Axiom. *Linguistic phenomena tend to follow Zipf's law.*

Lemma. *Good-Turing adjusted counts are pretty good.*

Proof. General consensus of language researchers after decades of use. □

Theorem. *Laplace adjusted counts are bad.*

Proof. (After Gale and Church, 1994.) Suppose Laplace adjusted counts were good. Then, by the Lemma and the transitivity of goodness, they would be similar to GT adjusted counts, that is,

$$\frac{(r+1)\cdot N}{N+V} \approx \frac{(r+1)\cdot n_{r+1}}{n_r}$$

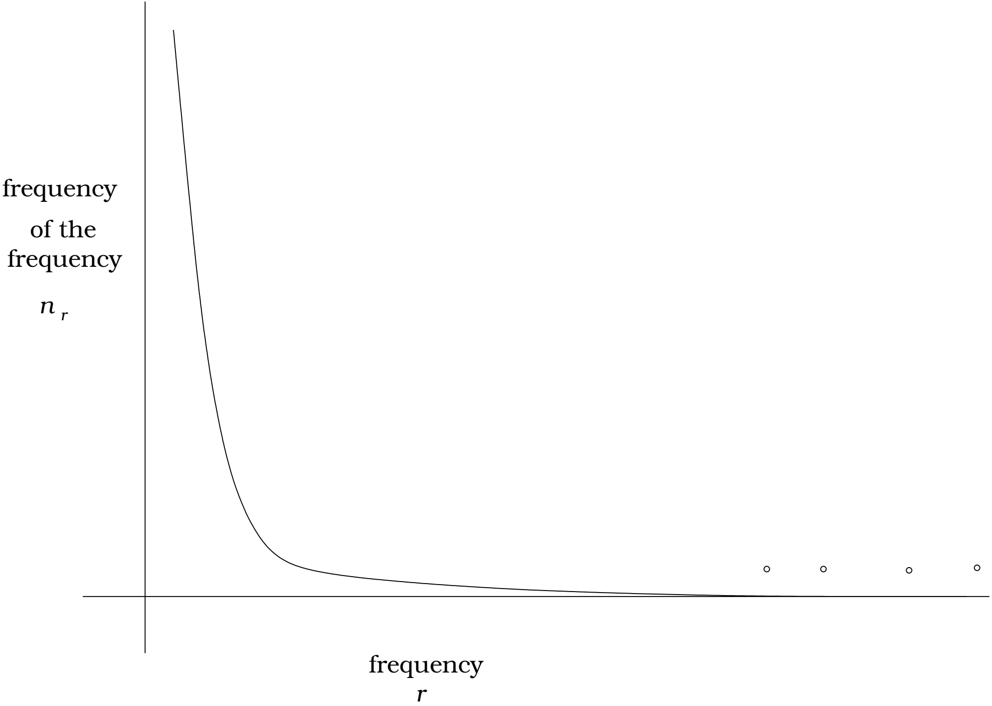
Let $\rho = \frac{N}{N+V}$. Then

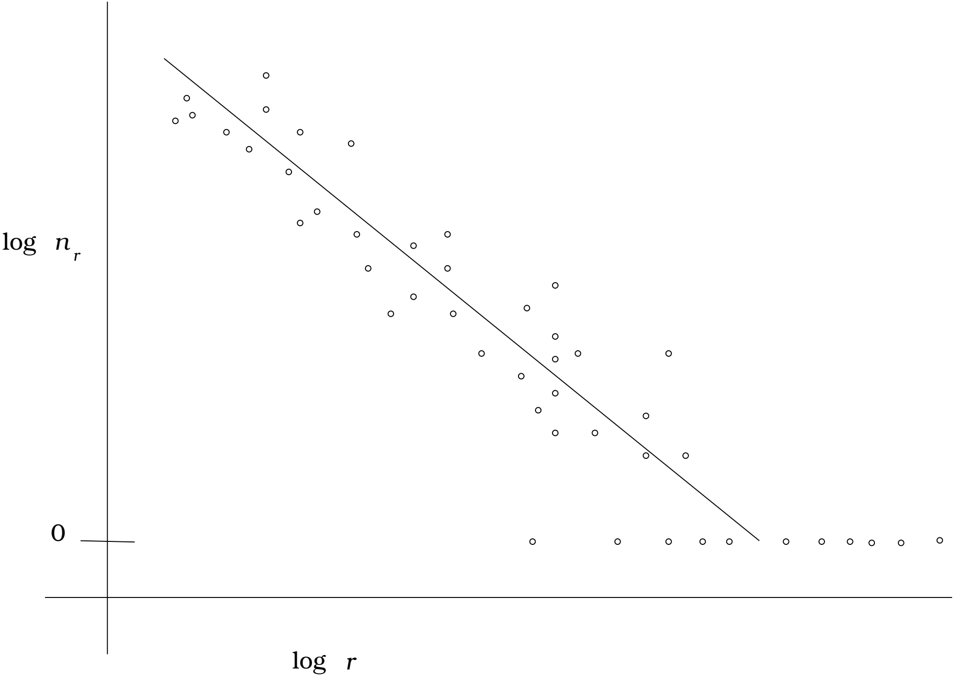
$$\begin{aligned}
n_1 &= \rho \cdot n_0 \\
n_2 &= \rho \cdot n_1 = \rho^2 \cdot n_0 \\
n_r &= \rho^r \cdot n_0 \\
\log n_r &= \log(\rho^r \cdot n_0) \\
&= \log \rho^r + \log n_0 \\
&= r \cdot \log \rho + \log n_0
\end{aligned}$$

