PRINCIPLES OF COMPLEX SYSTEMS

HW03 WRITE-UP

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

0.1 Problem 1

The complementary cumulative distribution function (CCDF) is, as its name suggests, the compliment of the cumulative distribution function. The CDF describes the probability of a random variable being less than or equal to a particular sample value. If we consider k, the number of times a word appears in a text corpus, and N_k, the frequency of words that appear K times, and we observe measures for both over a large text corpus (say a set of google articles containing close to 14 million words), then the cdf would describe, for any k, the probability that we expect to see a word that occurs less frequently, or with equal frequency, to that word. The CCDF would ask the complimentary question: *what is the probability that a word that occurs with a certain frequency occurs with probability* $X > k$*?* for some random variable X and count k. Thus, we can take a cummulative sum over the frequencies of words that occur k times; N_k to find the CDF, and then subtract those values from 1 to obtain a measure of the CCDF.

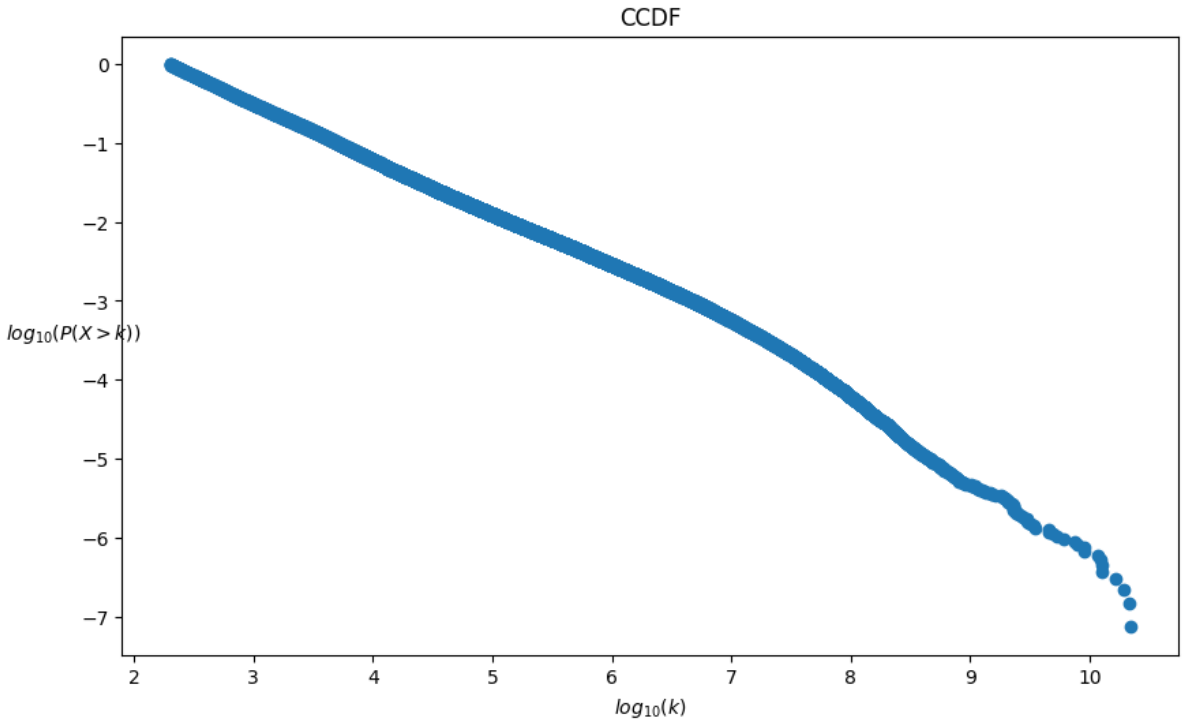

Figure 1: Complimentary cumulative distribution showing that it's more probable to see rare words that show up fewer times in a corpus than it is to see common words that occur more frequently

