Maximum Likelihood Estimation for the Improved Nanoparticle Tracking Analysis

Mihir Rajendra Athavale

Master of Science by Research

University of York

Physics

January 2022

Abstract

The Nanoparticle Tracking Analysis (NTA) is a method used to estimate the nanoparticle size distribution from their Brownian motion in suspension tracked using multiple images of the particles and the calculation of their hydrodynamic radii using the mean squared displacement data and the Einstein-Stokes equation. However, the distributions obtained by this conventional approach of NTA are usually broader because of the inability of NTA to track the particles over a large number of frames. To overcome this undesirable effect, a statistical parameter estimation method, Maximum Likelihood Estimation is implemented which takes into account the number of steps for each track along with the mean squared displacement values. To test the applicability of this method on the various particle size distributions profiles, the computer simulations are used to generate the different sets of random mean squared displacement and particle steps values by which it is possible to simulate the different experimental scenarios and obtain the different particle size distributions. The distributions obtained by the MLE are compared with the distributions obtained by the conventional method and by the use of Gaussian fitting, the distribution widths are also compared. Further, the application of this MLE method is tested on the actual experimental data obtained by the NTA system on the TiO_2 sample.

Keywords— Brownian Motion - Nanoparticle Tracking Analysis - Maximum likelihood Estimation

Acknowledgements

First of all, I wish to acknowledge my supervisor Professor Jun Yuan for his continuous support and guidance throughout the research without whom this thesis would not be possible. I would also like to thank Dr Chris Murphy and Dr Laurence Wilson for their valuable suggestions and the Physics department of the University of York for providing me with the facilities to conduct this research. Lastly, I would like to thank my parents and my friends for their love, constant support and encouragement. Special mention to my table tennis friends from York who made my time more enjoyable for the last seven-eight months after the lockdown.

Declaration

I declare that this thesis is a presentation of original work, and I am the sole author. This work has not previously been presented for an award at this, or any other university. All sources are acknowledged as References.

Table of Contents

List of Figures

List of Tables

List of Abbreviations

ENP's . . . . . Engineered Nanoparticles

SEM . . . . . Scanning Electron Microscopy

TEM . . . . . Transmission Electron Microscopy

AFM . . . . . Atomic Force Microscopy

DLS . . . . . Dynamic Light Scattering

SAXS . . . . . Small-Angled X-Ray Scattering

PTA . . . . . Particle Tracking Analysis

NTA . . . . . Nanoparticle Tracking Analysis

CCD . . . . . Charge Coupled Device

CMOS . . . . . Complementary Metal Oxide Semiconductor

MLE . . . . . Maximum Likelihood Estimation

PDF . . . . . Probability Density Function

MSD . . . . . Mean Squared Displacement

FoV . . . . . Field of View

MAP . . . . . Maximum A Posteriori

List of Constants and Symbols

α	Random variable or observations for which probability distribution can be found out
χ^2	Goodness of fit
η	Viscosity
γ	Hidden variable in the probability distribution
λ	Rate parameter
μ	Mean of the Gaussian distribution
ϕ	Transition probability
Ψ	Operator
σ	Standard deviation of the Gaussian distribution
σ^2	Variance of the Gaussian distribution
τ	Time duration between the captured frames
θ	Measurement of any quantity e.g. parameters of a pdf
θ_r	Mean of the distribution
b	Total number of bins
D	Diffusion constant
d^2	Squared displacement

E Expected frequency for the bin

$f(x,t)$ Probability density of a Brownian particle being at x at time t

H_0 Data that is known to follow a specific distribution

H_1 Data that need to be tested to see if it follows the same distribution

$H_{d^{2(*)}}$ Histogram formed by the MSD data

H_{ML} Histogram formed by the iterative solution of the MLE

j Number of iterations

k Number of steps taken by a particle for the observed tracks

k_β Boltzmann constant

L Likelihood

LL Log-likelihood

N Total number of particles

O Observed frequency for the bin

P_d Gamma PDF for an MSD value

P_m Overall particle size distribution

$P_r(r)$ Unnormalized particle radii PDF

P_z Probability of obtaining an MSD value

Q Expectation value for the log-likelihood

r	Hydrodynamic radius
T	Temperature of the solvent
u	Shape parameter for Gamma PDF
v	Rate parameter for Gamma PDF
z	Mean Squared Displacement or MSD

1 | Introduction

According to the European commission's recommendation [1], the definition of a nanomaterial is -

"A natural, incidental or manufactured material containing particles, in an unbound state or as an aggregate or as an agglomerate and where, for 50 % or more of the particles in the number size distribution, one or more external dimensions is in the size range 1 nm - 100 nm."

