Avalanche Dynamics in a Universal Amplitude Distribution for MALDI Peaks

Jennifer E. Dorand, College of William and Mary

Abstract

Avalanche models occur in a variety of physical processes, including the electron avalanches in "avalanche" detectors, which are commonly used to detect photons, electrons, or ions. We are interested in modeling this avalanche process to understand the statistics of amplitudes in MALDI-TOF spectra. MALDI-TOF mass spectrometry stands for matrix assisted-laser desorption time-of-flight mass spectrometry. During this process, proteins are accelerated by an electrostatic potential and identified when they cause an electron avalanche in the detector. The number of electrons at the end of the avalanche is recorded and a MALDI spectra is obtained for the sample. We propose to model this electron avalanche, finding an equation that expresses the probability distribution that for m initial ions at the start of the avalanche, there are n electrons at the end. By understanding the avalanche dynamics of the detector, a better estimate of peak height, and therefore a more accurate protein distribution can be obtained from the MALDI spectra, even for small data sets.

1 Introduction and Motivation

Avalanche models occur in a variety of physical processes, from grains of rice being added to a pile to avalanches in a MALDI-TOF detector. There are a variety of different equations that fit various avalanche processes. The process I will be modeling is from MALDI-TOF (matrix-assisted laser desorption ionization-time of flight) spectra. This spectrum is obtained from blood and tissue samples collected and prepared by research associates at Eastern Virginia Medical School. The samples are then analyzed using a MALDI-TOF instrument at EVMS.

Mass spectrometry is used to determine the chemical composition, inferred from the charge-to-mass ratio, of a sample. It has replaced chromatography, electrophoresis, and DNA sequencing because these methods work well for low thoroughput diagnostics only, or those where only small amounts of information can be gained from a sample. They are also costly and insensitive[7]. However, mass spectrometry is extremely rapid, can analyze data in a very short time period, and has sensitivities in the femtomolar, or 10^{-15} molar, range. It has become especially promising in cancer diagnostics[6]. Currently, the most reliable, sensitive, and widely available tests to detect cancer are protein-ligand assays, used to detect proteins in a sample, which require a single analyte, or one particular protein that is being examined, that has been authenticated in addition to a well-characterized, high-affinity antibody to detect the analyte. Thus, such methods require a great deal of time and effort for only one sample. Therefore, mass spectrometry has become the preferred method in proteomics research, including cancer detection[7].

MALDI-TOF is one of the most widely used mass spectrometry methods. The advantages of MALDI is that it is easy to use, is very sensitive, can be used for heavy particles (>20,000 kDa), is able to determine sample components' molecular mass values with extremely high accuracy and has a high tolerance to contaminants[6]. A MALDI matrix is first deposited on the sample, in this case, blood or tissue, and allowed to crystallize. The matrix is of low molecular weight to allow easy vaporization but large enough not to evaporate during sample preparation. It is acidic and acts as a proton source to aid in ionization of the sample. When mixed with the analyte, the matrix allows strong absorption in the UV light range. It contains an organic solvent to allow the hydrophobic molecules to dissolve but is an aqueous solution to allow hydrophilic molecules in the sample to dissolve as well. The mixture of matrix and analyte is then put onto a MALDI plate, a pre-made specifically designed surface for use with MALDI spectrometry, and the solvents vaporize, allowing the matrix to recrystallize along with the analyte molecules. A laser is then fired at the matrix, heating the matrix, and causing it to desorb. The matrix transforms the laser energy into excitation energy for the sample. The matrix allows efficient energy transfer and spares the analyte molecules from excessive direct energy that could cause them to decompose into smaller particles. Through collision processes during ionization, the proteins embedded in the sample become charged. There is a desorption, in which atoms are ejected from a solid target, of analyte and matrix ions from the surface of the mixture and a resulting plume of ions[2].

The resulting ions are then accelerated by an electrostatic potential. The velocity of the particle accelerated by the electric field of the analyzer depends on its charge to mass ratio (m/z).

$$E = qV = \frac{1}{2}mv^2, \tag{1}$$

where E is the energy of the particle, q is the charge of the particle, V is the electrostatic potential, m is the mass of the particle, and v is the velocity of the particle after it is accelerated. The electromagnetic energy of the particle is equal to the kinetic energy, and used to derive the m/z ratio by manipulating the equation into the form

$$\frac{q}{m} = \frac{v^2}{2V}. \tag{2}$$

The m/z ratio is derived from the time it takes the particle to reach the detector. The particles are detected when they cause an electron avalanche in the detector. The detector is a microchannel plate detector, used because Time of Flight Mass Spec requires high temporal resolution. The ions from the MALDI instrument first strike the converter, which converts these ions into electrons that then interact with two microchannel plates in a chevron configuration, which means they are in series with their channels directed in opposite directions[1].

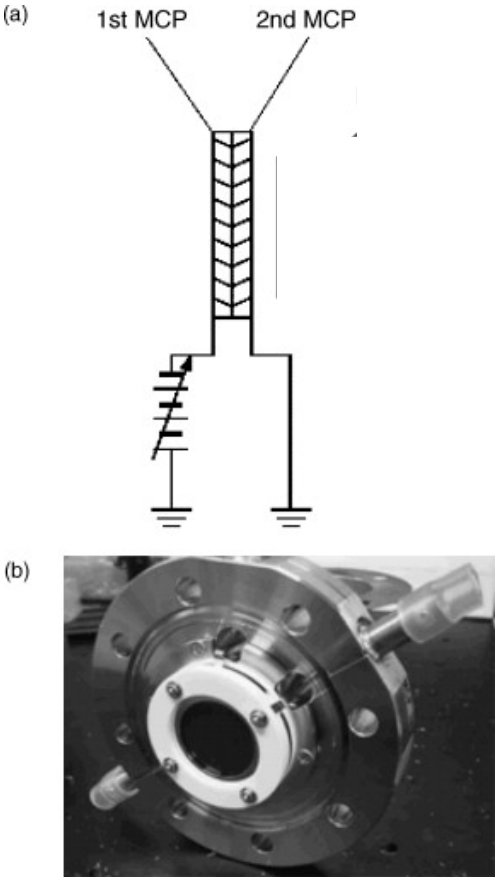

Figure 1: (a) A schematic diagram of a chevron detector. (b) A photograph of a chevron detector[6].

The microchannel plates are a mechanism to amplify an initial electron event into an avalanche. Each plate, consisting of a two dimensional periodic array of very small glass capillaries fused together, is sliced into a thin plate. An electron enters a channel and then emits another electron from the channel wall. Secondary electrons knocked from the walls of the channels are then accelerated by an electric field due to an applied voltage. The secondary electrons in turn continue

to cause new electrons to be emitted from the channel walls producing more secondary electrons. Thus an avalanche is created. A readout device, in this case, a metal anode, is located at the end of the detector to detect the avalanche[1]. The number of electrons at the end of the avalanche is recorded and a MALDI spectra is obtained for the sample based on each particle's m/z ratio and relative abundance[3].

After correction for background noise, the spectra are then analyzed. Each peak is normalized, based on calibration data. There is an observed variability in the detector, proposed to be due to the avalanche dynamics. The data shows a long tail at high amplitude and a hard zero at low amplitude, which yields an asymmetric distribution. This irregularity can cause poor estimates of peak height for small sample sizes. The avalanche dynamics studied are thought to affect quality control spectra. These spectra are obtained from blood sera from several hundred individuals that are randomly selected and believed to be healthy. The blood sera is pooled together and mixed completely. Many spectra are obtained from the pooled sera, which should have identical spectra. However, there are observed variations in the spectra. This may be explained by shot to shot variations of the laser or the biochemistry on the MALDI plate. Our hypothesis is that it is that the avalanche statistics, which produce a variable number of final electrons for each number of initial electrons, contributes to the observed variability in the acquired data[5]. Through Matlab modeling of the avalanche dynamics in the detector, a thorough mathematical understanding of the avalanche process can be gained. Through modeling and thoroughly understanding the avalanche process, better estimates of peak height can be obtained for small sample sizes. Thus, the data obtained might be corrected in order to obtain a more accurate spectrum.

2 Monte Carlo Simulation of Avalanches

The detector has 30 stages. In each stage, an electron can be reabsorbed into the wall of the detector, creating 0 from 1, remain unchanged, or knock another electron from the walls of the detector, producing 2 from 1. Each of these three options occurs with specific probabilities, denoted p_0, p_1, p_2. The detector outputs a gain of approximately 10^6 (i.e. 10^6 electrons at the end of the avalanche for every initial electron)[4]. By writing a program that repeats for 30 stages with varying absorption and creation probabilities, from 0 to 0.8, with increments of 0.01, probabilities were chosen that yield a gain of 10^6. A Matlab program was then written with a 10% probability of absorption, a 70% probability of creation, and a 20% chance that the electron remains unchanged. This gives a gain per stage of 1.6: given one electron, the expected number of electrons at the second stage is $n' = 0*p_0 + 1*p_1 + 2*p_2 = 1.6$. The proper probabilities characteristic of the detector are unknown, but initial probabilities were chosen and will be used throughout the rest of the calculations. Beginning with a specified number of electrons, the program randomly selects either -1, 0, or 1 for each electron in each stage of the detector with the specified probabilities and sums the number of electrons at the end of each stage, for 30 stages. Thus, the program outputs the number of electrons at the end of the detector beginning with 1-10 electrons. This process is then repeated 10,000 times to output a probability distribution that for n_i starting electrons, n_f electrons reach the end of the detector. The number of electrons at the end of the detector was counted in increments of 100,000 to obtain the probability distribution.

Plots were produced for a starting number of electrons from 1 to 10 using the data obtained from the Matlab program. The number of electrons was counted in increments of 100,000 to obtain a percentage of outputting n electrons.

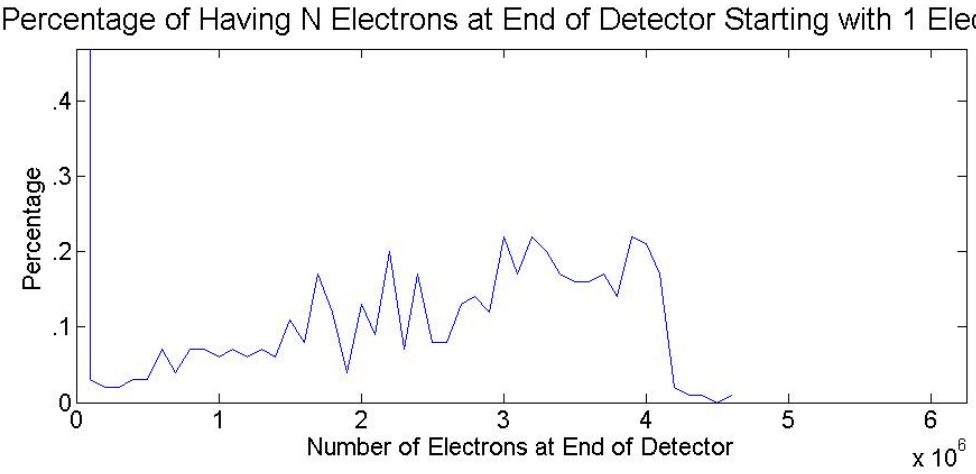

Figure 2: A figure of the avalanche data for an avalanche starting with 1 electron with a condensed range to illuminate curve behavior. The average number of electrons at the end of an avalanche beginning with one electron, denoted $\langle n \rangle_1 = 1.3 * 10^6$, which is observed in the figure.