Here, we plot the CCDF of word counts in the google corpus on a log-log plot to

highlight the fact that counts of rare words in a corpus are less unique than that of common words. The x-axis is the log of the values for k, and the y-axis is a log of the probabilities of seeing a word that occurs k times based on the corresponding value for N_k. The linear nature of the relationship in log-log space indicates a power-law distribution over k and N_k. This is also made evident by the fact that the probability mass function describes a power-law distribution between the two quantities:

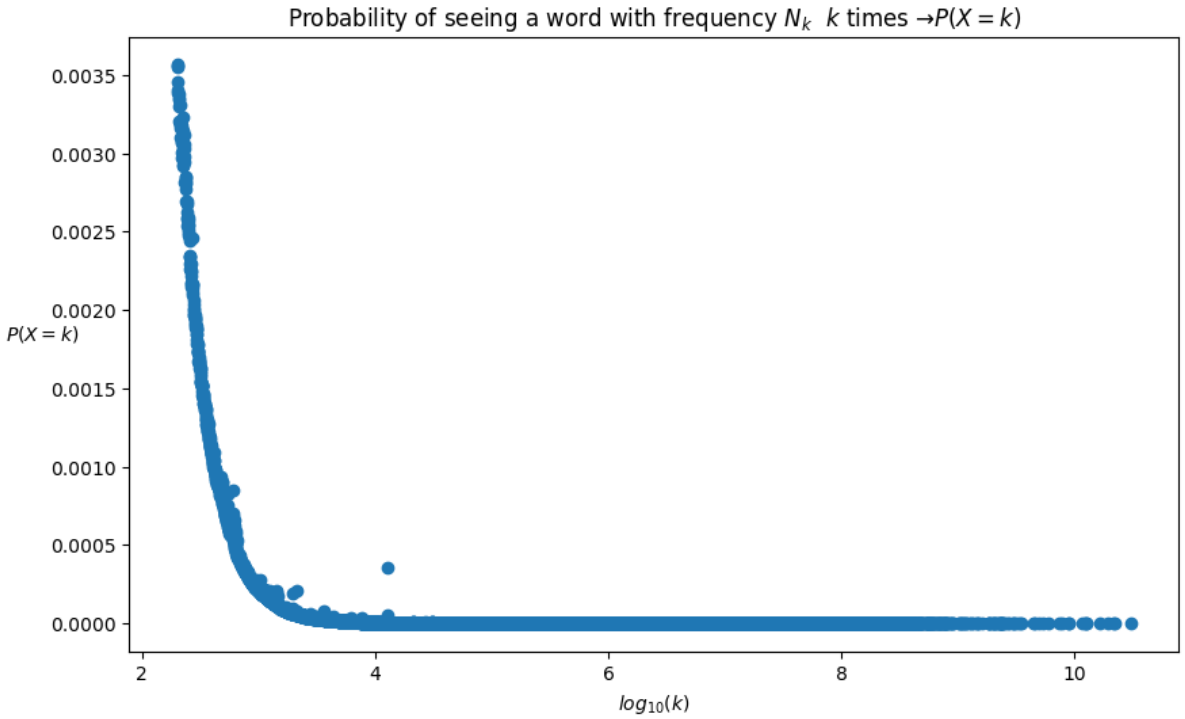

Figure 2: The probability of a word occuring k times

The next section shows an estimate of the best fit regression over CCDF curve. Two regimes are used to describe the relationship between word frequency k and the frequency of a word that occurs k times, N_k, as the data cannot be described perfectly with one regression.