The nanomaterials or nanoparticles can be manufactured according to the specific role (which are called ENP's or Engineered Nanoparticles). Hence, the usage of nanomaterials spans across various industries from medicine, mechanical industries, electronics to environmental preservation, air purification and energy harvesting [2].

The bulk materials (the materials which do not have nano dimensions) have constant physical properties regardless of the size but the properties of the same materials change as their size approaches the nanoscale and as the percentage of atoms at the surface of a material becomes significant. Many physical and chemical properties of nanomaterials or nanoparticles are size-dependent for example, the colour of colloidal gold nanoparticles are dependent on their size and shape [3]. Also, many technological and industrial processes involving colloidal nanoparticles are largely dependent on particle size for example, nanoparticle-based drug delivery [4]. Therefore, the effect of particle size distribution products and processes can be critical to the success of many manufacturing businesses. The size distribution may reveal for example that there are aggregates present in the solution [5].

There is a diverse range of nanoparticle size measurement techniques available

and each technique makes use of different analytical methods for particle size measurement. For example, SAXS (Small-Angle X-ray Scattering) determines the size distribution by measuring the intensities of X-rays scattered by a sample as a function of scattering angle. Some techniques measure the actual physical size of the nanoparticle i.e. the actual material interface and some techniques measure the hydrodynamic size of the nanoparticle i.e. they measure the size of the water layer attached to the particle along with the actual material interface. Table 1.1 gives an overview of the size range of different particle sizing techniques and how different technique uses different particle sizing methods.

Particle Size Measurement Technique	**Size Range**(μm)	**Size Type**	**Type of Quantity**
Microscopy-		Length/ Shape/ Structure	Number
Scanning Electron Microscopy (SEM)	0.01-500		
Transmission Electron Microscopy (TEM)	0.001-5		
Dynamic Light Scattering (DLS)	0.005-1	Hydro-dynamic	Scatter Intensity
Nanoparticle Tracking Analysis(NTA)	0.01-1	Hydro-dynamic	Number
Atomic Force Microscopy (AFM)	0.1-8	Length / Shape / Structure	Number
Small-Angled X-Ray Scattering (SAXS)	0.003-0.3	Scattering Cross-section	Model

Table 1.1: Comparison of some of the particle sizing techniques
[6]

Although Electron Microscopy techniques especially Transmission Electron Microscopy (TEM), is considered as "Gold-Standard" for the particle size measure-

ment [7], the most common techniques for sizing colloidal particles are light scattering techniques due to their relative simplicity and the existence of many commercially available instruments.

The most widely used particle size measurement technique which uses light scattering to determine particle size distribution is Dynamic Light Scattering (DLS) or Photon Correlation Spectroscopy. It works on the principle of the Brownian motion of nanoparticles in the suspension. Brownian motion is the random movement of particles occurring due to the constant collision of particles with the solvent molecules. If you know all the other parameters such as the viscosity of the solvent and the sample temperature then you can determine the hydrodynamic radius of that particle by measuring the diffusion of the particle. The relation between these quantities is given by the Einstein-Stokes equation 1.1. A detailed explanation of the physics of the Brownian motion has been given in section 2.1.

$$D = \frac{k_\beta T}{6\pi\eta r} \tag{1.1}$$

where,

D- Diffusion constant[m^2/s]-used to calculate the diffusion speed,

k_β- Boltzmann constant[$m^2.kg/Ks^2$],

T- Temperature of the solvent[K],

η- Viscosity[$Pa.s$],

r- Hydrodynamic radius[m^2/s].

