Principles of Complex Systems, CSYS/MATH 300
University of Vermont, Spring 2013
Assignment 5 • code name: Stayin' Alive (⊞)

Dispersed: Friday, February 22, 2013.
Due: By start of lecture, 11:30 am, Thursday, February 28, 2013.
Some useful reminders:
Instructor: Peter Dodds
Office: Farrell Hall, second floor, Trinity Campus
E-mail: peter.dodds@uvm.edu
Office hours: 1:00 pm to 4:00 pm, Wednesday
Course website: http://www.uvm.edu/∼pdodds/teaching/courses/2013-01UVM-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. The discrete version of HOT theory:

   From lectures, we had the follwoing.

   Cost: Expected size of 'fire' in a d-dimensional lattice:

   $$C_{\text{fire}} \propto \sum_{i=1}^{N_{\text{sites}}} (p_i a_i) a_i = \sum_{i=1}^{N_{\text{sites}}} p_i a_i^2,$$

   where a_i = area of ith site's region, and p_i = avg. prob. of fire at site in ith site's region.

   From lectures, the constraint for building and maintaining $(d-1)$-dimensional firewalls in d-dimensions is

   $$C_{\text{firewalls}} \propto \sum_{i=1}^{N_{\text{sites}}} a_i^{(d-1)/d} a_i^{-1},$$

   where we are assuming isometry.

   Using Lagrange Multipliers, safety goggles, rubber gloves, a pair of tongs, and a maniacal laugh, determine that:

   $$p_i \propto a_i^{-\gamma} = a_i^{-(2+1/d)}.$$

2. (3 + 3 + 3) *Highly Optimized Tolerance:*

   This question is based on Carlson and Doyle's 1999 paper "Highly optimized tolerance: A mechanism for power laws in design systems" [1]. In class, we made our way through a discrete version of a toy HOT model of forest fires. This paper revolves around the equivalent continuous model's derivation.

   Our interest is in Table I on p. 1415:

$p(x)$	$p_{\text{cum}}(x)$	$P_{\text{cum}}(A)$
$x^{-(q+1)}$	x^{-q}	$A^{-\gamma(1-1/q)}$
e^{-x}	e^{-x}	$A^{-\gamma}$
e^{-x^2}	$x^{-1}e^{-x^2}$	$A^{-\gamma}[\log(A)]^{-1/2}$

   and Equation 8 on the same page:

$$P_{\geq}(A) = \int_{p^{-1}(A^{-\gamma})}^{\infty} p(\mathbf{x})\mathrm{d}\mathbf{x} = p_{\geq}\left(p^{-1}\left(A^{-\gamma}\right)\right),$$

   where $\gamma = \alpha + 1/\beta$ and we'll write $P_{\geq}$ for P_{cum}.

   Please note that $P_{\geq}(A)$ for $x^{-(q+1)}$ is not correct. Find the right one!

   Here, $A(\mathbf{x})$ is the area connected to the point $\mathbf{x}$ (think connected patch of trees for forest fires). The cost of a 'failure' (e.g., lightning) beginning at $\mathbf{x}$ scales as $A(\mathbf{x})^{\alpha}$ which in turn occurs with probability $p(\mathbf{x})$. The function p^{-1} is the inverse function of p.

   Resources associated with point $\mathbf{x}$ are denoted as $R(\mathbf{x})$ and area is assumed to scale with resource as $A(\mathbf{x}) \sim R^{-\beta}(\mathbf{x})$.

   Finally, $p_{\geq}$ is the complementary cumulative distribution function for p.

   As per the table, determine $p_{\geq}(x)$ and $P_{\geq}(A)$ for the following (3 pts each):

   (a) $p(x) = cx^{-(q+1)}$,

   (b) $p(x) = ce^{-x}$, and

   (c) $p(x) = ce^{-x^2}$.

   Note that you should incorporate a constant of proportionality c, which is not shown in the paper.

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

   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 limit values should be distributed evenly in log space.

4. (3 + 3 + 3)

   Estimating the rare:

   Google's raw data is for word frequency $k \geq 200$ so let's deal with that issue now.

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

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

   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 (meaning log-log space).

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

   (c) Estimate:

      i. the hypothetical fraction of words that appear once out of all words (think of words as organisms here),

      ii. the hypothetical total number and fraction of unique words in Google's data set (think at the species level now),

      iii. and what fraction of total words are left out of the Google data set by providing only those with counts $k \geq 200$ (back to words as organisms).

References

[1] J. M. Carlson and J. Doyle. Highly optimized tolerance: A mechanism for power laws in designed systems. Phys. Rev. E, 60(2):1412–1427, 1999.