0.2 Problem 2

The intervals for which the two regimes are defined were approximated by sight (bad practice, but fair for the purposes of this assignment) to be $k \in [10^2, 10^4.3)$ and $k \in (10^7, 10^{11}]$ respectively. The values of γ are expressed in the figure below:

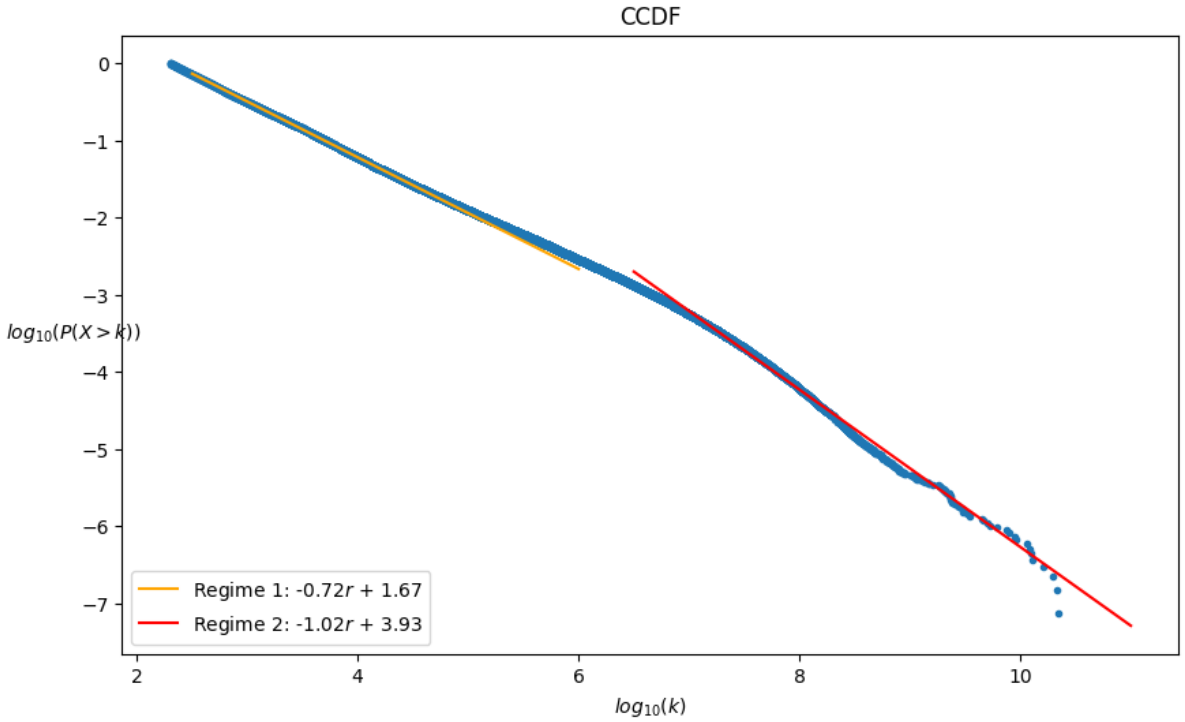

Figure 3: A plot showing the two linear regimes that best describe the data over two regions of the x-axis. The slopes of the two functions indicate a best estimate of the value for the exponent γ for the intervals $k \in [10^2, 10^4.3)$ and $k \in (10^7, 10^{11}]$ respectively.

The large size of the data set (13,588,391 word counts) implies that the data is representative of a larger population. Thus, we can approximate the 95% confidence interval of each regime by using the observed standard error from the regression. For the first regime, we are 95% confident that the mean value of γ is between -0.72328 and -0.72292. For the second, we are 95% confident that the mean value of γ is between -1.02118 and -1.01782.

0.3 Problem 3

Zipf's law says that there is an inverse relationship between the frequency of any word in the corpus and its rank amongst all frequencies. By sorting the data by its ranked frequency and plotting the rank r of each word against the number of times it occurs, k, then we get a curve as described below:

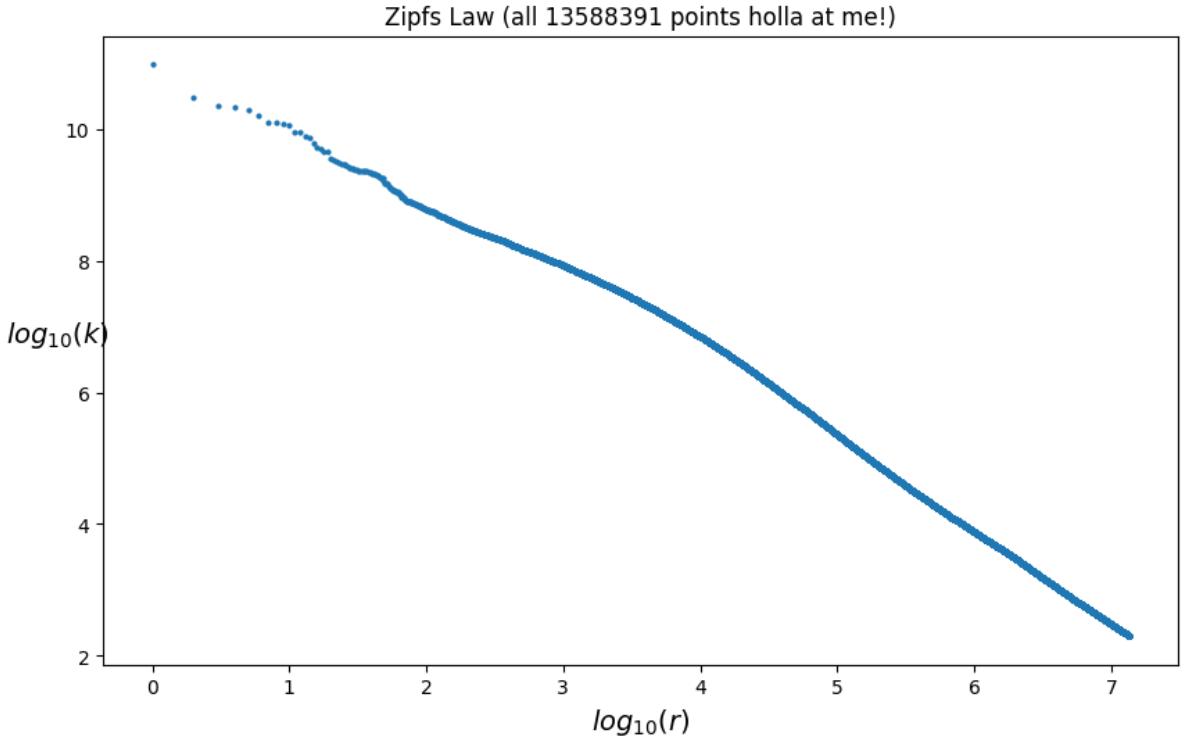

Figure 4: A plot of the rank of each word frequency against word frequency k

0.4 Problem 4

In the same manner as before, we can describe the curve of the "Zipfian" distribution function with two regimes. We will take our two linear regressions over the same interval for k.