The diffusion speed of the particle is determined by measuring the rate at which the intensity of the scattered light fluctuates using an autocorrelator. The size distribution obtained is a plot of the relative intensity of light scattered by particles of different sizes and is therefore known as an intensity-weighted size distribution [9]. DLS is a very useful technique for the rapid, low-cost characterisation of

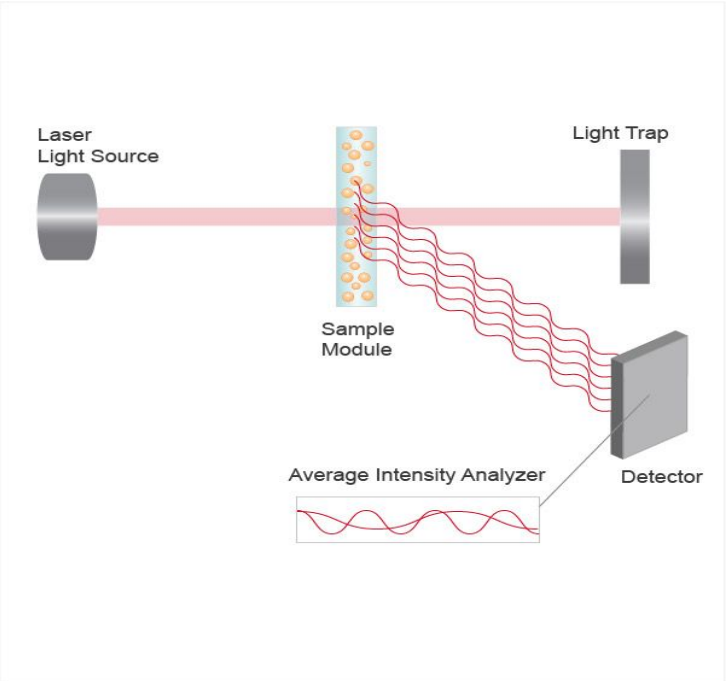


Figure 1.1: The Schematic of typical DLS experimental configuration. Diagram deduced from [8].

monodisperse and spherical particles. However, in the case of polydisperse or non-spherical particles, this technique is ill-suited because of the strong bias of the DLS technique towards the largest particles present in the sample as demonstrated by Filipe et al. [10] and Maguire et al. [11]. This bias comes from the fact that the measured particle size is dependent on the light scattered by that particular particle. This particle size dependence of light scattering is explained more clearly in the next chapter in subsection 2.2.2. Also, DLS is an ensemble technique i.e. it is not capable of single-particle tracking [12].

Particle Tracking Analysis (PTA) widely known as Nanoparticle Tracking Analysis (NTA) is a similar technique that also uses the principle of Brownian motion for particle size determination. The only difference is that in NTA light scattered by the particles are captured by using a CCD or CMOS camera and the Mean

Squared Displacement (MSD) is determined by locating the same particle in each frame of the video. (The detailed explanation of how NTA measures particle size is given in the section 2.2). NTA is the only commercial technique available for visualising and measuring the number concentration of particles directly. Due to the instrumentation and optical setup, NTA is capable of single-particle tracking and offers better size resolution than DLS for the highly polydispersed samples [10], [13]–[16]. One other advantage of NTA is unlike DLS which gives intensity-weighted size distribution, NTA gives number-weighted size distribution which is less sensitive to the variation of particle size in the distribution [17].

Despite all of these advantages, NTA has its own shortcomings such as limited concentration range, low reproducibility [10]. The focus of this research is on one particular weakness of the NTA system, the errors in the particle size distribution profile appearing due to the particle tracking approach of the system.

For the measurement of particle size, NTA calculates the particle's mean squared displacement value and as Brownian motion is a random process, it should be calculated for the very large number of frames (ideally infinite number of frames) to get an accurate estimate of the mean-squared displacement. This is not possible due to several difficulties such as the tendency of the particles to diffuse out of the camera's focal volume or the difficulty in the association of intensity peaks to the particles if two particles come very close together during tracking [18]. Hence in the current NTA, size distribution is only obtained after tracking individual particles for very small number of finite steps. As the particles are tracked for only a small number of finite steps, there will be statistical errors in each of the size estimates. This generally gives a broader size distribution estimate than actual size distribution [19], [20].

To overcome this difficulty, the alternative data processing method based on the

Maximum Likelihood Estimation given by J. Walker [18] has been explored in this thesis. In the current NTA system, to get the particle size distribution data only mean squared displacement data is used for the particle size distribution histogram. It does not make use of the other data which is readily available and can be critical in the accuracy of the size distribution, the number of steps (number of frames) for which particle is being tracked. The data processing method described in this thesis uses the Maximum Likelihood principle [21] and Walker's MLE algorithm [18] which not only makes use of the mean-squared displacement data but also the number of steps data for each particle and attempts to recover the original size distribution. The Brownian Motion simulation is done to compare the outputs of the new method with the original method as by using the simulations it was possible to imitate the different experimental scenarios more easily. Further, experimental results are also presented where this MLE algorithm is applied on the actual particle size distribution obtained from the NTA system for the TiO_2 sample.

