PRINCIPLES OF COMPLEX SYSTEMS

HW05 WRITE-UP

October 6, 2018

David W. Landay
University of Vermont
Graduate Student, Comp. Systems and Data Sci.

0.1 Problem 1

Simpson Concentration:

$$S = \sum_{i=1}^{n} p_i^2$$

$$D_S = \frac{1}{S} = \frac{1}{\sum_{i=1}^{n} p_i^2} \therefore \tag{1}$$

Gini Index:

$$G = 1 - S = 1 - \sum_{i=1}^{n} p_i^2$$

$$\implies S = 1 - G \rightarrow D_S = \frac{1}{1 - (1 - S)}$$

$$\implies D_S = D_G = \frac{1}{\sum_{i=1}^{n} p_i^2} \therefore \tag{2}$$

Shannon's Entropy:

$$H = -\sum_{i=1}^{n} p_i \ln p_i$$

$$= -\sum^{D_H} \frac{1}{D_H} \ln \frac{1}{D_H}$$

$$\implies H = -\ln \frac{1}{D_H}$$

$$\implies \frac{1}{D_H} = e^{-H}$$

$$\implies D_H = e^H = e^{-\sum_{i=1}^{n} p_i \ln p_i} \therefore \tag{3}$$

Renyi Entropy:

$$H_q^{(R)} = \frac{1}{q-1}\left(-\ln \sum_{i=1}^{n} p_i^q\right)$$

$$= \frac{1}{1-q}\left(\ln \sum_{i=1}^{n} p_i^q\right)$$

$$= \ln \sum_{i=1}^{n} \left(p_i^q\right)^{\frac{1}{1-q}}$$

$$\implies H_q^{(R)} = \ln \sum^{D_{H_q^{(R)}}} \left(\frac{1}{D_{H_q^{(R)}}}^{q}\right)^{\frac{1}{1-q}}$$

$$\implies e^{-H_q^{(R)}} = \left(\frac{1}{D_{H_q^{(R)}}}^{q}\right)^{\frac{1}{1-q}}$$

$$\implies D_{H_q^{(R)}} = e^{H_q^{(R)}} = e^{\ln \sum_{i=1}^{n} \left(p_i^q\right)^{\frac{1}{1-q}}} \quad \therefore \tag{4}$$

Generalized Tsallis Entropy:

$$H_q^{(T)} = \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right)$$

$$= \frac{1}{q-1}\left(1 - \sum^{D_{H_q^{(T)}}} \frac{1}{D_{H_q^{(T)}}}^q\right)$$

$$\implies H_q^{(T)} = \frac{1}{q-1}\left(1 - D_{H_q^{(T)}}^{1-q}\right)$$

$$\implies 1 - H_q^{(T)}(q-1) = D_{H_q^{(T)}}^{1-q}$$

$$\implies \left(1 - H_q^{(T)}(q-1)\right)^{\frac{1}{1-q}} = D_{H_q^{(T)}}$$

$$\implies \left(1 - \frac{1}{q-1}\left(1 - \sum_{i=1}^{n} p_i^q\right)(q-1)\right)^{\frac{1}{1-q}} = D_{H_q^{(T)}}$$

$$\implies D_{H_q^{(T)}} = \sum_{i=1}^{n}\left(p_i^q\right)^{\frac{1}{1-q}} \therefore \tag{5}$$

We would like to show that as the degree $q \to 1$ in the generalized form of Tsallis Entropy, the Diversity of the Tsallis Entropy matches with that of the generalized Shannon Entropy. We will begin by taking the limit as $q \to 1$ of Shannon's Entropy. Recall,

$$D_H = e^H = e^{-\sum_{i=1}^{n} p_i \ln p_i} = -\sum_{i=1}^{n} p_i^{p_i}$$

So, $q = p_i$, and as we take the limit as $p_i \to 1$, then D_H clearly goes to $-\infty$ because we have the negation of an infinite sum of 1's.