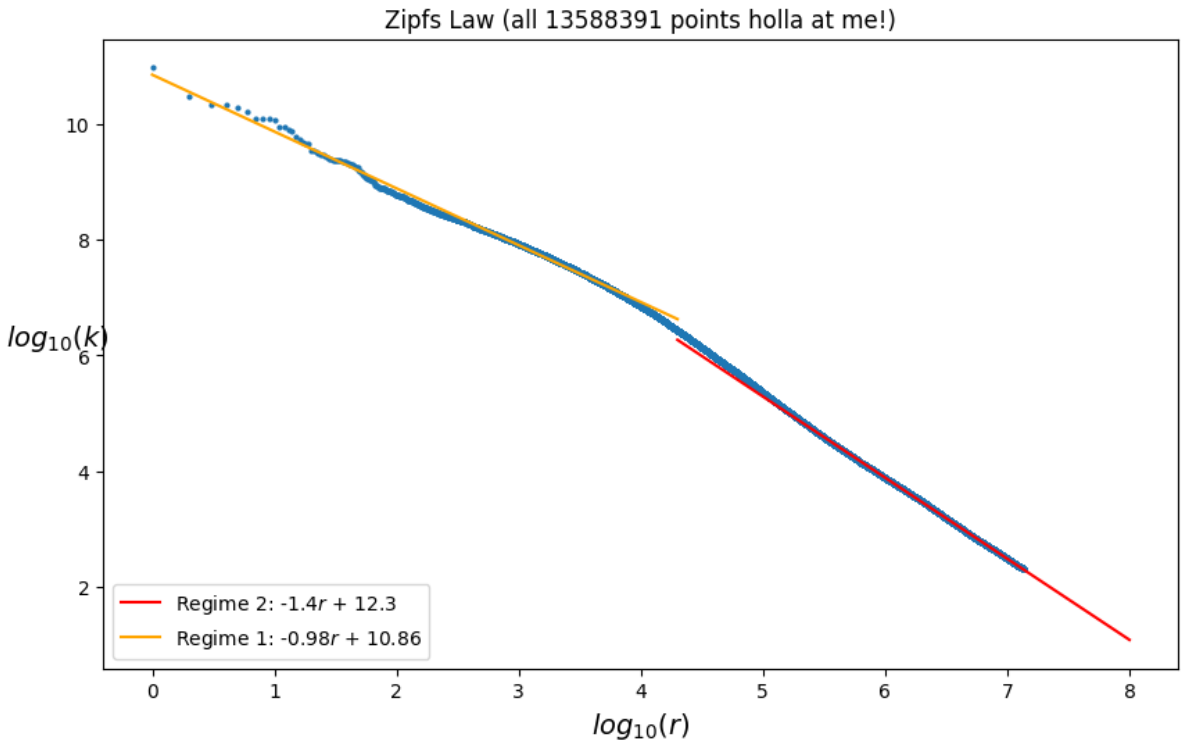

Figure 5: Two linear regressions over the ranges $k \in [10^2, 10^{4.3})$ and $k \in (10^7, 10^{11}]$ respectively

For the first regime, we are 95% confident that the mean value of α lies between -1.40286 and -1.40284. For the second regime, we are 95% confident that the mean value of α lies between -0.9849 and -0.98176.

0.5 Problem 5

In class, we reasoned that the Zipfian distribution function is the inverse of the CCDF. Hence, we expect that our estimates of γ are the inverse of our estimates of α for each regime. Thus, we want to check the relationship:

$$\alpha = \frac{1}{\gamma}$$

holds among our estimated values for the slope of each regime.

We find that for the first regime, our observed values for $\alpha = \frac{1}{\gamma_1} = -0.980869$, and $\gamma = \frac{1}{\alpha_1} = -0.712834$. For the second regime, our observed values for $\alpha = \frac{1}{\gamma_2} = -1.3829$, and

$\gamma = \frac{1}{\alpha_2} = -1.01695$. Hence, the estimates of γ and α for each linear regression conform well to the observed CCDF and Zipfian function.

0.6 Problem 6 - Random Walks

For large $t = 2n$, we will show

$$\mathbf{N}(0, 2k, 2n) = \binom{2n}{n+k} = \binom{2n}{n-k}$$

leads to a Gaussian distribution. We will let $\alpha = n - k$, and $\beta = n + k$. Then,

$$\mathbf{N}(0, 2k, 2n) = \frac{(2n)!}{(\alpha)!\,(\beta)!}$$

Using *Sterling's Approximation*, we can substitute $n!$ for $\sqrt{2\pi n}\,(\frac{n}{e})^n$ to get

$$\frac{\sqrt{4\pi n}\,(\frac{2n}{e})^{2n}}{\sqrt{2\pi\alpha}\,(\frac{\alpha}{e})^{\alpha}\,\sqrt{2\pi\beta}\,(\frac{\beta}{e})^{\beta}}$$

and we can simplify accordingly to obtain

$$= \frac{2\sqrt{\pi n}\,(\frac{2n}{e})^{2n}}{2\pi\sqrt{\alpha}\,(\frac{\alpha}{e})^{\alpha}\,\sqrt{\beta}\,(\frac{\beta}{e})^{\beta}}$$

$$= \frac{\sqrt{n}\,2^{2n}\,n^{2n}\,e^{-2n}}{\sqrt{\pi}\,\sqrt{\alpha}\,\alpha^{\alpha}\,e^{-\alpha}\,\sqrt{\beta}\,\beta^{\beta}\,e^{-\beta}}$$