The remainder of this thesis is organised as follows:

Chapter 2 gives the theoretical background behind the concepts appearing in the thesis such as Physics of Brownian Motion, Detailed description of the working of NTA system, Particle tracking softwares and detailed description of input parameters used, and Maximum Likelihood Estimation.

Chapter 3 describes the approach adopted in this research and the details of the data processing method used for the Maximum Likelihood Estimation.

Chapter 4 is based on the results of the comparison of different size distributions with different parameters and Brownian motion simulation results along with the size distribution profiles obtained with the maximum likelihood estimation.

Chapter 5 demonstrates the application of the MLE program to the experimental data obtained with NTA for the TiO_2 sample.

Chapter 6 describes the overall summary of the result and discusses the future scope of work in the improvement of the particle size distribution profile.

2 | Theoretical Background

2.1 Physics of Brownian Motion

The first detailed description for the Brownian motion was given by Robert Brown by studying the rapid oscillatory motion of pollen grains in aqueous suspension [22]. Following Brown's work, there were many theories about the cause of the phenomenon such as Gouy's demonstration of Brownian motion being a fundamental property of the matter before Einstein developed conclusive mathematical evidence for the random thermal motion of particles in a suspension due to diffusional motion [23]–[25]. The results that Einstein got were particularly of interest because further Smoluchowski and Langevin also managed to get the same mathematical results by different methods [26], [27]. The experimental proof of this theory was further given by Jean Perrin [28], [29].

According to Einstein, if a particle in a fluid without friction collides with another molecule randomly, then there is a change in the velocity of the molecule. But, if the fluid is very viscous, there will be a quick dissipation of the velocity and the net result will be a random change in the displacement of the particle. This process keeps repeating. So if we look at the overall motion, the particle is performing irregular motion where nothing can be predicted about the next step. The only thing we can predict is the probability of the particle covering a particular distance in time t. Because of the randomness of this motion, it is called a 'random walk' and in our case of Brownian particles, the steps of the walk are caused by molecular collisions.

Einstein further derived the expression for mean square displacement of particle utilizing Fokker-Planck equation (Fokker-Planck equation is a partial differential

equation and gives time evolution of probability density function of particle velocity under the influence of random forces [30]). To derive this equation, he made the following assumptions:

1) The motion of an individual particle is independent of all other particles.

2) Considering a large time interval, the motion of a particle at any particular instant is independent of that particle in any other instance. [23]

Einstein's method for the derivation of the mean-squared displacement expression can be given as follows:

We introduce a time interval τ which is very small if compared to the time interval t over which the whole system is observed, but large enough such that the observed particle motion during two consecutive time intervals τ can be considered as a mutually independent event as stated in the second assumption above.

Suppose there are f particles per unit volume between x and $x + dx$ at a time t (We are considering only one dimension here for simplicity. The result can be generalized to any higher dimension by projecting motion onto all one-dimensional orthogonal dimensions). After time interval τ that we have introduced earlier, we will consider exactly similar spatial element at the point x^*. We also consider that, the probability of particles entering from the neighboring spatial element is a function of only spatial distance $x^* - x$ and the time difference τ between the successive observations. Let's say this probability function is $\phi(x^* - x, \tau)$. This also includes the case if particles were in x^* if we set $x^* - x = 0$. Here, ϕ is the transition probability. Transition probability is the probability of jump from one point to another point and $f(x, t)$ is the probability density that a Brownian particle is at x at time t [31], [32]. As particles need to come from some spatial

element, the density will be given as

$$f(x^*, t+\tau) = \int_{-\infty}^{\infty} f(x,t)\phi(x^* - x, \tau)dx \tag{2.1}$$

The above equation is a special form of the Chapman-Kolmogorov equation or Master equation [31] which usually describes the time evolution of the probability for Markov processes.