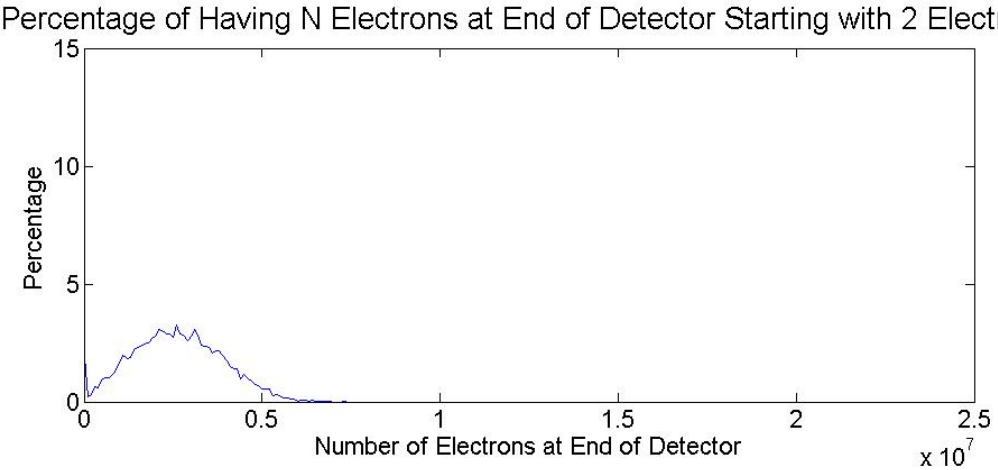

Figure 3: Percentage of avalanches having n_f electrons at the end of an avalanche beginning with 2 electrons. The average number of electrons at the end of an avalanche beginning with two electrons, denoted $\langle n \rangle_2 = 2.6 * 10^6 = 2\langle n \rangle_1$, is observed in the figure.

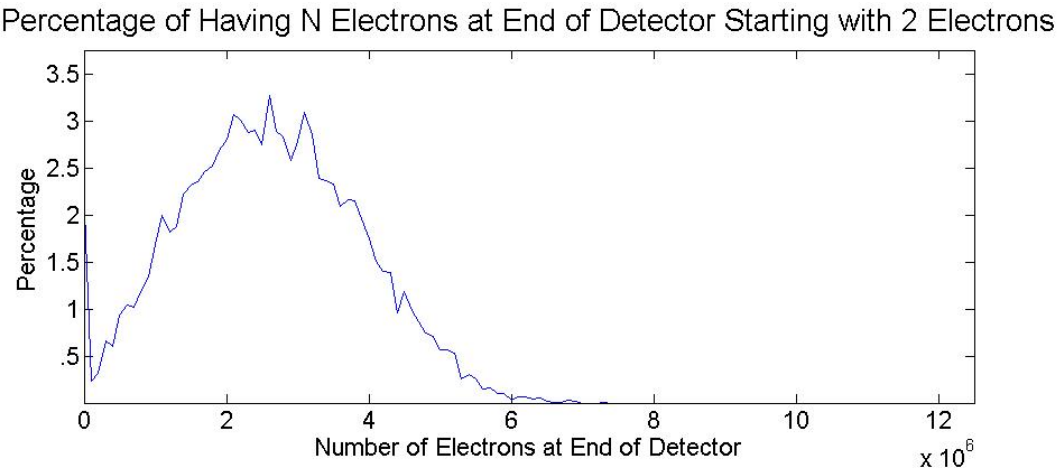

Figure 4: A figure of the avalanche data for an avalanche starting with 2 electrons with a condensed range to illuminate curve behavior. The average number of electrons at the end of an avalanche beginning with two electrons, denoted $\langle n \rangle_2 = 2.6*10^6 = 2\langle n \rangle_1$, is observed in the figure.

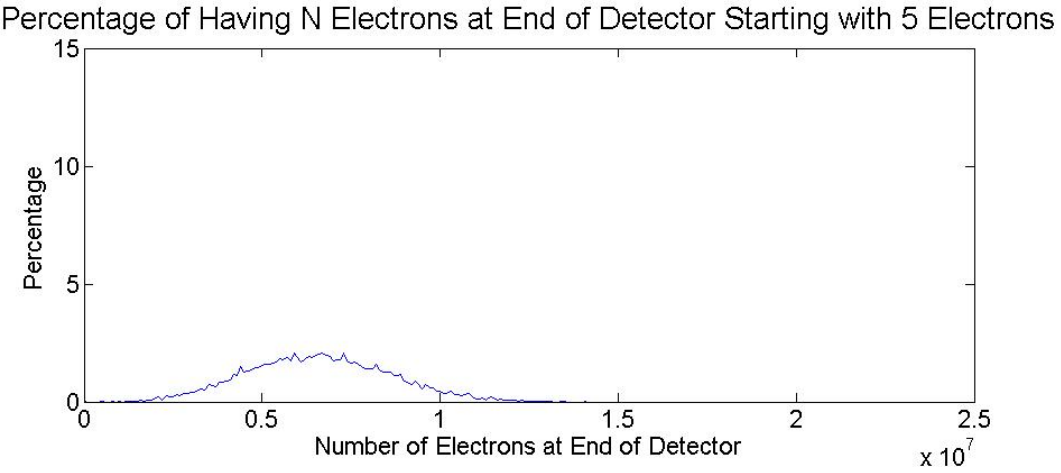

Figure 5: Percentage of avalanches having n_f electrons at the end of an avalanche beginning with 5 electrons. The average number of electrons at the end of an avalanche beginning with five electrons, denoted $\langle n \rangle_5 = 6.6*10^7 = 5\langle n \rangle_1$, is observed in the figure.

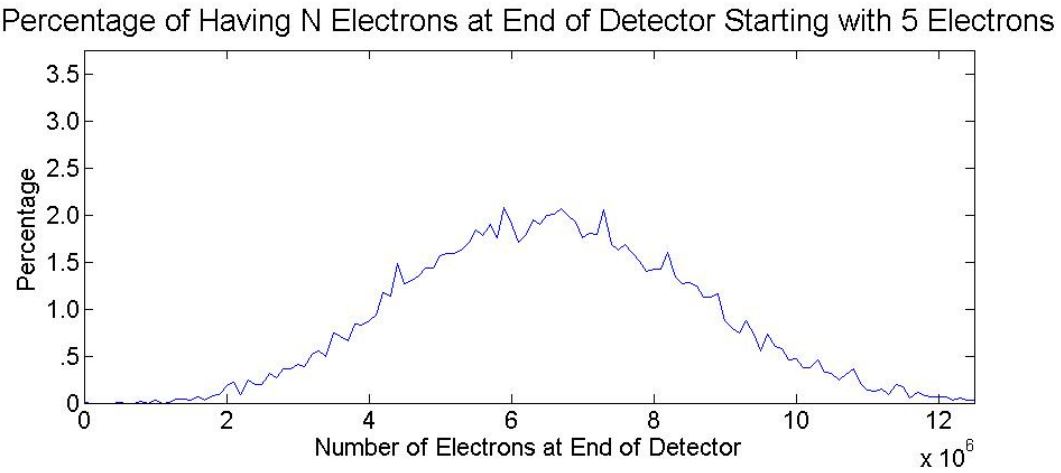

Figure 6: A figure of the avalanche data for an avalanche starting with 5 electrons with a condensed range to illuminate curve behavior. The average number of electrons at the end of an avalanche beginning with five electrons, denoted $\langle n \rangle_5 = 6.6 * 10^7 = 5 \langle n \rangle_1$, is observed in the figure.

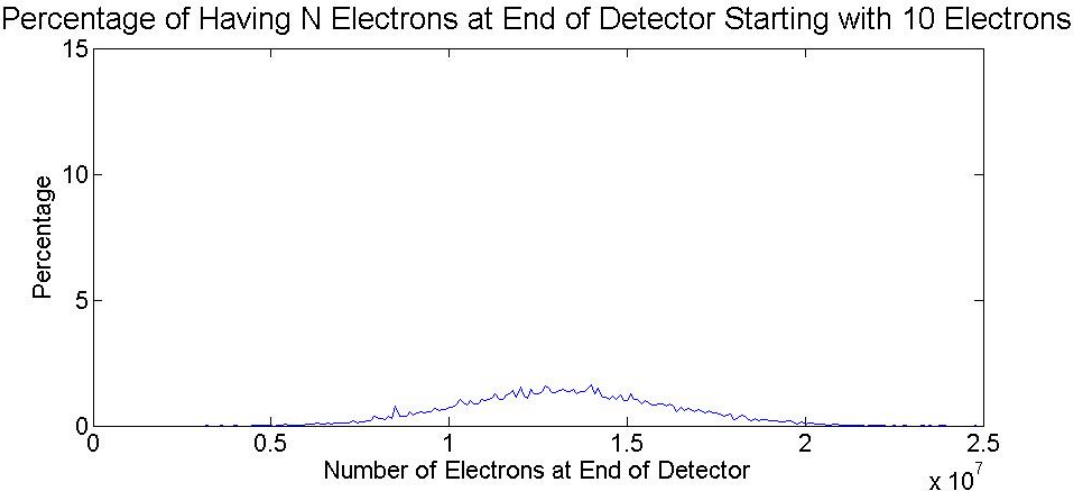

Figure 7: Percentage of avalanches having n_f electrons at the end of an avalanche beginning with 10 electrons. The average number of electrons at the end of an avalanche beginning with ten electrons, denoted $\langle n \rangle_{10} = 1.3 * 10^7 = 10 \langle n \rangle_1$, is observed in the figure.

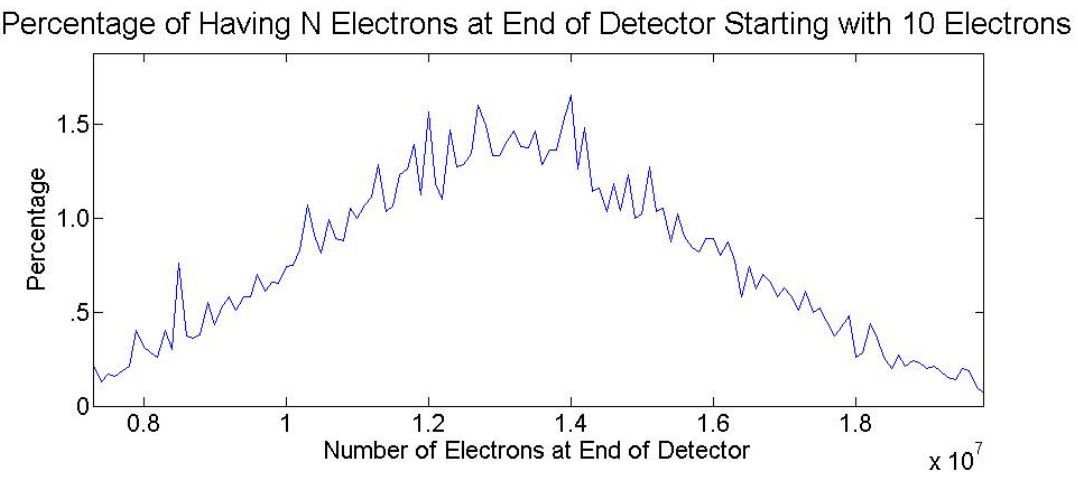

Figure 8: A figure of the avalanche data for an avalanche starting with 10 electrons with a condensed range to illuminate curve behavior. The average number of electrons at the end of an avalanche beginning with ten electrons, denoted $\langle n \rangle_{10}=1.3*10^7=10\langle n \rangle_1$, is observed in the figure.

