Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Fall 2011
Assignment 5

Dispersed: Monday, October 31, 2011.
Due: By start of lecture, 11:30 am, Thursday, November 10, 2011.
Some useful reminders:
Instructor: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds@uvm.edu
Office hours: 12:50 pm to 3:50 pm, Wednesday
Course website: http://www.uvm.edu/∼pdodds/teaching/courses/2011-08UVM-300

All parts are worth 3 points unless marked otherwise. Please show all your working clearly and list the names of others with whom you collaborated.

Graduate students are requested to use LaTeX (or related TeX variant).

---

1. (3 + 3 points) *Zipfarama via Optimization:*

   Complete the Mandelbrotian derivation of Zipf's law by minimizing the function

   $$\Psi(p_1, p_2, \ldots, p_n) = F(p_1, p_2, \ldots, p_n) + \lambda G(p_1, p_2, \ldots, p_n)$$

   where the 'cost over information' function is

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

   and the constraint function is

   $$G(p_1, p_2, \ldots, p_n) = \sum_{i=1}^{n} p_i - 1 = 0$$

   to find

   $$p_j = (j + a)^{-\alpha}$$

   where $\alpha = H/gC$.

   3 sub points: When finding λ, find an expression connecting λ, g, C, and H.

   Hint: one way may be to substitute the form you find for $\ln p_i$ into H's definition (but do not replace p_i).

   Note: We have now allowed the cost factor to be $(j + a)$ rather than $(j + 1)$.

2. (3 + 3)

   (a) For $n \to \infty$, use some computation tool (e.g., Matlab, an abacus, but not a clever friend who's really into computers) to determine that $\alpha \simeq 1.73$ for $a = 1$. (Recall: we expect $\alpha < 1$ for $\gamma > 2$)

   (b) For finite n, find an approximate estimate of a in terms of n that yields $\alpha = 1$.

   (Hint: use an integral approximation for the relevant sum.)

   What happens to a as $n \to \infty$?

3. (3 + 3 + 3) The next two questions continue on with the Google data set we first examined in Assignment 1.

   Estimating the rare:

   Google's raw data is for word frequency $k \geq 200$.

   From Assignment 2, we had for word frequency in the range $200 \leq k \leq 10^7$, a fit for the CCDF of

   $$N_{\geq k} \sim 3.46 \times 10^8 k^{-0.66054},$$

   ignoring errors.

   (a) Using the above fit, create a complete hypothetical N_k by expanding N_k back for $k = 1$ to $k = 199$, and plot the result in double log space.

   (b) Compute the mean and variance of the reconstructed distribution.

   (c) Estimate the fraction of 'words' that appear once, and the total number of unique words in Google's data set.

   Add question on filling in N_k for small k and then computing the mean and variance.

   Show how to compute mean and variance properly.

4. Using the CCDF and standard linear regression, measure the exponent $\gamma - 1$ as a function of the upper limit of the scaling window, with a fixed lower limit of $k_{\min} = 200$.

   Please plot γ as a function of $k_{\max}$, including 95% confidence intervals.

   Note that the break in scaling should mess things up but we're interested here in how stable the estimate of γ is up until the break point.

   Comment on the stability of γ over variable window sizes.

   Pro Tip: your upper values should be distributed evenly in log space.