Let's say the displacement is $X = x - x^*$ and the x^* is constant therefore, $dX = dx$. and then, the equation 2.1 becomes

$$f(x^*, t+\tau) = \int_{-\infty}^{\infty} f(x^* + X, t)\phi(X, \tau)dX$$

As Brownian particle has no memory, positive and negative displacements are equally likely and the function $\phi(X, \tau)$ will be even i.e. $\phi(X, \tau) = \phi(-X, \tau)$

Now suppose τ is small then we will expand left hand side in powers of τ and right-hand side in the powers of small displacement X, then the equation will be

$$f(x^*, t) + \tau\frac{\partial f}{\partial t} + ... = \int_{-\infty}^{\infty} \left[f(x^*, t) + X\frac{\partial f}{\partial x} + \frac{X^2}{2!}\frac{\partial^2 f}{\partial x^2}... \right] \phi(X, \tau)dX$$

The right hand side of the equation is then given as

$$= f(x^*, t)\int_{-\infty}^{\infty} \phi(X, \tau)dX + \frac{\partial f}{\partial x}\int_{-\infty}^{\infty} X\phi(X, \tau)dX + \frac{1}{2!}\frac{\partial^2 f}{\partial x^2}\int_{-\infty}^{\infty} X^2\phi(X, \tau)dX + ...$$

For the next step, we will use the fact that the probability function sums to unity in the first term. Also, f is independent of X, hence can be taken out of the integral over X. As ϕ is an even probability density function i.e. $\phi(X, \tau) = \phi(-X, \tau)$,

Also,

$$\int_{-\infty}^{\infty} \phi(X,\tau)dX = 1 \tag{2.2}$$

$$\int_{-\infty}^{\infty} X\phi(X,\tau)dX = 0 \tag{2.3}$$

$$\int_{-\infty}^{\infty} X^2\phi(X,\tau)d\tau = < X^2 > \tag{2.4}$$

All higher order terms such as $< X^4 >$ on the right hand side are of order τ^2 therefore,

$$\tau\frac{\partial f}{\partial t} = \frac{1}{2}\frac{\partial^2 f}{\partial x^2}\int_{-\infty}^{\infty} X^2\phi(X,\tau)d\tau \tag{2.5}$$

The integral right in equation 2.5 represents the mean-square displacement as it is the sum of the squares of the displacements each multiplied by the probability of its occurrence [31]. Therefore,

$$\frac{\partial f}{\partial t} = \frac{< X^2 >}{2\tau}\frac{\partial^2 f}{\partial x^2} \tag{2.6}$$

Equation 2.6 has the same form as the thermal diffusion equation hence we can define an effective diffusion constant *D*. Therefore,

$$\frac{\partial f}{\partial t} = D\frac{\partial^2 f}{\partial x^2} \tag{2.7}$$

where, $D = \frac{<X^2>}{2\tau}$

Equation 2.7 represents the diffusion equation in one dimension and *D* is the translational diffusion coefficient [23], [33]. The solution of this equation can be given as,

$$f(x,t) = \frac{1}{\sqrt{4\pi Dt}}\exp\left(\frac{-x^2}{4Dt}\right), -\infty < x < \infty \tag{2.8}$$

where, D is the diffusion coefficient. The detailed derivation to obtain this solu-

tion from equation 2.7 is given in Appendix A. The above equation has the form of the Gaussian PDF (equation B.3) with variance $\sigma^2 = 2Dt$ and mean $\mu = 0$.

Finally, Mean square displacement (MSD) of the particle can now be calculated with the known expression for f given in equation 2.8 and from the starting condition given for ϕ in equation 2.4 ,

$$< x^2 >= \int_{-\infty}^{\infty} x^2 f(x,t)dx = 2Dt \tag{2.9}$$

Therefore,

$$M.S.D. = z =< (x(t) - x(0))^2 >= 2Dt \tag{2.10}$$

Equation 2.10 gives the mean squared displacement of particles for one-dimensional Brownian motion.

For the case of n-dimensional Brownian motion, suppose the position co-ordinates are given by $x_1, x_2, x_3, .., x_n$. By using the product rule in probability, the n-variable PDF will be given by the products of the fundamental solution in each variable. Therefore, equation 2.8 will change into

$$f(x,t) = f(x_1,t)f(x_2,t)...f(x_n,t) = \frac{1}{\sqrt{(4\pi Dt)^n}} \exp\left(\frac{-x^2}{4Dt}\right) \tag{2.11}$$

In this case, MSD (z) will be given as

$$z =< (x_1(t) - x_1(0))^2 + (x_2(t) - x_2(0))^2 + ... + (x_n(t) - x_n(0))^2 >$$

As all co-ordinates are independent, their total displacement from the reference point is also independent. Therefore,

$$z =< (x_1(t) - x_1(0))^2 > + < (x_2(t) - x_2(0))^2 > +...+ < (x_n(t) - x_n(0))^2 >$$