Beginning with 1 electron, there is a high probability of having 0 electrons at the end of the detector. Starting with 2 electrons, a clearly defined distribution arises clustered around 2.5 million. As the beginning number of electrons increases, the average value of the distribution increases and the range increases, spreading the curve.

3 Maximum Entropy Analysis of the Monte Carlo Results

In order to study maximum entropy methods, a preliminary example was first evaluated. Let $A_1, A_2,\ldots, A_M$ be an alphabet with M symbols to represent the output of an experiment with M possible outcomes. The probabilities that each of these outcomes will occur are $p_1, p_2,\ldots, p_m$, where

$$\sum p_k = 1 \text{ with } p_k \neq 0 \text{ and k=1...M.} \tag{3}$$

In this thought experiment, there are N independent observations, where a random symbol from the alphabet is observed, or a random output of the experiment is observed, where N can be greater than, equal to, or less than m. These observations are $A_{i_1}, A_{i_2}, ..., A_{i_N}$ or written more compactly $i_1, i_2,\ldots,i_N$.

An example of the observed outcomes would be 5, 32, 6,..., M-1, 16, where the integers correlate to the outcomes of the experiment.

There are M^N possible symbol strings. We next determine an equation for how many symbol strings have A_k occur N_k times, where:

$$\sum N_k = N. \tag{4}$$

This equation is a multinomial expression

$$\frac{N!}{N_1!N_2!...N_M!}, \tag{5}$$

which is the number of symbol strings with A_k occurring N_k times.

Next we must find how many observed symbol strings are consistent with our assignment of the probabilities p_1, p_2,…, p_M. Consistency means

$$N_k \sim p_k N. \tag{6}$$

We can then define a function $\Omega_N(p)$ that is the number of strings of length N consistent with p_1, p_2,…,p_M:

$$\Omega_N(p) = \frac{N!}{(p_1N)!(p_2N)!...(p_MN)!}. \tag{7}$$

Using Stirling's approximation, we find that

$$\Omega_N(p) \sim e^{NH(p)} \quad \text{as } N \rightarrow \infty. \tag{8}$$

Here

$$H(p) = -\sum_{k=1}^{M} p_k \ln(p_k), \tag{9}$$

and is the Shannon entropy, it is the "growth rate" of $\Omega_N(p)$ as N is increased. It is called an entropy because entropy is used to find the number of microstates consistent with the macro knowledge and external constraints of a system.

Maximum entropy methods are used to assign the probabilities of a probability distribution when the probabilities are unknown, but some properties of these probabilities are known. For example, suppose the probabilities p_1, p_2,…, p_M are unknown quantities but the following are known about them:

$$\sum_{k=1}^{M} p_k = 1 \tag{10}$$

$$\sum_{k=1}^{M} k p_k = \mu, \tag{11}$$

where μ is the mean of the probability distribution. The probabilities must be chosen subject to these constraints, while biasing the result as little as possible. Thus, the probabilities must be assigned by maximizing the number of symbol strings consistent with these constraints. The

number of strings grows like the entropy, hence the name maximum entropy methods. The probabilities are then assigned in order to maximize the Shannon entropy while satisfying the constraints.

Example 1:

In this example, we have no knowledge other than the fact that p_k are probabilities. We will use the method of Lagrange multipliers to maximize the Shannon entropy by setting the derivatives with respect to each variable equal to zero. We define a quantity

$$\varepsilon(p,\lambda_0) = -\sum_{k=1}^{M} p_k \ln(p_k) + \lambda_0 (\sum_{k=1}^{M} p_k - 1). \tag{12}$$

Taking the derivative with respect to λ_0 gives the constraint

$$\frac{\partial \varepsilon}{\partial \lambda_0} = 0 \Rightarrow \sum_{k=1}^{M} p_k = 1. \tag{13}$$

Using the fact that the p_k are independent:

$$\frac{\partial p_k}{\partial p_j} = \delta_{jk}, \tag{14}$$

We find:

$$\frac{\partial \varepsilon}{\partial p_j} = 0 \Rightarrow -\sum_{k=1}^{M} \delta_{jk} \ln(p_k) - \sum \frac{p_k}{p_k} \delta_{jk} + \lambda_0 \sum_{k=1}^{M} \delta_{jk}, \tag{15}$$

Therefore

$$-\ln(p_j) - 1 + \lambda_0 = 0, \tag{16}$$

which gives

$$\ln(p_j) = \lambda_0 - 1. \tag{17}$$

Therefore p_j is independent of j. We now fix p_k by normalizing:

$$\sum_{k=1}^{M} p_k = 1 \Rightarrow p_k = \frac{1}{M} \tag{18}$$

In order for the entropy to maximized, each possible symbol string or experimental outcome will appear with equal probability.

Example 2: Given knowledge of the mean, we then use another Lagrange multiplier to add the constraint

$$\sum_{k=1}^{M} kp_k = \mu. \tag{19}$$

So that now we have two Lagrange multipliers:

$$\varepsilon(p, \lambda_0, \lambda_1) = -\sum_{k=1}^{M} p_k \ln(p_k) + \lambda_0 (\sum_{k=1}^{M} p_k - 1) + \lambda_1 (\sum_{k=1}^{M} kp_k - \mu). \tag{20}$$

This gives:

$$\frac{\partial \varepsilon}{\partial \lambda_0} = 0 \Rightarrow \sum_{k=1}^{M} p_k = 1, \tag{21}$$

and

$$\frac{\partial \varepsilon}{\partial \lambda_1} = 0 \Rightarrow \sum_{k=1}^{M} kp_k = \mu, \tag{22}$$

for our two constraints. Now taking the derivative with respect to p_j we find

$$\frac{\partial \varepsilon}{\partial p_j} = -\ln(p_j) - 1 + \lambda_0 + \lambda_1 j = 0, j = 1,..., M. \tag{23}$$

Therefore:

$$\ln(p_j) = -1 + \lambda_0 + \lambda_1 j, \tag{24}$$

or

$$p_j = e^{-1+\lambda_0+\lambda_1 j}. \tag{25}$$

In this case, the probability distribution is an exponential in j. We can fix the values of λ_0 and λ_1 in the following manner. Writing

$$(e^{-1+\lambda_0})(e^{\lambda_1})^j = \alpha\beta^j, \tag{26}$$

we normalize to get

$$\alpha\sum_{j=1}^{M}\beta^{j}=1, \tag{27}$$

And fix the mean by requiring:

$$\alpha\sum_{j=1}^{M}j\beta^{j}=\mu \tag{28}$$

$$\alpha=\frac{1}{\sum_{j}\beta^{j}} \tag{29}$$

and

$$\mu=\frac{\sum_{j}j\beta^{j}}{\sum\beta_{j}}=\frac{\beta\frac{d}{d\beta}\sum_{j=1}^{M}\beta^{j}}{\sum_{j=1}^{M}\beta^{j}}. \tag{30}$$

We know that

$$\sum_{j=1}^{M}\beta^{j}=1+\beta+\ldots+\beta^{M}. \tag{31}$$

and

$$\beta\sum_{j=1}^{M}\beta^{j}=\beta+\ldots+\beta^{M}+\beta^{M+1}. \tag{32}$$

Thus, we find

$$(1-\beta)\sum\beta^{j}=1-\beta^{M+1} \tag{33}$$

and finally,

$$\sum_{j=1}^{M}\beta^{j}=\frac{1-\beta^{M+1}}{1-\beta}. \tag{34}$$

Now we must apply this to our avalanche problem. We can assign each outcome of our avalanche a symbol. The following statements:

A_0=0 electrons at the end of the avalanche given 1 electron at the beginning of the avalanche

A_1=1 electron at the end of the avalanche given 1 electron at the beginning of the avalanche

A_2=2 electrons at the end of the avalanche given 1 electron at the beginning of the avalanche

.

.

.

A_M=M electrons at the end of the avalanche given 1 electron at the beginning of the avalanche

Therefore, we know the mean value and the maximum number of electrons at the end of the avalanche. Now to treat the probabilities as a continuum and, thus, p_j becomes p(x). Our variational principal becomes

$$\varepsilon([p],\lambda_0,\lambda_1) = -\int_0^\infty dx p(x)\ln(p(x)) + \lambda_0(\int_0^\infty dx p(x) - 1) + \lambda_1(\int_0^\infty dx x p(x) - \mu) \tag{35}$$

Where the notation ε([p], λ_0, λ_1) indicates that ε is a function of λ_0 and λ_1, but a functional of p(x) (i.e. a function of an infinite number of variables, one for each value of x).

Taking the derivative with respect to λ_0 we find

$$\frac{\partial\varepsilon}{\partial\lambda_0} = 0 \Rightarrow \int_0^\infty dx p(x) = 1, \tag{36}$$

taking the derivative with respect to λ_1 we find

$$\frac{\partial\varepsilon}{\partial\lambda_1} = 0 \Rightarrow \int_0^\infty dx x p(x) = \mu. \tag{37}$$

We then introduce a variation on p(x)

$$p(x) \to p(x) + \xi\eta(x) \tag{38}$$

and compute

$$\lim_{\xi\to 0}\frac{\varepsilon[p+\xi\eta]-\varepsilon[p]}{\xi} \tag{39}$$

$$\frac{\partial\varepsilon[p+\xi\eta(x)]}{\partial\xi}|_{\xi=0} = \int dx\eta(x)[-\ln(p(x)) - 1 + \lambda_0 + \lambda_1 x], \tag{40}$$

where the quantity in brackets is defined to be the variational derivative of ε with respect to p(x) denoted

$$\frac{\delta\varepsilon}{\delta p(x)} = -\ln(p(x)) - 1 + \lambda_0 + \lambda_1 x = 0. \tag{41}$$

We set

$$\frac{\delta\varepsilon}{\delta p(x)} = 0 \tag{42}$$

and find

$$\ln(p(x)) = -1 + \lambda_0 + \lambda_1 x. \tag{43}$$

Therefore, p(x) is exponential in x. The most general form is given by

$$p(x) = \frac{1}{\mu} e^{-\frac{x}{\mu}}. \tag{44}$$

We now can use this expression that represents the probability of having x electrons at the end of the avalanche beginning with 1 electron to find expressions for avalanches starting with any number of electrons.