$$= \frac{\sqrt{n}\,2^{2n}\,n^{2n}\,e^{-2n}}{\sqrt{\pi}\,\sqrt{\alpha}\,\alpha^{\alpha}\,e^{-\alpha}\,\sqrt{\beta}\,\beta^{\beta}\,e^{-\beta}}$$

$$= \frac{\sqrt{n}\,2^{2n}\,n^{2n}\,e^{-2n}}{\sqrt{\pi}\,\sqrt{\alpha}\,\alpha^{\alpha}\,\sqrt{\beta}\,\beta^{\beta}\,e^{-\alpha-\beta}}$$

We can then substitute back $n - k$ and $n + k$ for α and β respectively, and then simplify to obtain

$$= \frac{n^{\frac{1}{2}} 2^{2n} n^{2n} e^{-2n}}{\sqrt{\pi}(n-k)^{\frac{1}{2}}(n-k)^{n-k}(n+k)^{\frac{1}{2}}(n+k)^{n+k} e^{-(n-k)-(n+k)}}$$

$$= \frac{n^{\frac{1}{2}} 2^{2n} n^{2n} e^{-2n}}{\sqrt{\pi}(n-k)^{n+\frac{1}{2}-k}(n+k)^{n+\frac{1}{2}+k} e^{-2n}}$$

$$= \frac{n^{\frac{1}{2}} 2^{2n} n^{2n}}{\sqrt{\pi}(n-k)^{n+\frac{1}{2}-k}(n+k)^{n+\frac{1}{2}+k}}$$

$$= \frac{n^{\frac{1}{2}} 2^{2n} n^{2n}}{1} \frac{1}{\sqrt{\pi}(n-k)^{n+\frac{1}{2}-k}(n+k)^{n+\frac{1}{2}+k}}$$

$$= \frac{n^{\frac{1}{2}} 2^{2n} n^{2n}}{1} \frac{1}{\sqrt{\pi}\, n^{n+\frac{1}{2}-k}(1-\frac{k}{n})^{n+\frac{1}{2}}(1-\frac{k}{n})^{-k} n^{n+\frac{1}{2}+k}(1+\frac{k}{n})^{n+\frac{1}{2}+k}}$$

$$= \frac{n^{\frac{1}{2}} 2^{2n} n^{2n}}{1} \frac{1}{\sqrt{\pi}(\frac{n^{n+\frac{1}{2}}}{n^{-k}}\, n^{n+\frac{1}{2}} n^{k})(1-\frac{k}{n})^{n+\frac{1}{2}}(1-\frac{k}{n})^{-k}(1+\frac{k}{n})^{n+\frac{1}{2}+k}}$$

$$= \frac{n^{\frac{1}{2}} 2^{2n} n^{2n}}{\sqrt{\pi}\, n^{2n+1}(1-\frac{k^2}{n^2})^{n+\frac{1}{2}}(1+\frac{k}{n})^{k}(1-\frac{k}{n})^{-k}}$$

$$= \frac{2^{2n}}{\sqrt{\pi n}(1-\frac{k^2}{n^2})^{n+\frac{1}{2}}(1+\frac{k}{n})^{k}(1-\frac{k}{n})^{-k}}$$

We want to utilize the Taylor expansion of the natural logarithm of the expression, which says that $\ln(1+z) \to z$ as $z \to 0$. Under our initial assumption that $k \ll n$, $\frac{k}{n} \to 0$, so

we may use the taylor

$$= \ln\left(\frac{2^{2n}}{\sqrt{\pi\, n}\,(1-\frac{k^2}{n^2})^{n+\frac{1}{2}}\,(1+\frac{k}{n})^{k}\,(1-\frac{k}{n})^{-k}}\right)$$