Therefore, in the case of n-dimensions MSD will be given by

$$M.S.D. = z = 2nDt \tag{2.12}$$

In summary of the above derivations, the equation 2.12 gives the relation between the mean square displacement and the diffusion constant and equation 2.6 defines the probability distribution function of finding particles with a displacement x over a time period 'tau'(τ). The former is the basis of the Einstein-Stokes equation that is the key equation for the conventional NTA approach of obtaining the particle size distribution and the latter is the basis for the Monte Carlo simulation and the basis for the MLE approach.

2.2 Nanoparticle Tracking Analysis(NTA)

Particle Tracking Analysis or Nanoparticle Tracking Analysis (NTA) is one of the few systems which can be used for measuring nanoparticles in suspension as well as to visualise them. It utilizes both the properties of light scattering and Brownian motion and can be used to determine fluorescence, particle concentration as well as zeta potential (surface charge). The size range of the particles which can be measured falls between 10-1000 nm of diameter but this also depends on the light-scattering properties of the material from which the particle is formed [34]. There is a wide range of commercial NTA systems available such as Nanosight developed by Malvern Panalytical [35], ZetaView developed by Particle Metrix [36] etc. There is also a special NTA system [17] which is developed in the Department of Physics of the University of York for the low-concentration operation. This system is flexible as it is a completely open system and made from off-the-shelf components which are cheap and easily available.

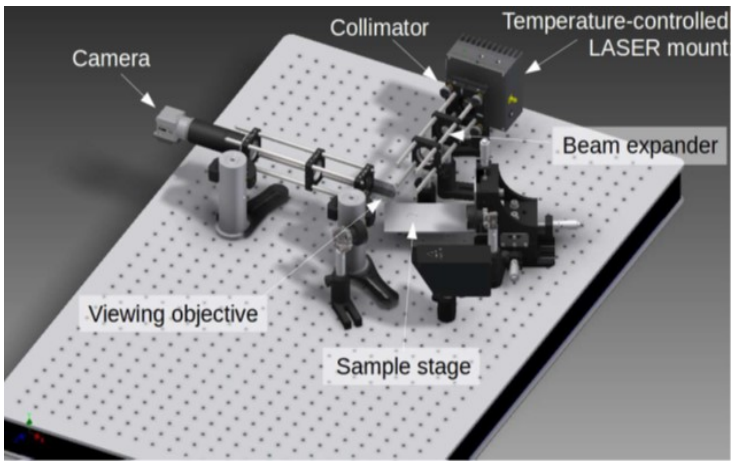


Figure 2.1: 3D rendering of the low-concentration NTA system set-up[17].

2.2.1 Principle of Operation

For the size determination, NTA uses the particle tracks of the individual particles obtained by tracking the Brownian motion of particles.

To observe the particle movements in suspension, a sample is inserted into a chamber that is illuminated by the laser. The scattered light from particles is collected through an objective lens and then focus onto a CCD or CMOS camera. Through this camera, the movement of the particles is recorded in the video. The rough schematic of this optical setup is given in Figure 2.2.

This video is analysed frame by frame and in each frame, the particle centres are located and identified through algorithms such as Crocker and Grier algorithm [38]. The identified locations of each particle are followed in each frame of the video. To make a track of any particular particle, NTA takes into account a threshold distance [6] or maximum jump distance [17] to determine if the particle tracked in the successive frame is the same particle that was tracked in the previous frame or a different particle. If a particle is identified within the threshold distance of the previously detected particle in the successive frame, then these two

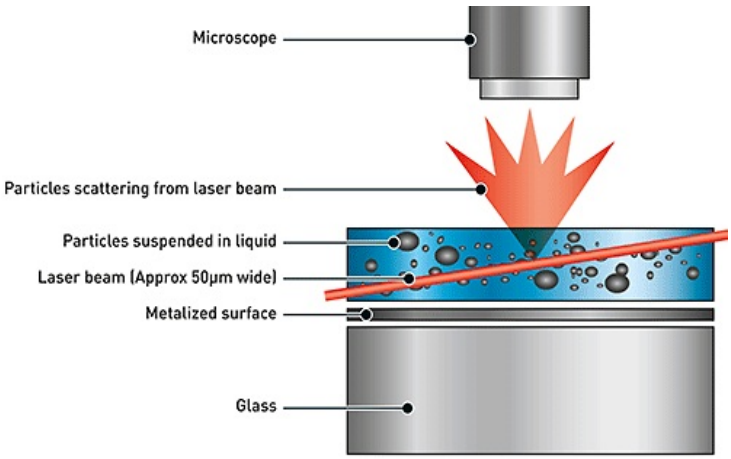


Figure 2.2: Optical set-up for a typical NTA system. The glass and metalized surface as shown in figure is not necessary in general [37].