First

$$p(x) = p(x \mid 1), \tag{45}$$

where x should be interpreted as the statement

$$p_1(x \quad electrons \quad out \mid 1 \quad electron \quad in \quad and < x >_1 = \mu dx) = \frac{1}{\mu} e^{-\frac{x}{\mu}} dx. \tag{46}$$

Here x is the number of electrons out that lie between x and x+dx and $<x>_1$ is the average number of electrons out, given 1 electron at the beginning of the avalanche.

Now we find an expression that represents the equation for the probability that x electrons are found at the end of the avalanche with two electrons at the beginning and an average output for the single electron initiated avalanche of μ

$$p_2(x \quad electrons \quad out \mid 2 \quad electrons \quad in \quad and < x >_1 = \mu). \tag{47}$$

We assume that the two avalanches are additive and independent. The joint probability of x_1 electrons in avalanche 1 and x_2 electrons in avalanche 2 is then derived.

$$p(x_1, x_2 \mid 1 \;\; electron \;\; starting \;\; both \;\; avalanches \;\; and \;\; < x >_1 = \mu \;\; for \;\; both) = p(x_1 \mid 1e^-, \mu)\, p(x_2 \mid 1e^-, \mu).$$

(48)

This can then be written as

$$p(x_1 + x_2 = x \mid 2 \;\; additive \;\; independent \;\; avalanches \;\; with < x >_1 = \mu \;\; and \;\; 1 \;\; electron \;\; starting \;\; each) = p(x \mid 2e^- s, \mu).$$

(49)

Where $x_1+x_2=x$ represents the length of a line between x_1 and x_2 that grows linearly with x.

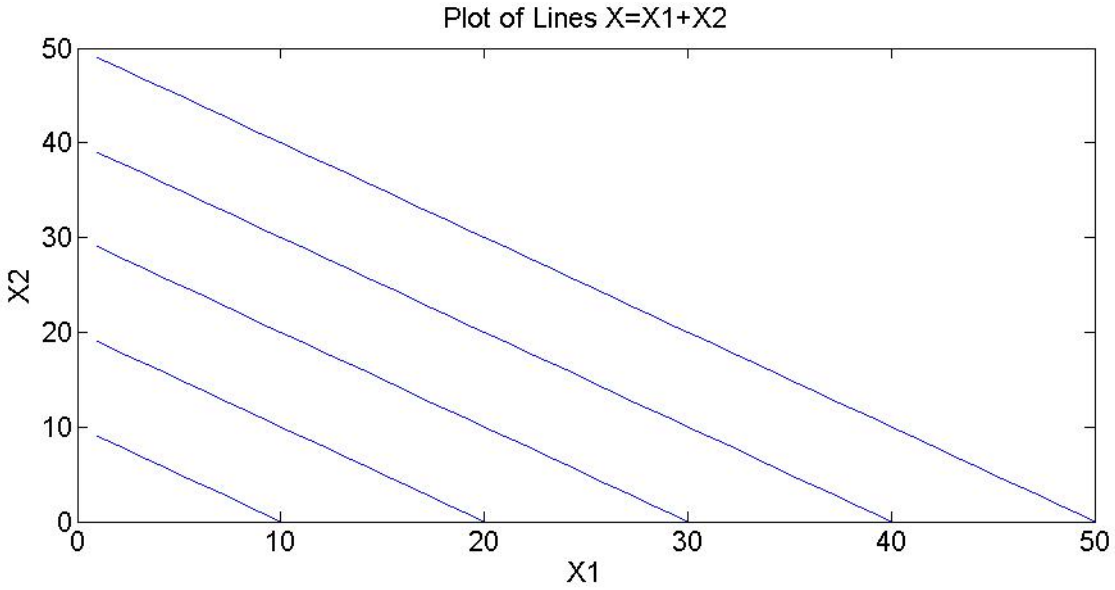

Figure 9: Plot demonstrating the additive properties of x1 and x2. The lines here represent x=x1+x2, where the length of the lines are growing linearly with x.

$$\int_{dx_1 dx_2} p(x_1, x_2 \mid 1, \mu) = p(x_1 + x_2 = x \mid 2e^- s, \mu)$$

(50)

$$\int dx_1 dx_2 \delta(x - (x_1 + x_2))\, p(x_1, x_2 \mid 1e^-, \mu) = \int_0^\infty dx_1 \int_0^\infty dx_2 \delta(x - (x_1 + x_2))\, p(x_1, x_2 \mid 1e^-, \mu)$$

(51)

$$x_2 = x - x_1$$

(52)

$$p(x_1 + x_2 = x \mid 2e^- s, \mu) = \int dx_1 \, p(x_1 \mid 1e^-, \mu) \, p((x - x_1) \mid 1e^-, \mu) \tag{53}$$

where

$$p(x \mid 1e^-, \mu) = \frac{1}{\mu} e^{-\frac{x}{\mu}} \tag{54}$$

Thus,

$$p(x_1 + x_2 = x \mid 2e^- s, \mu) = \frac{1}{\mu} e^{-\frac{x_1}{\mu}} \frac{1}{\mu} e^{-\frac{(x - x_1)}{\mu}} \tag{55}$$

$$= \frac{1}{\mu^2} e^{-\frac{x}{\mu}} \tag{56}$$

$$= \frac{1}{\mu^2} e^{-\frac{x}{\mu}} \int_0^x dx \tag{57}$$

$$= \frac{x}{\mu^2} e^{-\frac{x}{\mu}} \tag{58}$$

From figure 9 it is shown that the integrals over x1 and x2 convert to a finite integral over x, due to the change of variables. Treating each avalanche as separate and independent we can determine an expression for the probability distribution starting with any number of electrons. We claim this expression to have the form

$$p(x_1 + x_2 + \ldots + x_N = x \mid Ne^- s, \mu) = \frac{c}{\mu} \left(\frac{x}{\mu} \right)^{N-1} e^{\frac{-x}{\mu}} \tag{59}$$

where the normalization c is fixed by requiring

$$\int_0^\infty \frac{c}{\mu} \left(\frac{x}{\mu} \right)^{N-1} e^{-\frac{x}{\mu}} dx = 1. \tag{60}$$

As N approaches infinity $p(x|Ne^-,\mu)$ will look like a Gaussian centered around μ and the width of the Gaussian will be easily predicted.

In order to determine c, we must take this integral, which returns a gamma function

$$\Gamma(z) = \int_0^\infty t^{-1+z} e^{-t} dt \tag{61}$$

$$\Gamma(z) = (z-1)! \tag{62}$$

Therefore

$$\int_0^\infty \frac{c}{\mu}\left(\frac{x}{\mu}\right)^{N-1} e^{-\frac{x}{\mu}} dx = 1 \tag{63}$$

becomes

$$c\Gamma(N) = 1 \tag{64}$$

$$c(N-1)! = 1 \tag{65}$$

$$c = \frac{1}{(N-1)!} \tag{66}$$

Therefore

$$p(x \mid Ne^{-}s, \mu) = \frac{1}{(N-1)!\mu}\left(\frac{x}{\mu}\right)^{N-1} e^{-\frac{x}{\mu}} \tag{67}$$

The resulting figures using this formula to determine the probability of having x, which represents n_f, electrons at the end of an avalanche beginning with N electrons are shown below. These figures represent the theoretical probability densities and curves we would expect our avalanche data to possess.

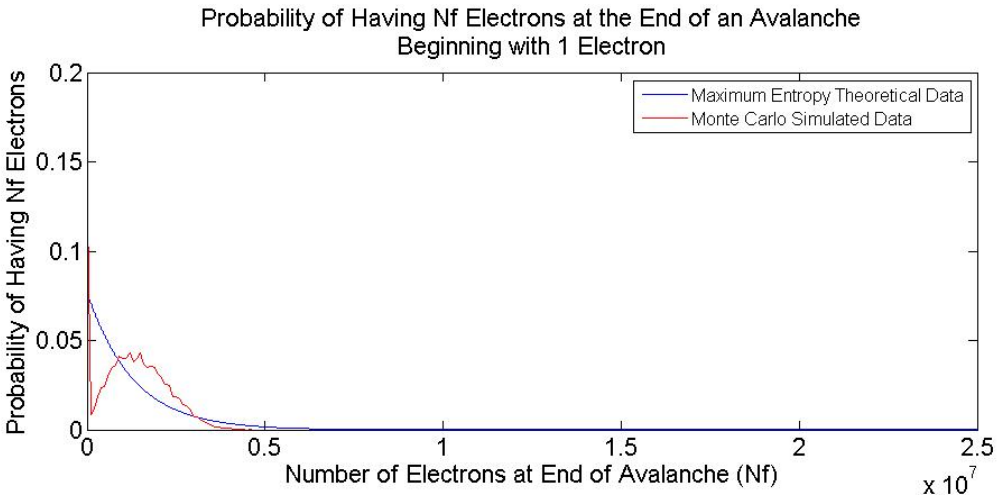

Figure 10: A figure comparing the theoretical probabilities based on maximum entropy calculations with the probabilities based on Monte Carlo simulations of having a certain number of electrons at the end of an avalanche beginning with 1 electron. In the maximum entropy calculations, the probabilities were multiplied by a bin width of 100,000 to match that of the Monte Carlo calculations.

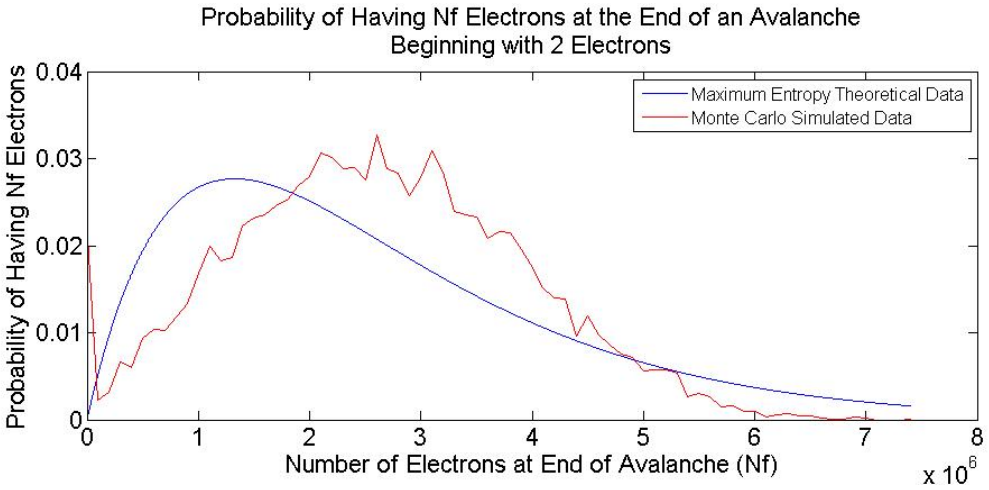

Figure 11: A figure comparing the theoretical probabilities based on maximum entropy calculations with the probabilities based on Monte Carlo simulations of having a certain number of electrons at the end of an avalanche beginning with 2 electrons. In the maximum entropy calculations, the probabilities were multiplied by a bin width of 100,000 to match that of the Monte Carlo calculations.

