entropy. This is the "cross-entropy" of one distribution to the other (but not symmetrical!). Entropy, or self-entropy, is always smaller than cross-entropy.

$$\sum_x p(x) ln \frac{q(x)}{p(x)} \leq \sum_x p(x)(1 - \frac{q(x)}{p(x)}) \quad (1)$$

Why? Look at the plot of $ln(x)$, and compute its first and second derivatives, and its value at (1,0).

$$= \sum_x p(x) - \sum_x p(x) \frac{q(x)}{p(x)} = 1 - 1 = 0. \quad (2)$$

So $\sum_x p(x)ln(\frac{q(x)}{p(x)} \leq 0$, which is to say, the cross-entropy always exceeds the entropy that isn't cross, when we use natural logs as our base.

But we can maintain the inequality when we switch to base 2 logs (which is what we use with plogs), since it just amounts to multiplying both sides by a constant. First we get:

$$\sum_x p(x) ln\, q(x) \leq \sum_x p(x) ln\, p(x) \tag{3}$$

and then we multiply by -1:

$$\sum_x p(x) plogp(x) \leq \sum_x p(x) plog\, q(x) \tag{4}$$

The Kullback-Leibler divergence $D_{KL}(p, q)$ is defined as KL divergence

$$\sum_x p(x)\, ln\, \frac{p(x)}{q(x)} \tag{5}$$

You see that it's the difference between the cross-entropy and the self-entropy—pay careful attention to the *absence* of a minus before the sum.

$$\prod_{i=1}^{i=len(string)} S[i] = \prod_{l \in lexicon} l^{count_S(l)}. \quad (6)$$

$$logprob(S) = \sum_{lexicon} count_S(l) logprob(l). \quad (7)$$

$$plog(S) = \sum_{lexicon} count_S(l) plog(l). \quad (8)$$

If we divide through by the length of our string, we get the average which is Shannon's entropy:

$$entropy(S) = \sum_{lexicon} freq_S(l)\, plog(l). \quad (9)$$

This is more familiar if we write $-\sum p(x) logp(x)$.

cross-entropy of two distributions

$$-\sum_{x \in X} \mathrm{p(x)} \log \mathrm{q(x)}. \tag{10}$$

cross-entropy is less than self-entropy

- p() and q() are two different distributions.
- How do $-\sum$ p(x) log p(x) and $-\sum$p(x) log q(x) compare?
- $-\sum$ p(x) log p(x) $+\sum$p(x) log q(x) $=\sum$ p(x) log $\frac{q(x)}{p(x)}$
- Suppose we use natural logs: then we know that $ln(x) \leq (x-1)$.
- $\sum$ p(x) log $\frac{q(x)}{p(x)} \leq \sum$ p(x) [$\frac{q(x)}{p(x)}$ - 1] $=$ $\sum p(x) - \sum q(x) = 1 - 1 = 0$
- So $-\sum$ p(x) log p(x) (the entropy) is always smaller than the cross-entropy $-\sum$p(x) log q(x)