$$\ln(2) - \ln(\sqrt{\pi\, n}) - \ln\left((1-\frac{k^2}{n^2})^{n+\frac{1}{2}}\right) - \ln\left((1+\frac{k}{n})^{k}\right) - \ln\left((1-\frac{k}{n})^{-k}\right)$$

$$= \ln(2^{2n}) - \ln(\sqrt{\pi\, n}) - \ln\left((1-\frac{k^2}{n^2})^{n+\frac{1}{2}}\right) - k\ln(1+\frac{k}{n}) + k\ln(1-\frac{k}{n})$$

$$= \ln(2^{2n}) - \ln(\sqrt{\pi\, n})\,\frac{k^2}{n^2}\,(n+\frac{1}{2}) - k\,\frac{k}{n} - k\,\frac{k}{n}$$

$$= \ln(2^{2n}) - \ln(\sqrt{\pi\, n}) + \frac{k^2}{n} + \frac{k^2}{2n^2} - k\,\frac{k}{n} - k\,\frac{k}{n}$$

$$= \ln(2^{2n}) - \ln(\sqrt{\pi\, n}) - \frac{k^2}{n} + \frac{k^2}{2n^2}$$

Raising the base of the natural logarithm, e, to the entire expression, we get

$$= \frac{2^{2n}\, e^{-\left(\frac{k^2}{2n^2} + \frac{2n\, k^2}{2n^2}\right)}}{\sqrt{\pi\, n}}$$

$$= \frac{2^{2n}\, e^{-\frac{k^2 + 2n\, k^2}{2n^2}}}{\sqrt{\pi\, n}}$$

$$= \frac{2^{2n}\, e^{-\frac{k^2\,(2n+1)}{2n^2}}}{\sqrt{\pi\, n}} \approx \frac{2^{2n}\, e^{-\frac{k^2\,(2n)}{2n^2}}}{\sqrt{\pi\, n}}, \text{ for } k \ll n$$

$$= \frac{2^{2n}\, e^{-\frac{2\,k^2}{2n} \times \frac{2}{2}}}{\sqrt{\pi\, n}} = \frac{2^{2n}\, e^{-\frac{4\,k^2}{2(2n)}}}{\sqrt{\pi\, n}} = \frac{2^{2n}\, e^{-\frac{4\,k^2}{2(2n)}}}{\sqrt{\pi\, n}}$$

$$= \frac{2^{2n}\, e^{-\frac{(2k)^2}{2(2n)}}}{\sqrt{\pi\, n}} = \frac{2 \times 2^{2n}\, e^{-\frac{(2k)^2}{2(2n)}}}{2 \times \sqrt{\pi\, n}}$$

$$= \frac{2^{2n+1}\, e^{-\frac{(2k)^2}{2(2n)}}}{\sqrt{4\,\pi\, n}} = \frac{2^{t+1}\, e^{-\frac{x^2}{2t}}}{\sqrt{4\,\pi\, \frac{t}{2}}}$$

$$= \frac{2^{t+1}}{\sqrt{2\,\pi\, t}}\, e^{-\frac{x^2}{2t}}$$

Notice that the number of discrete random walks is equal to $2^{2n+1} = 2^{t+1}$. So, we may divide by the total number of random walks and thus confirm that

N $(0, 2k, 2n)$ leads to a Gaussian distribution for large $t = 2n$, given by:

Pr $(x_t \equiv x) \simeq \frac{1}{\sqrt{2\,\pi\, t}}\, e^{\frac{-x^2}{2t}}$ $\therefore$

0.7 Problem 7

If a $1-d$ random walk starts at position $x = i$, and ends at position $x = j$, then the displacement after t time steps is $i - j$.