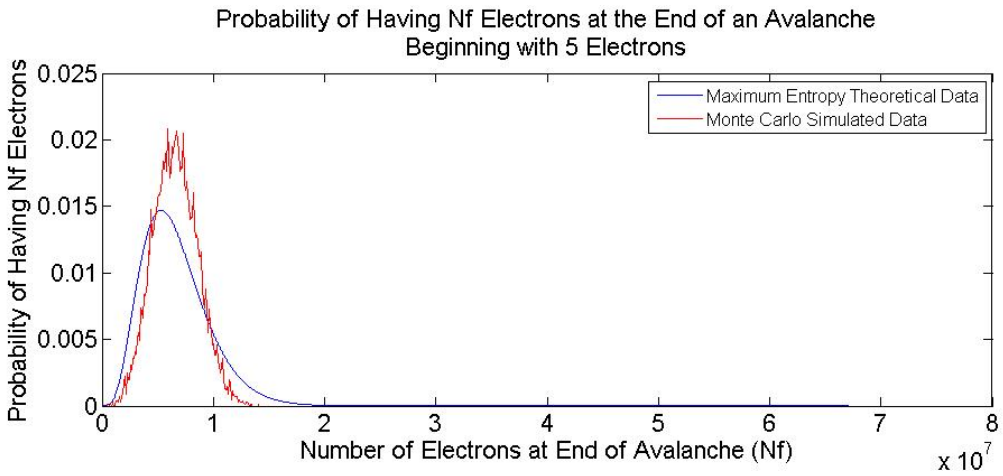

Figure 12: A figure comparing the theoretical probabilities based on maximum entropy calculations with the probabilities based on Monte Carlo simulations of having a certain number of electrons at the end of an avalanche beginning with 5 electrons. In the maximum entropy calculations, the probabilities were multiplied by a bin width of 100,000 to match that of the Monte Carlo calculations.

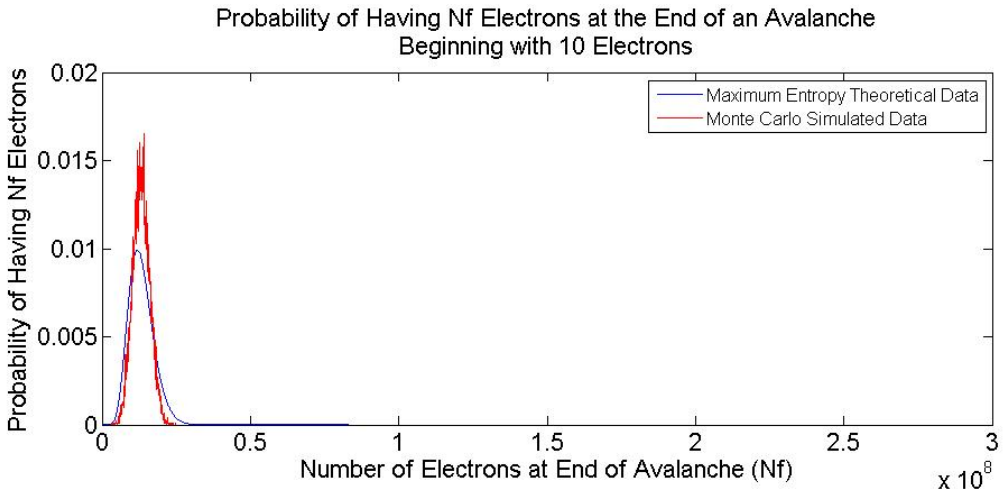

Figure 13: A figure comparing the theoretical probabilities based on maximum entropy calculations with the probabilities based on Monte Carlo simulations of having a certain number of electrons at the end of an avalanche beginning with 10 electrons. In the maximum entropy calculations, the probabilities were multiplied by a bin width of 100,000 to match that of the Monte Carlo calculations.

Figures 10-13 are from maximum entropy calculations and model the Monte Carlo simulations that were performed. The fit for avalanches beginning with 1 electron is not a good fit, and though the Monte Carlo simulations show a basic exponential shape, the dip in the middle is

not present in the maximum entropy calculations. For avalanches beginning with two electrons, the maximum entropy calculations are shifted to the left and have a longer tail with a less steep slope. The fit has the general shape of the simulations, but it can not be concluded that the fit is good. For avalanches beginning with 5 electrons, the fit has the same general shape, but the peak height is too low, as is the case for avalanches beginning with 10 electrons. Therefore, the maximum entropy calculations are not good approximations for the Monte Carlo simulations.

4 Markov Analysis of Avalanche Dynamics

As another approach to the study of avalanches, an equation to model the evolution of the probability distribution from one stage to the next was obtained. Next, the chance that n electrons will reach the end of the detector can be predicted from an avalanche beginning with any number of electrons. Starting at step t, with a number of ni electrons, step t+1 will have nf electrons, from an equation of the form

$$n_f = -n_0 + n_2 + n_i \tag{68}$$

where n_0 represents the number of reabsorbed electrons, n_2 the number of electrons that knock off another electron and n_1 the number of unchanged electrons.

The probability that n_f electrons are produced at the t+1 stage that began with n_i electrons will be obtained. The equation will be a sum over the following quantity:

$$p(n_f \mid n_i) = \sum_{n_0 n_1 n_2}{}' \frac{n_i!}{n_0! n_1! n_2!} p_0^{n_0} p_1^{n_1} p_2^{n_2} \tag{69}$$

With p_0 representing the probability that an electron is absorbed, p_1 the probability an electron is unchanged, and p_2 the probability that an electron produces another electron.

The previous equation rests on the following conditions. The number of electrons reabsorbed, unchanged, and added, must add to the initial number of electrons. The difference between the number of electrons added during this time-step and the number of electrons reabsorbed must be equal to the difference between the final number of electrons and the number of electrons at the beginning of the time-step:

$$n_0 + n_1 + n_2 = n_i \tag{70}$$

$$n_2 - n_0 = n_f - n_i \tag{71}$$

Because of these conditions n_0 and n_2 can be represented in terms of n_i, $n_{f,}$ and $n_{1.}$

$$n_0 = n_i - \frac{n_f - n_1}{2} \tag{72}$$

$$n_2 = \frac{n_f - n_1}{2} \tag{73}$$

The equation

$$P(n_0 n_1 n_2 | n_i p_0 p_1 p_2) \tag{74}$$

represents the probability of obtaining a particular number of reabsorbed, unchanged, and knocked off electrons, knowing the number of initial electrons and the probability of an electron being reabsorbed, unchanged, or knocking off another electron. A particular sequence probability is represented by

$$p_0^{n_0} p_1^{n_1} p_2^{n_2} \tag{75}$$

which is multiplied by the number of sequences of n_i electrons with a particular n_0, n_1, and n_2. Thus,

$$\frac{n_i!}{n_0! n_1! n_2!} p_0^{n_0} p_1^{n_1} p_2^{n_2} \tag{76}$$

represents the probability of having a particular absorbed, unchanged, and created combination starting with n_i initial electrons and the assigned probabilities. This quantity is then summed over all possible numbers of absorbed, unchanged, and created electrons starting with n_i electrons and having n_f electrons at the next time step.

For example, starting with n_i=2, n_f=0, 1, 2, 3, 4. For n_f=0, n_0=2, n_1=0, and n_2=0. Thus the equation becomes

$$\frac{2!}{2!0!0!}.1^2.2^0.7^0 = .01 \tag{77}$$

With n_f=1, n_0=1, n_1=1, and n_2=0. The probability of this particular sequence is then

$$\frac{2!}{1!1!0!}.1^1.2^1.7^0 = .04 \tag{78}$$

With n_f=2, n_0=0, n_1=2, and n_2=0 or n_0=1, n_1=0, and n_2=1, with a probability equation of

$$\frac{2!}{0!2!0!}.1^0.2^2.7^0 + \frac{2!}{1!0!1!}.1^1.2^0.7^1 = .04 + .14 = .18 \tag{79}$$

If n_f=3, n_0=0, n_1=1, and n_2=1. The probability of this particular sequence is then

$$\frac{2!}{0!1!1!}.1^0.2^1.7^1 = .28 \tag{80}$$

When n_f=4, n_0=0, n_1=0, and n_2=2, with a probability equation of

$$\frac{2!}{0!0!1!}.1^0.2^0.7^2 = .49 \tag{81}$$

The probabilities of all possible n_f with an initial n_i of 2 add to 1, as expected. This calculation for all n_i and n_f was performed using Matlab for n_i from 1 to 150 to form the Markov matrix discussed below.

A Markov matrix was constructed that summarizes the single step transition probabilities. It is a left stochastic matrix, a square matrix whose columns consist of nonnegative real numbers whose sum is 1. The matrix is a transition matrix for a finite Markov chain, or, more specifically, describes a Markov chain over a finite state space. Starting with n_i electrons, the matrix will summarize the probability that you get n_f electrons out in the next time step.[3] The Markov matrix equation of interest is

$$p'_f = M_{fi}\, p_i \tag{82}$$

where p'_f is a vector with the probabilities of obtaining n_f electrons and p_i is a vector with the probabilities of starting with n_i electrons. M_{fi} is the Markov matrix of interest. In order for this matrix to map probabilities to probabilities, the column sum is always unity

$$\sum_f M_{fi} = 1 \; for \; all \; i. \tag{83}$$

The Markov matrix was first constructed as a 10 by 10 matrix, calculating each entry for a specific n_i, the number of electrons at the beginning of the time step, and n_f, the number of electrons at the end of the time step. Each entry was calculated using equation (69) where p_0 is the probability of an electron being reabsorbed, p_1, the probability and electron is unchanged, p_2, the probability that an electron produces another electron, n_0, the number of reabsorbed electrons, n_1, the number of electrons unchanged, and n_2, the number of electrons that create another electron. For each matrix entry of having n_f electrons at the end of a time step beginning with n_i electrons, the probabilities from each combination of n_0, n_1, and n_2 are summed. It is ensured that the columns add to one.

This matrix was then expanded into a 150 by 150 matrix and a heat map was produced from the resulting matrix. Selecting columns and rows from the matrix, Gaussian plots were produced to model the behavior across a row or down a column. In order to better approximate the infinite matrix, with a goal of finding an accurate equation to model the matrix's behavior, Stirling's approximation was used for n_0, n_1, and n_2 greater than 10. Stirling's approximation is used to approximate factorials and has the form

EMBED Equation.3

(84)

Now each term in the matrix for n_0, n_1, and n_2 greater than 10 is represented by

$$p(n_f \mid n_i) = \sum_{n_0 n_1 n_2} \frac{\sqrt{2\pi n_i}\left(\frac{n_i}{e}\right)^{n_i}}{\sqrt{2\pi n_0}\left(\frac{n_0}{e}\right)^{n_0}\sqrt{2\pi n_1}\left(\frac{n_1}{e}\right)^{n_1}\sqrt{2\pi n_2}\left(\frac{n_2}{e}\right)^{n_2}} p_0^{n_0} p_1^{n_1} p_2^{n_2} \tag{85}$$

This was determined to be an accurate approximation. In order to model the avalanche for multiple time steps, this matrix was multiplied by itself, and heat maps were produced for the results. The resulting matrices from these multiplications were found to also display Gaussian behavior across the rows and down the columns. Heat maps were produced for each iteration of the multiplication.