For the generalized Tsallis Entropy, we can use L'Hôspital's rule to find the limit of

$D_{H_q^{(T)}}$ as $q \to \infty$. We find the limit to be

$$\lim_{q\to\infty} D_{H_q^{(T)}} = \lim_{q\to\infty} \sum_{i=1}^{n} \left(p_i^q \right)^{\frac{1}{1-q}}$$

$$\frac{\mathbf{d}D_{H_q^{(T)}}}{\mathbf{d}p_i}, \to \sum_{i=1}^{n} \frac{1}{1-q} \left(p_i^q \right)$$

$$\frac{\mathbf{d^2}D_{H_q^{(T)}}}{\mathbf{d}p_i^2}, \to \sum_{i=1}^{n} \frac{q}{1-q} \left(p_i^{q-1} \right)$$

but, as $q \to 1,\ p_i^{q-1} = 0$

$$\implies \frac{\mathbf{d^2}D_{H_q^{(T)}}}{\mathbf{d}p_i^2}, \to \sum_{i=1}^{n} \frac{q}{1-q} = -\infty \therefore \tag{1.f}$$

We can see that for Tsallis Entropy, as $q \to \infty$, we get an infinity sum of negative values. Hence, both entropy measures tend to $-\infty$ as $q \to 1$

0.2 Problem 2

We want to minimize:

$$\Psi(\, p_1, p_2, \ldots, p_n\,) \;=\; \mathbf{F}(\, p_1, p_2, \ldots, p_n\,) \;+\; \lambda\, \mathbf{G}(\, p_1, p_2, \ldots, p_n\,) \tag{1}$$

We are given the constraint function

$$\mathbf{G}(\, p_1, p_2, \ldots, p_n\,) \;=\; \sum_{i=1}^{n} p_i \;-\; 1, \;\; \text{which equals 0 for large } n, \tag{2}$$

and that

$$\mathbf{F}(\, p_1, p_2, \ldots, p_n\,) \;=\; \frac{C}{H} \;=\; \frac{\sum_{i=1}^{n} p_i \ln(i \;+\; a)}{-g \sum_{i=1}^{n} p_i \ln(p_i)} \tag{3}$$

where C and H are the cost and entropy, respectively. From the constraint (2) we are given that $\sum_{i=1}^{n} p_i \;=\; 1$. In order to minimize (1), we want to find a value(s) $i \;=\; j$, such that a p_j minimizes Ψ. Hence, it suffices to find when

$$\frac{\partial \Psi}{\partial p_j} \;=\; 0 \;=\; \frac{\partial \mathbf{F}}{\partial p_j} \;+\; \lambda \times 1 \;\Longrightarrow\; \lambda \;=\; -\frac{\partial \mathbf{F}}{\partial p_j} \tag{4}$$

Hence, we expect to find the value of λ with respect to H and C. We will first calculate $\frac{\partial \mathbf{F}}{\partial p_j}$, which is governed by

$$\frac{\partial C}{\partial p_j} \;=\; \ln(\, j + a\,)$$

$$\frac{\partial H}{\partial p_j} \;=\; -g \ln(\, p_j\,) \;-\; g$$

The chain rule gives us that

$$\frac{\partial \mathbf{F}}{\partial p_j} \;=\; \frac{H \frac{\partial C}{\partial p_j} \;-\; C \frac{\partial H}{\partial p_j}}{H^2} \tag{5}$$

So from (4) and (5), we get that

$$\lambda = -\frac{\partial \mathbf{F}}{\partial p_j} = -\frac{H\,\frac{\partial C}{\partial p_j} - C\,\frac{\partial H}{\partial p_j}}{H^2}$$

$$\implies -\lambda\,H^2 = H\,\frac{\partial C}{\partial p_j} - C\,\frac{\partial H}{\partial p_j}$$

$$\implies -\lambda\,H^2 = H\ln(j+a) - C\Big(-g\ln p_j - g\Big)$$

$$\implies -\lambda\,H^2 = H\ln(j+a) + gC\Big(\ln p_j + 1\Big)$$

$$\implies \frac{-\lambda\,H^2 - H\ln(j+a) - 1}{gC} = \ln p_j$$

$$\implies p_j = e^{\frac{-\lambda\,H^2 - H\ln(j+a) - 1}{gC}}$$

$$\implies p_j = e^{-\frac{\lambda\,H^2}{gC}}\,e^{-\frac{H\ln(j+a)}{gC}}\,e^{-1}$$

$$\implies p_j = e^{-1-\frac{\lambda\,H^2}{gC}}\,(j+a)^{-\frac{H}{gC}}\ \therefore \tag{6}$$

