1. (3+3 points) Simon’s model I:

For Herbert Simon’s model of what we’ve called Random Competitive Replication, we found in class that the normalized number of groups in the long time limit, \( n_k \), satisfies the following difference equation:

\[
\frac{n_k}{n_{k-1}} = \frac{(k-1)(1-\rho)}{1 + (1-\rho)k}
\]  

(1)

where \( k \geq 2 \). The model parameter \( \rho \) is the probability that a newly arriving node forms a group of its own (or is a novel word, starts a new city, has a unique flavor, etc.). For \( k = 1 \), we have instead

\[
n_1 = \rho - (1-\rho)n_1
\]  

(2)

which directly gives us \( n_1 \) in terms of \( \rho \).

(a) Derive the exact solution for \( n_k \) in terms of gamma functions and ultimately the beta function.

(b) From this exact form, determine the large \( k \) behavior for \( n_k \) (\( \sim k^{-\gamma} \)) and identify the exponent \( \gamma \) in terms of \( \rho \).
Note: Simon’s own calculation is slightly awry. The end result is good however.

Hint—Setting up Simon’s model:

The hint’s output including the bits not in the video:
2. (3+3 points) Simon’s model II:

(a) A missing piece from the lectures: Obtain $\gamma$ in terms of $\rho$ by expanding Eq. 1 in terms of $1/k$. In the end, you will need to express $n_k/n_{k-1}$ as $(1 - 1/k)^\rho$; from here, you will be able to identify $\gamma$. Taylor expansions and Procrustean truncations will be in order. This (dirty) method avoids finding the exact form for $n_k$.

(b) What happens to $\gamma$ in the limits $\rho \to 0$ and $\rho \to 1$? Explain in a sentence or two what’s going on in these cases and how the specific limiting value of $\gamma$ makes sense.

3. (6 + 3 + 3 points)

In Simon’s original model, the expected total number of distinct groups at time $t$ is $\rho t$. Recall that each group is made up of elements of a particular flavor.

In class, we derived the fraction of groups containing only 1 element, finding

$$n^{(g)}_1 = \frac{N_1(t)}{\rho t} = \frac{1}{2 - \rho}$$

(a) (3 + 3 points)

Find the form of $n^{(g)}_2$ and $n^{(g)}_3$, the fraction of groups that are of size 2 and size 3.

(b) Using data for James Joyce’s Ulysses (see below), first show that Simon’s estimate for the innovation rate $\rho_{\text{est}} \simeq 0.115$ is reasonably accurate for the version of the text’s word counts given below.

Hint: You should find a slightly higher number than Simon did.

Hint: Do not compute $\rho_{\text{est}}$ from an estimate of $\gamma$.

(c) Now compare the theoretical estimates for $n^{(g)}_1$, $n^{(g)}_2$, and $n^{(g)}_3$, with empirical values you obtain for Ulysses.

The data (links are clickable):

- Matlab file (sortedcounts = word frequency $f$ in descending order, sortedwords = ranked words):
  http://www.uvm.edu/~pdodds/teaching/courses/2013-08UVM-300/docs/ulysses.mat

- Colon-separated text file (first column = word, second column = word frequency $f$):
  http://www.uvm.edu/~pdodds/teaching/courses/2013-08UVM-300/docs/ulysses.txt
4. (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

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

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 points: When finding \( \lambda \), find an expression connecting \( \lambda, g, C, \) 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 \( (j + 1) \).
5. (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 $n \to \infty$?