Though a Markov matrix is finite, this matrix will be manipulated to approximate an infinite matrix, following the same conditions of a Markov matrix, which will serve as the transition between two infinite probability vectors, each represented by a finite approximation. Multiplying the matrix for multiple time steps will be used to determine the behavior of the electron avalanche throughout the 30 stages of the detector. The asymptotic behavior of the matrix from the heat maps will be determined in order to formulate an equation to approximate the behavior of the avalanche that can be applied to a growing number of electrons through the 30 stages. A function that represents the probability distribution will be acquired, where the probabilities must add to 1. The function at the next time step, with n_f electrons is then represented by the sum

$$f(n_f, t+1) = \sum_{n_i=0}^{\infty} p(n_f \mid n_i) f(n_i, t) \tag{86}$$

In order to form an equation to approximate the matrix and determine the most probable number of electrons at time step t+1, knowing the number of electrons at time step t, the natural logarithm of each term in each matrix entry was taken. So now, the dominant term in the sum is represented by the equation

$$\ln\left(\frac{\sqrt{2\pi n_i}\left(\frac{n_i}{e}\right)^{n_i}}{\sqrt{2\pi n_0}\left(\frac{n_0}{e}\right)^{n_0}\sqrt{2\pi n_1}\left(\frac{n_1}{e}\right)^{n_1}\sqrt{2\pi n_2}\left(\frac{n_2}{e}\right)^{n_2}} p_0^{n_0} p_1^{n_1} p_2^{n_2}\right) \tag{87}$$

Given n_i and n_f, with n_1 as our only free variable, we can write n_0 and n_2 as in equations 72 and 73.

Next, the derivative of each term in this function with respect to n_1 was taken and set to zero to determine the value of n_1 where the maximum term in the sum occurs for each term in the matrix.

$$\frac{d}{dn_1}\left[\ln\left[f\left(n_1\right)\right]\right]=0 \tag{88}$$

Where

$$f(n_1)=\frac{\sqrt{2\pi n_i}\left(\frac{n_i}{e}\right)^{n_i}}{\sqrt{2\pi\left(n_i-\frac{n_f-n_1}{2}\right)}\left(\frac{n_i-\frac{n_f-n_1}{2}}{e}\right)^{n_i-\frac{n_f-n_1}{2}}\sqrt{2\pi n_1}\left(\frac{n_1}{e}\right)^{n_1}\sqrt{2\pi\left(\frac{n_f-n_1}{2}\right)}\left(\frac{\frac{n_f-n_1}{2}}{e}\right)^{\frac{n_f-n_1}{2}}}$$

$$*\ {p_0}^{n_i-\frac{n_f-n_1}{2}}\,{p_1}^{n_1}\,{p_2}^{\frac{n_f-n_1}{2}} \tag{89}$$

After substituting in the equations for n_0 and n_2 in terms of n_1, n_i, and n_f, so the function has only these three variables. Consequently, for each term in the matrix, with known n_i and n_f, n_1 is the only variable. .

The natural log of this function is

$$\ln\left[f\left(n_1\right)\right]=\frac{1}{2}\ln\left(2\pi n_i\right)+n_i\ln\left(n_i\right)-n_i-\frac{1}{2}\ln\left(2\pi\left(n_i-\frac{n_f-n_1}{2}\right)\right)$$
$$-\left(n_i-\frac{n_f-n_1}{2}\right)\ln\left(n_i-\frac{n_f-n_1}{2}\right)+n_i-\frac{n_f-n_1}{2}-\frac{1}{2}\ln\left(2\pi n_1\right)-n_1\ln\left(n_1\right)+n_1$$
$$-\frac{1}{2}\ln\left(2\pi\left(\frac{n_f-n_1}{2}\right)\right)-\left(\frac{n_f-n_1}{2}\right)\ln\left(\frac{n_f-n_1}{2}\right)+\frac{n_f-n_1}{2}+\left(n_i-\frac{n_f-n_1}{2}\right)\ln\left(p_0\right)$$
$$+n_1\ln\left(p_1\right)+\left(\frac{n_f-n_1}{2}\right)\ln\left(p_2\right) \tag{90}$$

The derivative of this function is

$$\frac{d[\ln[f(n_1)]]}{dn_1} = -\frac{1}{4n_i - 2(n_f - n_1)} - \frac{1}{2}\ln\left(n_i - \frac{n_f - n_1}{2}\right) - \frac{1}{2n_1} - \ln(n_1) + \frac{1}{2(n_f - n_1)}$$
$$+ \frac{1}{2}\ln\left(\frac{n_f - n_1}{2}\right) + \frac{1}{2}\ln(p_0) + \ln(p_1) - \frac{1}{2}\ln(p_2)$$

(91)

A Matlab program was then written to determine the maximum value of n_1 using the above equation for specified values of n_f and n_i. Another program was then written to determine the maximum n_1 value for all n_f and n_i values. This program creates a three-dimensional matrix $p(n_f, n_i, n_1)$ that represents all $\ln[f(n_1)]$ values. It then finds the maximum value of the matrix with respect to the third dimension, or n_1. In other words, for each pair of n_i and n_f, the program determines the maximum value of $f(n_1)$. The matrix index of this value is determined, since the index represents the value of n_1, and in this case, $[n_1]$max. A matrix is then created of size (n_f, n_i) to store the index values for all n_1. However, the minimum value n_1 can have is 0. Matrix indices begin at 1, so a matrix of ones must be subtracted from the index matrix in order to determine the true maximum values of n_1.

Next, an approximation was found for the function $\ln[f(n_1)]$.

$$\ln[f(n_1)] \approx \ln[f(n_{1_{max}})] + \frac{1}{2}(n_1 - n_{1_{max}})^2 \frac{\partial^2 \ln(f)}{\partial {n_1}^2}\Big|_{n_{1_{max}}}$$

(92)

Where

$$\frac{\partial^2 \ln(f)}{\partial {n_1}^2} = \frac{2}{[4n_i - 2(n_f - n_1)]^2} - \frac{1}{4n_i - 2(n_f - n_1)} + \frac{1}{2{n_1}^2} - \frac{1}{n_1} - \frac{1}{2(n_f - n_1)^2} - \frac{1}{2(n_f - n_1)}$$

(93)

After exponentiating each side of this approximation, an approximation for $f(n_1)$ was acquired.

$$f(n_1) \approx f(n_{1_{max}}) e^{\frac{-(n_1 - n_{1_{max}})^2}{2{\sigma_n}^2}}$$

(94)

Where

$${\sigma_n}^2 = -\frac{1}{\frac{\partial^2 f}{\partial {n_1}^2}\Big|_{n_{1_{max}}}}$$

(95)

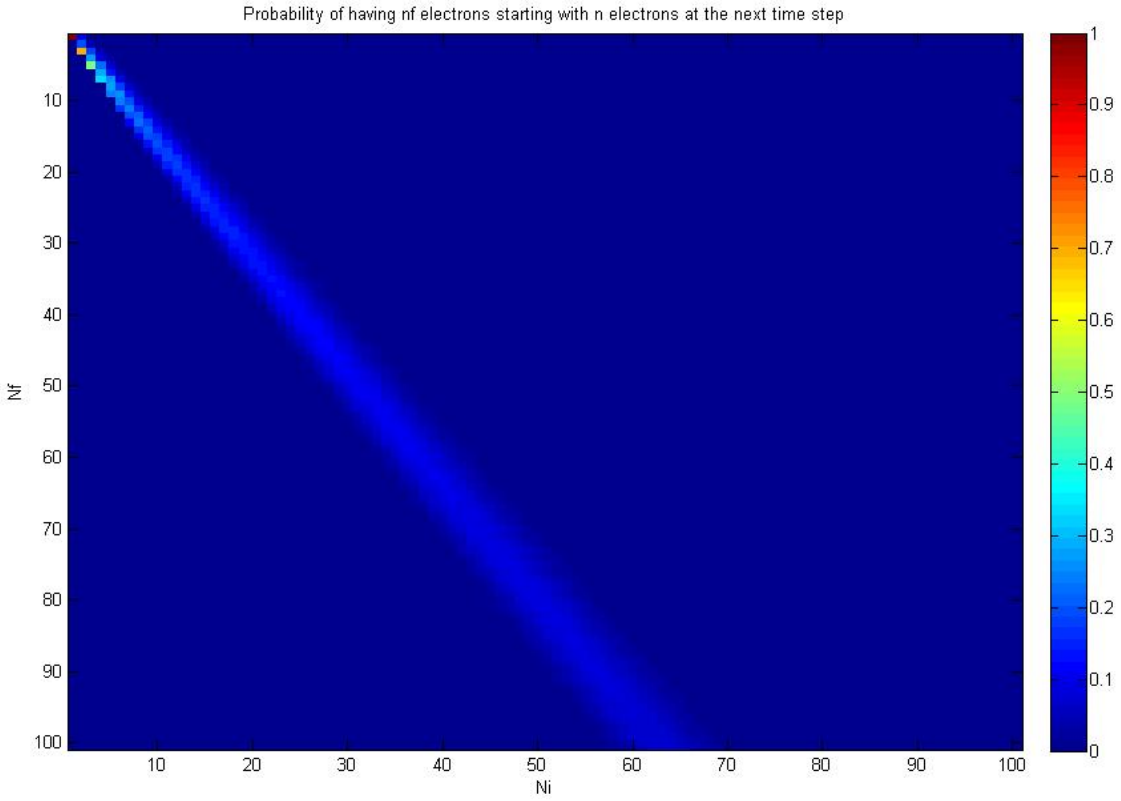

Figure 14: Heatmap of Markov Matrix showing the probability of having n_f electrons at the next time step starting with n_i electrons