To solve for p_j directly, we want solve for λ in terms of H and C. Turning our attention back to (3), we have that the minimal value of $\mathbf{F}$ is

$$\frac{C}{H} = \frac{\sum_{i=1}^{n} p_i \ln(j+a)}{-g\sum_{i=1}^{n} p_i \ln(p_j)}$$

We can substitute what we found for (6) for p_j to get

$$H = -g\sum_{i=1}^{n} p_i \ln\Big(e^{-1-\frac{\lambda\,H^2}{gC}}\,(j+a)^{-\frac{H}{gC}}\Big) \tag{7}$$

For simplicity, we will let $\alpha = \frac{H}{gC}$. Hence, we have

$$H = -g \sum_{i=1}^{n} p_i \ln\left(e^{-1 - \frac{\lambda H^2}{gC}} (j + a)^{-\alpha} \right)$$

$$= -g \sum_{i=1}^{n} p_i \left(-1 - \frac{\lambda H^2}{gC} - \alpha \ln(j + a) \right)$$

$$= -g \left(-\cancel{\sum_{i=1}^{n} p_i} - \cancel{\sum_{i=1}^{n} p_i} \frac{\lambda H^2}{gC} - \alpha \sum_{i=1}^{n} p_i \ln(j + a) \right)$$

$$= -g \left(-1 - \frac{\lambda H^2}{gC} - \alpha C \right)$$

Substituting back for α we get . . .

$$H = -g \left(-1 - \frac{\lambda H^2}{gC} - \frac{H}{g\cancel{C}} \cancel{C} \right)$$

$$H = g + \cancel{g} \frac{\lambda H^2}{\cancel{g}C} + \cancel{g} \frac{H}{\cancel{g}}$$

$$\implies 0 = g + \frac{\lambda H^2}{C}$$

$$\implies -g = \frac{\lambda H^2}{C}$$

$$\implies \lambda = \frac{-g C}{H^2} \therefore \tag{8}$$

(8) gives us a solution to p_j by substituting λ for (6). When we do this, we arrive at:

$$p_j = e^{-1 - \frac{\cancel{-gC}}{\cancel{H^2}} \cancel{H^2}} (j + a)^{-\frac{H}{gC}}$$

$$\implies p_j = e^{-\cancel{\lambda} - (-\cancel{\lambda})} (j + a)^{-\frac{H}{gC}}$$

$$\implies p_j = (j + a)^{-\alpha} \therefore \tag{9}$$

0.3 Problem 3

We want to minimize the sum of probabilities from $n = 1 \to \infty$ by finding a value for α that does this. We have that $\sum_{n=0}^{\infty} \frac{1}{(n+a)^{\alpha}}$ for $a = 1$, which is in the form of a Hurwitz-zeta function:

$$\zeta(n, a) = \sum_{n=1}^{\infty} \frac{1}{(n + a)^{\alpha}}$$

If we set $a = 1$, then our summation looks like

$$\frac{1}{1+1} + \frac{1}{2+1} + \ldots - \frac{1}{0+1}$$

$$= \frac{1}{2} + \frac{1}{3} + \ldots - 1$$

From problem 2, we know there is a value $j \in n$ such that ζ is minimized. In addition, we were given a constraint function (2), which says that the summation of each probability p_i

must be equal to 1. Hence,

$$\frac{1}{1+1} + \frac{1}{2+1} + \ldots - \frac{1}{0+1} = 1$$

$$= \frac{1}{2} + \frac{1}{3} + \ldots - 1 = 1$$

$$\implies \frac{1}{2} + \frac{1}{3} + \ldots = 2$$

$$\implies \sum_{n=1}^{\infty} \frac{1}{(n+1)^{\alpha}} = 2 \therefore$$

Hence, to find the α that minimizes the summation for $a = 1$, we want to find the α that corresponds to the value of $\zeta(n, a) = 2$. To do this, we can iterate over values of n and values of α, and identify which α corresponds to $\zeta(n, a) = 2$. The figure below shows a sweep over a range of α's:

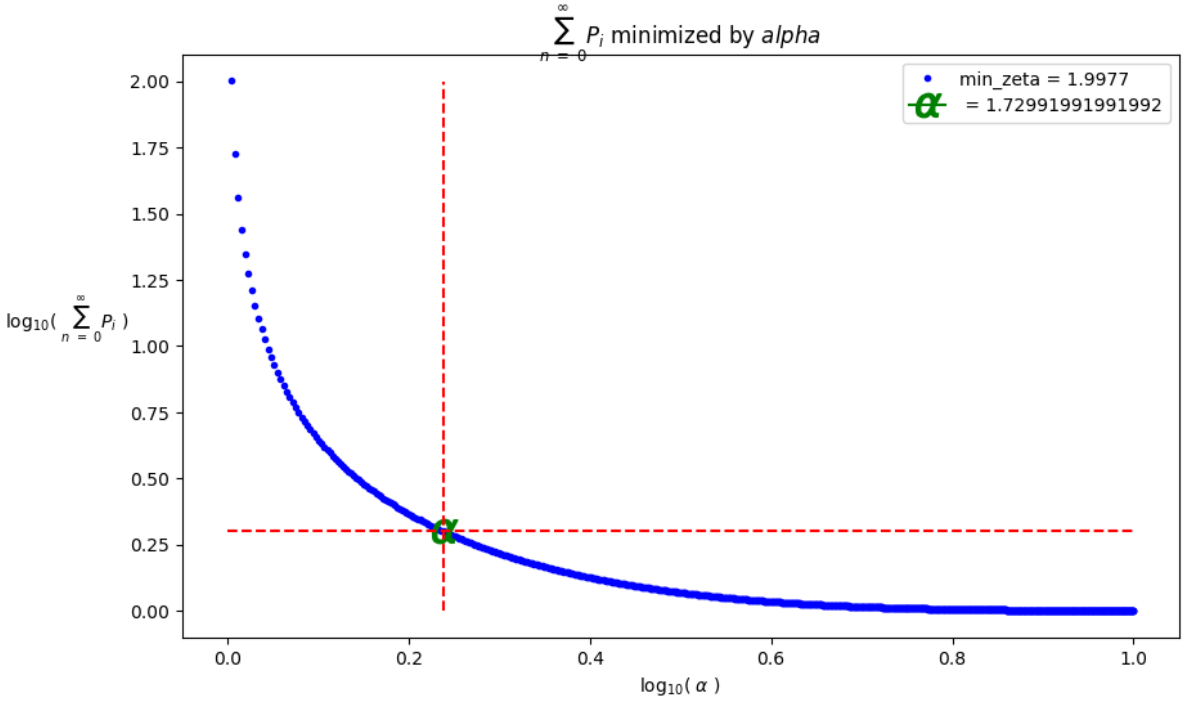

Figure 1: Computer simulation to find α, which minimizes $\zeta(n, a)$. For $a = 1$, we find that $\zeta(n, a) = 2$ is best approximated by $\alpha = 1.72992 \approx 1.73$

For a finite value of n, the value for a is given by:

$$1 = \sum_{i=1}^{n} \frac{1}{(i + a)^1}$$

$$1 = \sum_{i=1}^{n} \frac{1}{(i + a)}$$

If we integrate with respect to i over the bounds of our sum, we can solve for a directly. We have that

$$1 = \int_1^n p_i di \rightarrow 1 = \ln(i + a)\Big|_1^n$$

$$\implies 1 = \ln(n + a) - \ln(1 + a)$$

$$\implies 1 = \ln\left(\frac{n + a}{1 + a}\right)$$

$$\implies e = \frac{n + a}{1 + a}$$

$$\implies e(1 + a) - a = n$$

$$\implies a = \frac{n - e}{e - 1} \therefore \tag{3.b}$$

0.4 Problem 4

Google's raw data is for word frequency $k \geq 200$. However, from homework 2, we found a regression model intended to generate the whole CCDF. Apparently, the fit for the CCDF that we found was

$$N_{\geq k} = 3.46 \times 10^8 k^{-0.661}$$

Using this generating function, we can hypothesize what the CCDF would look like for rarer words; ones that appear with a frequency smaller than 200.