But Zipf's law ($r \cdot f = c$) predicts

$$\begin{aligned}
r \cdot n_r &= c && \text{for some } c \\
\log(r \cdot n_r) &= \log c = d && \text{for some } d \\
\log r + \log n_r &= d \\
\log n_r &= d - \log r
\end{aligned}$$

Thus Laplace smoothing assumes $\log n_r$ is linearly related to r (n_r and r makea a straight line on a semi-log graph), but Zipf's law predicts that $\log n_r$ is linearly related to $\log r$ (n_r and r make a straight line on a log-log graph). Hence Laplace smoothing is not good. At least, it is not Good-Turing. □


frequency
of the
frequency
n_r
frequency
r



log n_r
0
log r

Katz's k cut off for $k = 5$, intuitive but wrong version:

adjusted count	$\frac{n_1}{n_0}$	$\frac{2 \cdot n_2}{n_1}$	$\frac{3 \cdot n_3}{n_2}$	$\frac{4 \cdot n_4}{n_3}$	$\frac{5 \cdot n_5}{n_4}$	$\frac{6 \cdot n_6}{n_5}$	6	$\cdots\cdots\cdots$	100
count	0	1	2	3	4	5	6	$\cdots\cdots\cdots$	100
	use GT						use MLE		

Katz's k cut off, constrained to make it a probability function:

$$\underbrace{1 = \sum_{w} P(w)}_{\text{need}} = \underbrace{\sum_{w \mid c(w)=0} P(w)}_{\text{unseen words, keep GT}} + \underbrace{\sum_{w \mid 1 \leq C(w) \leq k} P(w)}_{\text{rare words, adjust GT}} + \underbrace{\sum_{w \mid C(w) > k} P(w)}_{\text{common words, keep MLE}}$$

Katz's k cut off, constrained to make it a probability function:

$$\sum_{w \mid c(w)=0} P_{GT}(w) \quad = \quad \sum_{w \mid 1 \leq w \leq k} (P_{MLE}(w) - P_{GTS}(w))$$

$$\underbrace{\frac{n_1}{N}}_{\text{total GT prob of unseens}} \quad = \quad \underbrace{\sum_{i=1}^{k}}_{\text{summation over frequencies, not types}} \left(\underbrace{\frac{n_i \cdot i}{N}}_{\text{total MLE prob for freq } i} - \underbrace{\mu}_{\text{scaling factor, what we want to find}} \cdot \frac{(i+1) \cdot n_{i+1}}{N} \right)$$

$$= \quad \sum_{i=1}^{k} n_i \cdot \left(\frac{i}{N} - \frac{\mu \cdot (i+1) \cdot n_{i+1}}{N \cdot n_i} \right)$$

$$n_1 \quad = \quad \sum_{i=1}^{k} i \cdot n_i - \mu \cdot \sum_{i=1}^{k} \frac{n_i + 1}{n_i}$$

Katz's k cut off, formlua to grab off the shelf and use:

$$P_{GT-Katz}(w) = \begin{cases} \frac{n_1}{N \cdot n_0} & \text{if } C(w) = 0 \\ \frac{(r+1) \cdot \frac{n_{r+1}}{n_r} - r \cdot \frac{(k+1) \cdot n_{k+1}}{n_1}}{N \cdot \left(1 - \frac{(k+1) \cdot n_{k+1}}{n_1}\right)} & \text{if } 1 \leq r = C(w) \leq k \\ \frac{C(w)}{N} & \text{otherwise} \end{cases}$$

Linear interpolation, book version (combining uni-, bi-, and trigrams):

$$P_{LI}(w_n \mid w_{n-2}w_{n-1}) = \lambda_1 \cdot P(w_n \mid w_{n-2}w_{n-1}) + \lambda_2 \cdot P(w_n \mid w_{n-1}) + \lambda_3 \cdot P(w_n)$$

Simplest example (combining unigram MLE and constant):

$$P_{LI}(w) = \lambda \cdot P_{MLE}(w) + (1 - \lambda)\frac{1}{v}$$

General form (any k language models):

$$P_{LI}(w) = \sum_{j=1}^{k} \lambda_j P_j(w)$$