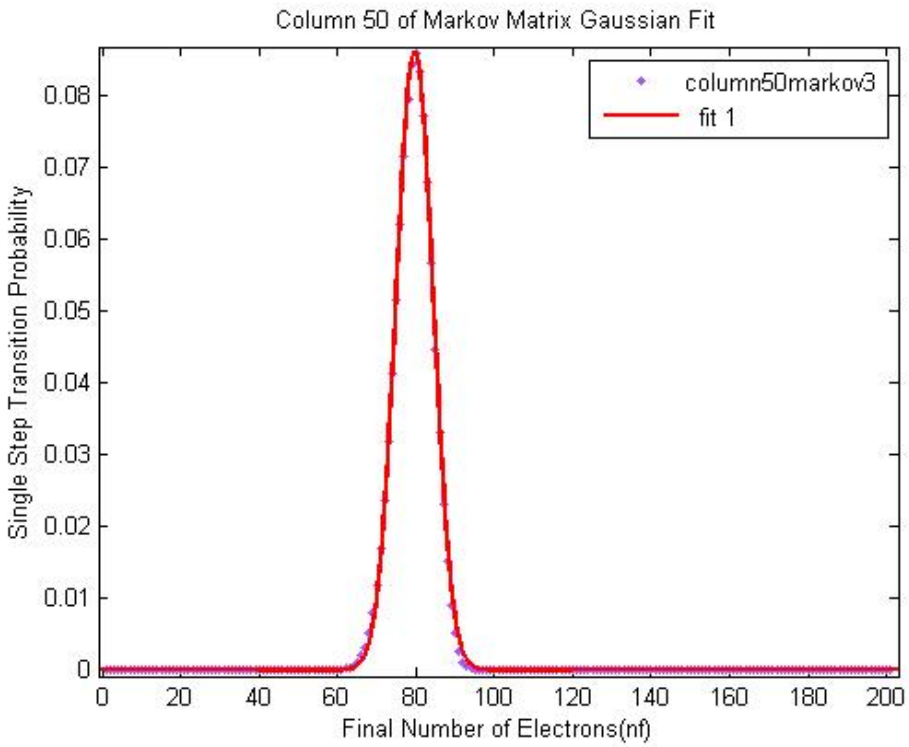

Figure 15: Gaussian Plot of Column 50 of the Markov Matrix demonstrating matrix behavior down a column

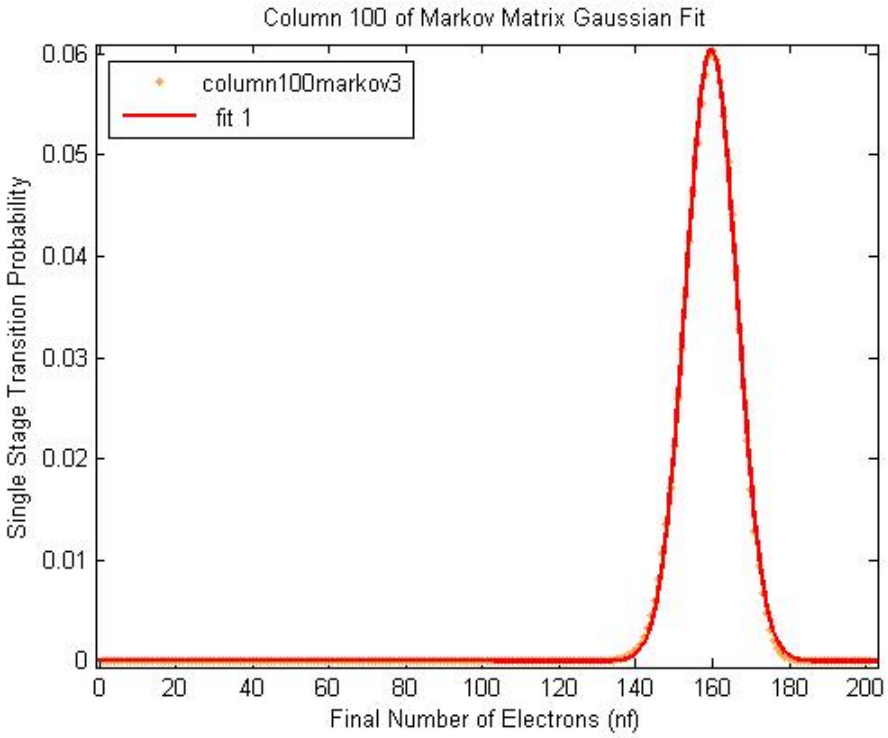

Figure 16: Gaussian Plot of Column 100 of Markov Matrix demonstrating matrix behavior down a column

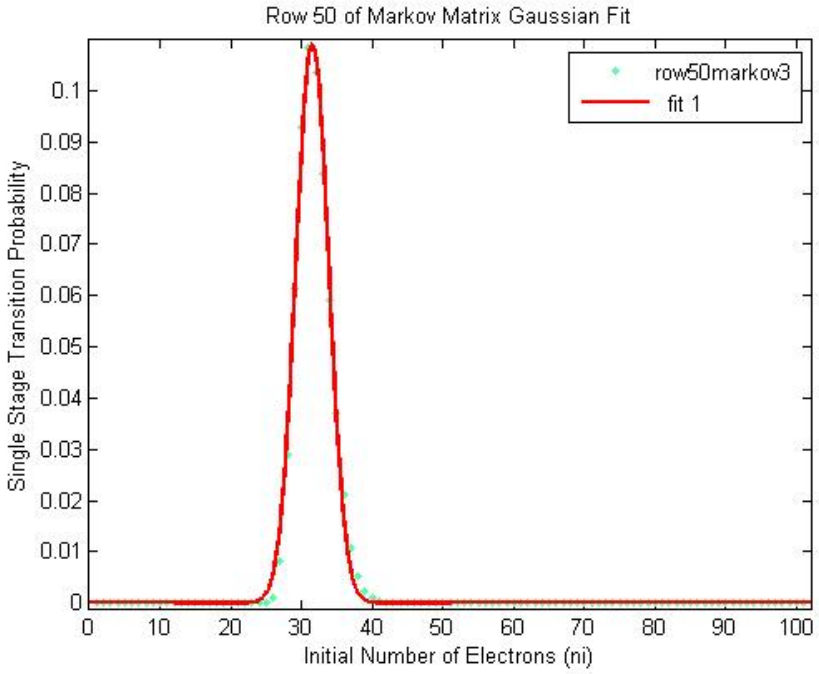

Figure 17: Gaussian Plot of Row 50 of Markov Matrix demonstrating matrix behavior across a row

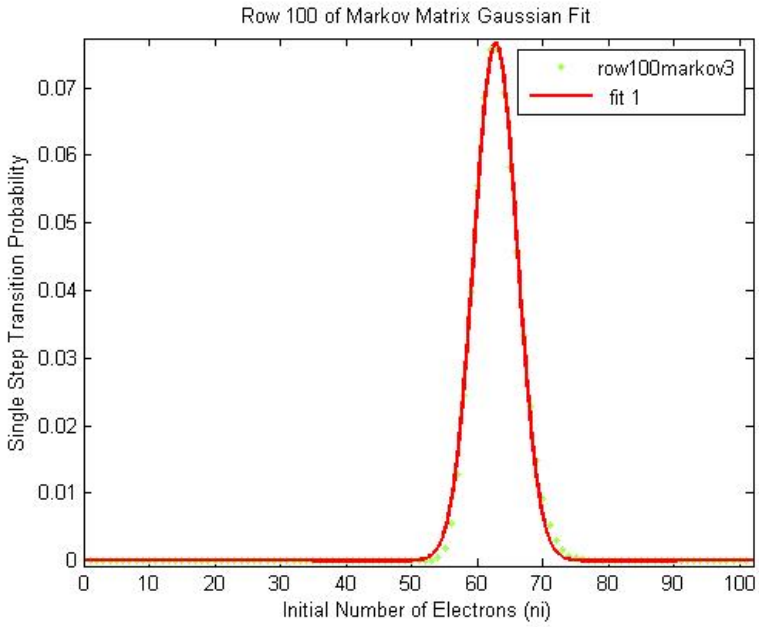

Figure 18: Gaussian Plot of Row 100 of Markov Matrix demonstrating matrix behavior across a row

Figures 15-18 demonstrate that the Markov matrix displays Gaussian behavior across its columns and rows. This demonstration of Gaussian behavior will be useful in developing a model to represent the probability distribution of electrons through 30 time steps. Gaussian behavior was also demonstrated for the Markov matrix representing 2 time steps.

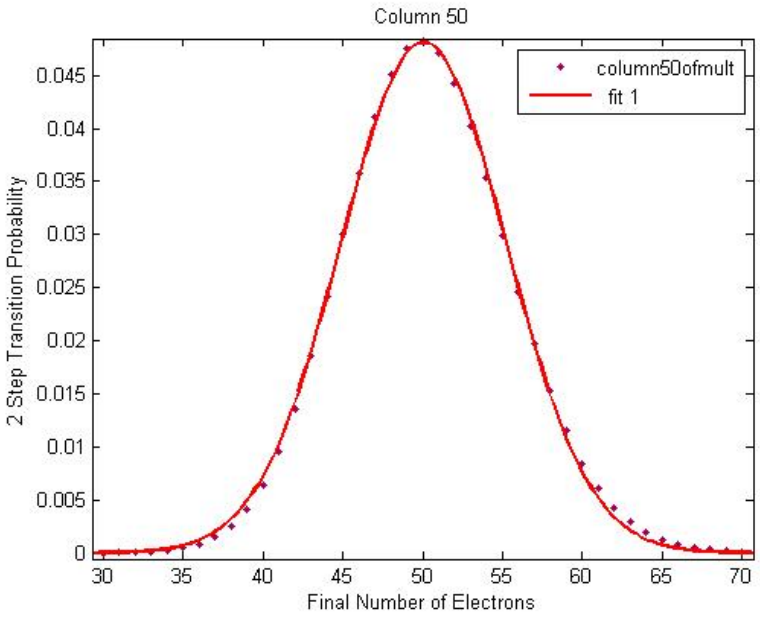

Figure 19: Gaussian Plot of Column 50 of Markov Matrix squared, representing 2 time steps, demonstrating matrix behavior down a column

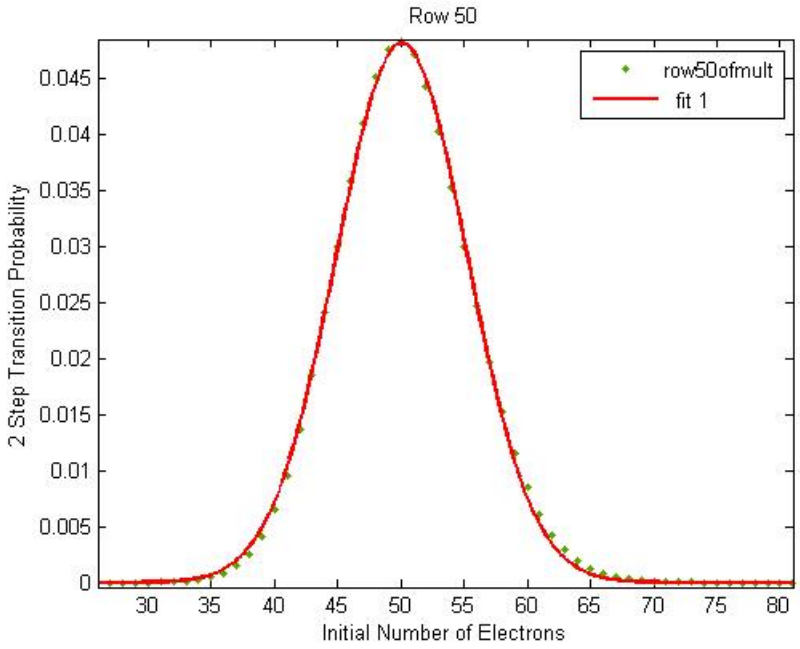

Figure 20: Gaussian Plot of Row 50 of Markov Matrix squared, representing 2 time steps, demonstrating matrix behavior across a row