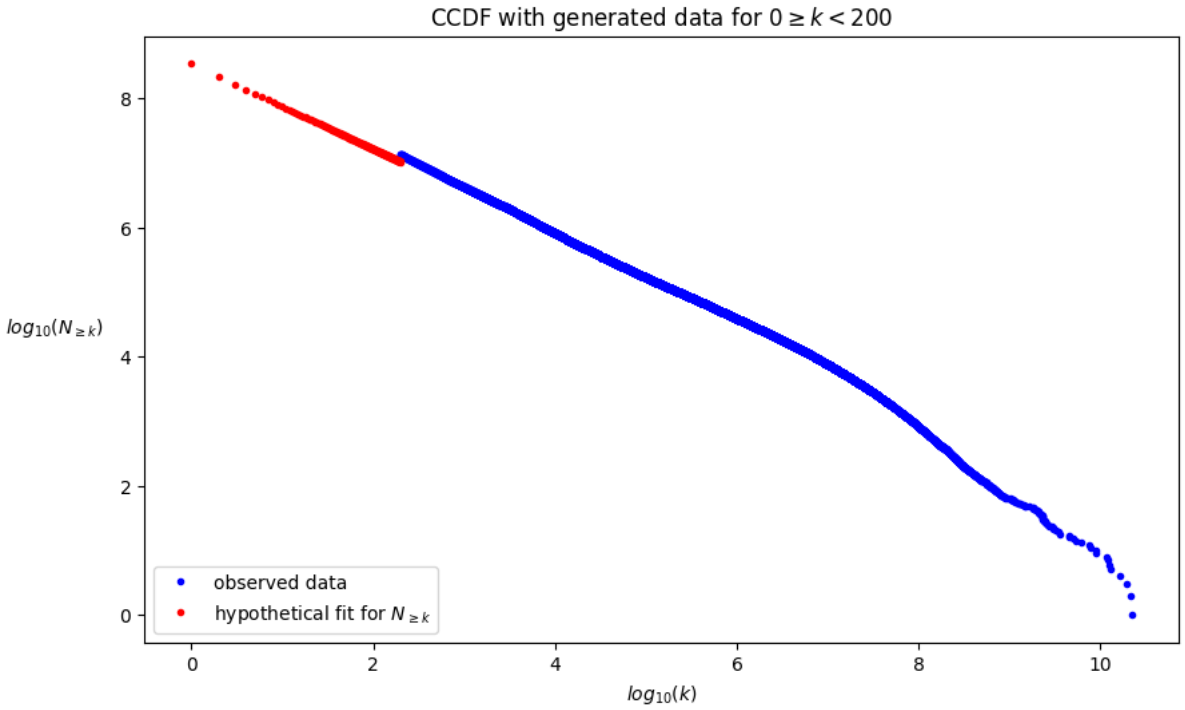

Figure 2: The hypothesized data for words that appear $k < 200$ with the observed data