Let, $P =$ the number of positive steps the walker takes after t time steps,
and let $N =$ the number of negaive steps the walker takes after t time steps

We know that

$$P + N = t \text{ and, } P - N = i - j \qquad \text{(by definition)}$$

So,

$$N = t - P,$$

$$P = N + j - i$$

$$= t - P + j - i$$

$$\implies 2P = t + j - i$$

$$\implies P = \frac{t + j - i}{2}$$

$$\implies N = \frac{t + i - j}{2}$$

$$\implies \mathbf{N_P}(i, j, t) = \begin{pmatrix} t \\ P \end{pmatrix}, \ \mathbf{N_N}(i, j, t) = \begin{pmatrix} t \\ N \end{pmatrix}$$

We will now show that $\mathbf{N_P}(\, i,\, j,\, t\,) \;=\; \binom{t}{P} \;=\; \mathbf{N_N}(\, i,\, j,\, t\,) \;=\; \binom{t}{N}$

$$\binom{t}{p} \;=\; \frac{t!}{P!\,(\,t\,-\,P\,)!} \;=\; \frac{t!}{P!\,N!} \qquad \text{(by definition)}$$

$$\binom{t}{N} \;=\; \frac{t!}{N!\,(\,t\,-\,N\,)!} \;=\; \frac{t!}{N!\,P!} \qquad \text{(by definition)}$$

$$\Longrightarrow \binom{t}{p} \;=\; \binom{t}{N} \;\therefore$$

Hence, the number of distinct $1-d$ random walks;

$$\mathbf{N}(\, i,\, j,\, t\,) \;=\; \binom{t}{\frac{(\,t\,+\,j\,-\,i\,)}{2}}$$

0.8 Problem 8

$$N(i,\, j,\, t) = \binom{t}{\frac{(t+j-i)}{2}}$$

$$N_{fr}(2n) = N(1, 1, 2n-2) - N(-1, 1, 2n-2)$$

$$= \binom{2n-2}{n-1} - \binom{2n-2}{n-2}$$

$$= \frac{(2n-2)!}{(n-1)!\,(n-1)!} - \frac{(2n-2)!}{(n-2)!\,(n)!}$$

$$= \frac{(2n-2)!}{\frac{(n)!^2}{n^2}} - \frac{(2n-2)!}{\frac{(n)!^2}{n^2-n}}$$

$$= \frac{(2n-2)!\,n^2}{(n)!^2} - \frac{(2n-2)!\,(n^2-n)}{(n)!^2}$$

$$= \frac{\frac{(2n)!}{(2n)(2n-1)}\,n^2}{(n)!^2} - \frac{\frac{(2n)!}{(2n)(2n-1)}\,(n^2-n)}{(n)!^2}$$

$$= \frac{(2n)!}{(4n-2)\,(n)!^2}$$

Using Sterling's approximation, we can substitue $\sqrt{2\,\pi\,n}\,(\,\frac{n}{e}\,)^n$ for $n!$ to obtain

$$=\frac{\sqrt{2\,\pi\,2n}\,(\,\frac{2n}{e}\,)^{2n}}{(\,4n-\,2\,)\,\sqrt{2\,\pi\,n}^{2}\,(\,\frac{n}{e}\,)^{2n}}$$

$$=\frac{\sqrt{2\,\pi\,2n}\,(\,2n\,)^{2n}}{(\,4n-\,2\,)\,2\pi n\,n^{2n}}$$

$$=\frac{\sqrt{4\,\pi\,n}\,(\,2n\,)^{2n}}{(\,2n-\,1\,)\,4\pi n\,n^{2n}}$$

$$=\frac{(\,2n\,)^{2n}}{(\,2n-\,1\,)\,\sqrt{4\,\pi\,n}\,n^{2n}}$$

$$=\frac{4^n}{(\,2n-\,1\,)\,\sqrt{4\,\pi\,n}}$$

$$=\frac{4^n}{(\,4n-\,2\,)\,\sqrt{\pi\,n}}, \text{Assume that } n \text{ is large;}$$

$$\Longrightarrow \frac{4^n}{(\,4n-\,2\,)\,\sqrt{\pi\,n}} \;\simeq\; \frac{4^n}{4n\,\sqrt{n}}$$

$$=\frac{4^n}{4n^{\frac{3}{2}}}$$

If we take $\lim_{n\to\infty}$ of the expression, we find that the dominant, lead, term is 4^n $\therefore$