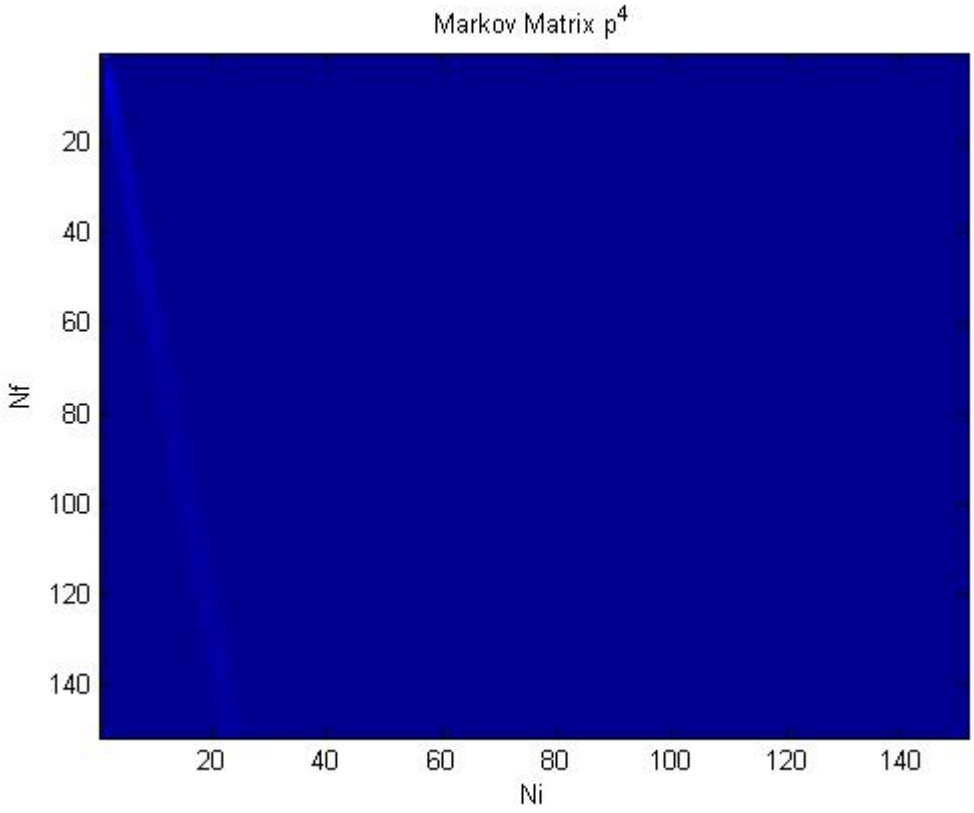

Figure 21: Heat map of Markov Matrix raised to the 4th power, representing 4 time steps

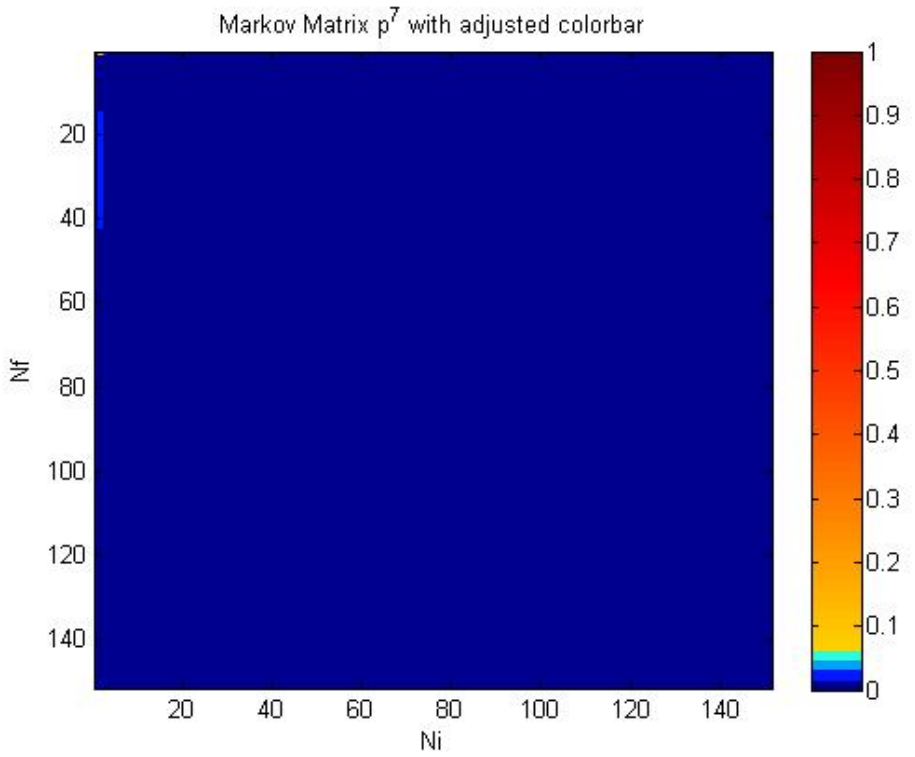

Figure 22: Heat map of Markov Matrix raised to the 7th power, representing 7 time steps, with an adjusted colorbar to better reveal matrix behavior

5 Summary and Conclusions

The avalanche statistics in a MALDI-TOF detector were examined by comparing Monte Carlo simulations with maximum entropy analysis and the use a Markov matrix approximation, both in hopes of determining a simple and accurate model for the electron avalanches occurring in the detector. Monte Carlo simulations were first performed for avalanches with a beginning number of electrons from 1 to 10, with specified absorption, stable, and creation probabilities p_0, p_1, and p_2 chosen to yield a gain of 10^6 at the end of the 30 stages in the detector. Maximum entropy analysis was performed in the hopes of finding a simple approximation for the Monte Carlo simulations. The maximum entropy calculations for avalanches beginning with 1 electron do not approximate the Monte Carlo simulations well, and though the Monte Carlo simulations show a basic exponential shape, the dip after the initial exponential decay is not present in the maximum entropy calculations. For avalanches beginning with two electrons, the peak of the maximum entropy calculations is shifted to the left, though the fit has the general shape of the simulations, but with a longer exponential decay. It can not be concluded that the maximum entropy calculations approximate the Monte Carlo simulations well. For avalanches beginning with 5 electrons, the fit has the same general shape and the average values are very close, but the peak height is too low, as is the case for avalanches beginning with 10 electrons. Thus, this method is not a good model for the Monte Carlo simulated avalanches.

Markov matrix analysis was used in the hopes of determining an equation to model the evolution of the probability distribution from one stage to the next. The Markov model was able to produce accurate heat maps, demonstrating a gain of 1.6 in each stage of the detector, demonstrated through the heat maps produced. We were also able to determine the behavior of the rows and columns of the Markov matrix, which were closely approximated to be Gaussian in shape. With further work, the demonstration of Gaussian behavior could be used to develop a model to represent the probability distribution of electrons through 30 time steps. Gaussian behavior was also observed down the columns and across the rows of a Markov matrix representing 2 time steps.

Multiplying the Markov matrix for multiple detector stages will be used to determine the behavior of the electron avalanche throughout the 30 stages of the detector. However, multiplying the matrix for multiple time steps did not reveal matrix behavior and more work needs to be done to modify the matrix to better reveal matrix behavior through more stages. The asymptotic behavior of the matrix from the heat maps could be determined in order to formulate an equation to approximate the behavior of the avalanche that could be applied to a growing number of electrons through the 30 stages. With more time, using further analysis of the Markov matrix to determine a simple approximation of the matrix could be performed. Using Stirling's approximation and reworking the equations for each matrix entry to be in terms of the only free variable, n_1, calculations were performed to determine the maximum value of n_1. In future work, the value of each matrix entry can be approximated using equation 88 to determine a simple and accurate approximation of the Markov matrix. A function that represents the probability distribution will be acquired where the probabilities must add to 1. The function at the next time step, with n_f electrons is then represented by equation (86). Accurate approximations of the functions representing the probability distribution will be found that will represent the infinite sum in a manageable manner. These functions will be used to determine the expected number of electrons at the end of an avalanche for any n_i and will be used for accurate peak height analysis.

With further work, an accurate model of the avalanche statistics in the detector can be obtained. This model will be used to understand variations in quality control spectra. After the variations in quality control spectra, which should be identical across all spectra, are understood, variations resulting from avalanche statistics can be applied to other spectra for use in biomarker analysis. Accurate peak heights are essential for proper data analysis. Biomarker analysis can be used to compare tissue or blood samples, such as by evaluating the discrepancies between spectra obtained from a cancerous sample to that from healthy tissue. By accumulating data from multiple samples, proteins occurring in diseased individuals, but not in healthy ones, can be determined, or discrepancies in the abundance of certain proteins can be analyzed. However, for rare conditions, few samples can be obtained. By understanding the avalanche dynamics of the detector, a better estimate of peak height, and therefore a more accurate protein distribution can be obtained, even for small data sets. Through accurate measurements of protein distribution, researchers will gain a greater understanding of diseases, particularly those that are rare, which can be applied to more accurately diagnose diseases, develop treatment options targeted to proteins and to develop ways to treat patients with uncommon conditions[7].

References

1 "Microchannel Plates and MCP Detectors and Imaging Systems." www.dmphotonics.com. Del Mar Photonics. 18 Mar 2009 <http://www.dmphotonics.com/MCP_MCPImageIntensifiers/ mcp_references.htm>.

2 Ashcroft, Alison E. "An Introduction to Mass Spectrometry." astbury.leeds.ac.uk. The University of Leeds. 13 Mar 2009 <http://www.astbury.leeds.ac.uk/facil/MStut/ mstutorial.htm>.

3 Chaurand, P., Norris,J. L., Cornett,D. S., Mobley,J. A., and Caprioli,R. M. "New Developments in Profiling and Imaging of Proteins from Tissue Sections by MALDI Mass Spectrometry." J. Proteome Res. 5.11 (2006): 2889-900. .

4 Cooke,William E. Personal communication to the author. 5 Oct 2008.

5 Cooke,William E., Tracy,Eugene R., Weaver,Dennis M. and Marchitelli,Robert. (unpublished).

6 Hazama, Hisanao, Aoki,Jun, Nagao,Hirofumi, Suzuki,Ren, Tashima,Toshio, Fujii,Ken-ichi, Masuda,Katsuyoshi, Awazu,Kunio, Toyoda,Michisato, and Naito,Yasuhide. "Construction of a Novel Stigmatic MALDI Imaging Mass Spectrometer." Applied Surface Science 255.4 (2008): 1257-63.

7 Petricoin, Emanuel F., and Liotta,Lance A. "Mass Spectrometry-Based Diagnostics: The Upcoming Revolution in Disease Detection." Clinical Chemistry 49.4 (2003): 533-4. <http://www.clinchem.org/cgi/reprint/49/4/533>.