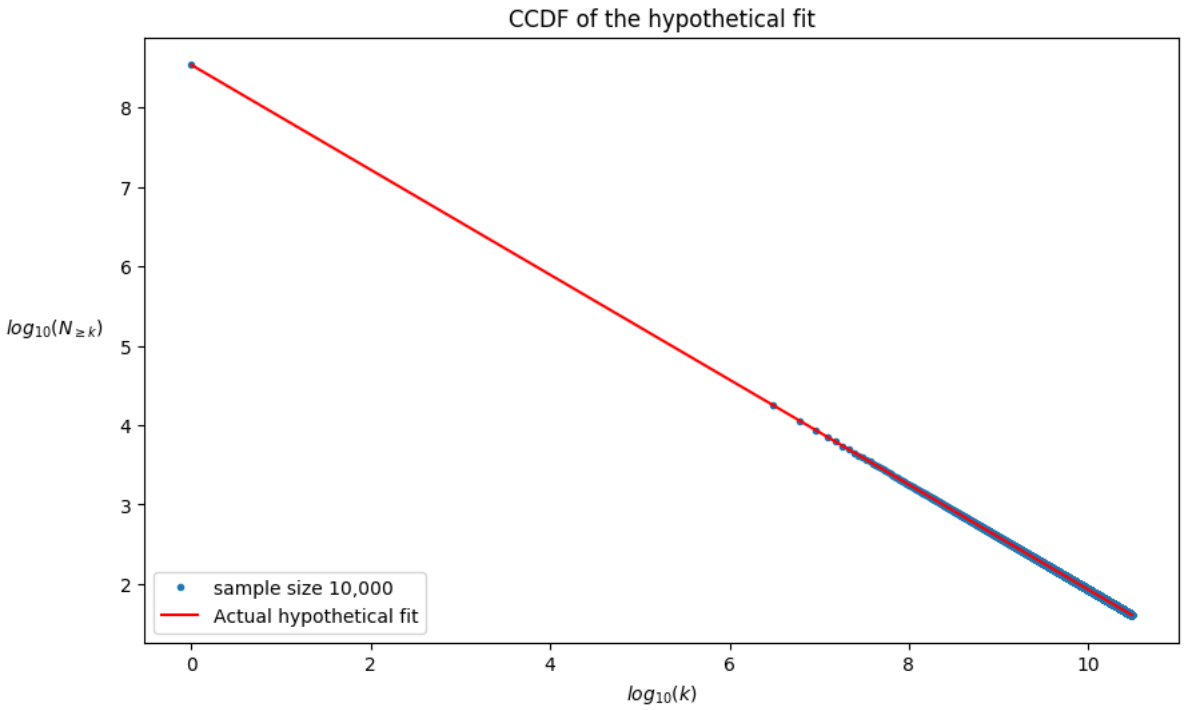

Figure 3: The hypothesized data for words that appear $k < 200$

Calculating N_k for the hypothetical data is easy now that we have the CCDF. To do it, we simply subtract the number of words of size $k-1$ that appear N_{k-1} times from the CCDF we hypothesized for groups of size $k < 200$. Since we can calculate N_k for the hypothetical data, we can now compute the fraction of words that appear k times for any group of size k. We can think of this as estimating the number of words of size k are in the corpus, divided by the total number of words in the corpus. This looks like:

$$n_k = \frac{k\, N_k}{\sum_1^{k_{max}} k\, N_k} \tag{4.c.i}$$

So, for the fraction of words in the google corpus that appear 1 time, we hypothesize that $n_1 = 0.00015167 \approx 0$.

We calculated all hypothetical N_k, so the hypothetical total number of unique words is governed by the denominator term in (4.c.i)

$$\sum_1^{k_{max}} k\, N_k$$

We find that that the hypothesized number of words in the corpus is $\sum_1^{k_{max}} k\, N_k = 838492320161$. The hypothesized number of unique words is thus the number of words of size k that appear only once (with $N_k = 1$). Thus, the fraction of all words that are unique is the number of unique words divided by the total number of words in the corpus. The total number of unique words, and fraction of unique words I calculated was

$$N_{unique} = 145424.0$$

and,

$$n_{unique} = 1.7343510071989458 \times 10^{-07} \tag{4.c.ii}$$

If we choose not to include the hypothesized data, we are only neglecting the total number of words that appear greater than or equal to 200 times, divided by the number of words in the hypothesized set. Hence we are calculating

$$\frac{\sum_{i=200}^{k_{max}} k\, N_k}{\sum_{i=1}^{k_{max}} k\, N_k} \tag{4.c.iii}$$

We estimate this to be ≈ 0.99541. This indicates that the google corpus accounts for almost all of the words. Thus, the shear size of the corpus gives us a very accurate estimate of the true distribution of words in the english language (I assume the corpus was an english one), so the heavy tail distribution is justified.