particles are recognised as the same particle and a particle track is made. This process is done for all the successive frames in the video which then determines the number of jumps for any particular particle. The track for any particular particle terminates if there are no particles in the successive frame or two or more particles are detected within the threshold distance. From these tracks, a mean squared displacement of particles in two dimensions is calculated. (As the recorded video will be in 2-D). If we put n=2 in the equation 2.12, this calculated MSD can then be converted into diffusion constant by the formula

$$z = \overline{(x, y)^2} = 4D\tau \tag{2.13}$$

where x and y are the coordinates of a particle in any particular frame and τ is the duration of time between the frames.

From the calculated MSD of a particle and therefore essentially from the diffusion constant (equation 2.7) D, it is possible to calculate the individual particle size by using the Einstein-Stokes equation (1.1) which is obtained in terms of molecular

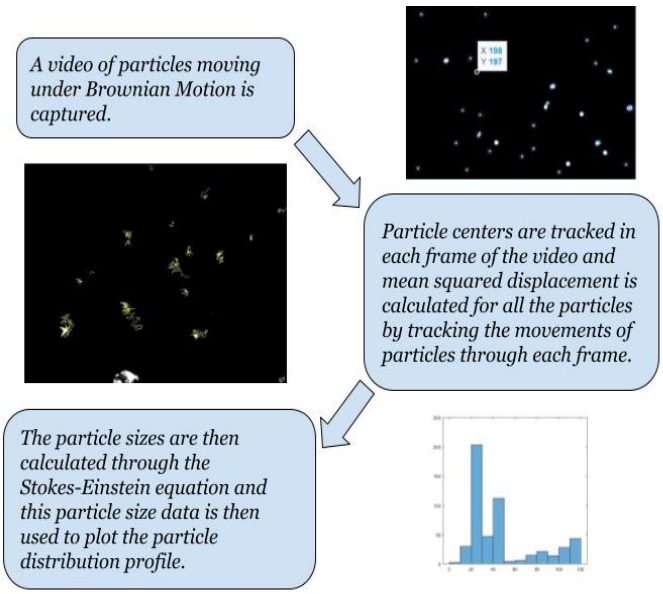


Figure 2.3: The step by step process of how NTA measures particle size where first image on the top-right shows the detected particle-like structures and their co-ordinates, second image on the center-left shows the particle tracks [17] and third image on the bottom-right shows the particle distribution profile of Frequency vs particle radius.

quantities by substituting *D* into equation 2.8 and given as,

$$D = \frac{k_\beta T}{6\pi\eta r} \tag{2.14}$$

From equation 2.14, equation 2.13 can be written as

$$\overline{(x,y)^2} = z = \frac{4k_\beta T\tau}{6\pi\eta r} \tag{2.15}$$

Rearranging the above equation for the hydrodynamic radius *r*,

$$r = \frac{4k_\beta T\tau}{6\pi\eta z} \tag{2.16}$$

These calculated particle sizes are then plotted as a histogram of radii to get the particle distribution profile. Figure 2.3 gives the pictorial representation of this

process.

2.2.2 How NTA measures different properties

Along with determining particle size, NTA is also capable of determining particle concentration, intensity and fluorescence detection.

Concentration Particle concentration can be easily calculated by the NTA system as NTA is a microscope-based technique. By the known magnification value and therefore the observable area or the field of view of the particle images, first, the particle centres in the field of view are counted. By also estimating the depth of field, the volume is determined within which particles are being tracked. The particle concentration can be determined by dividing the number of particle centres by the volume which is usually given in cm^{-3} or mL^{-1}.

Fluorescence If the sample of nanoparticles contains a sub-population of nanoparticles that absorbs the light and emit a longer wavelength, it is possible to detect those particles by insertion of a suitable optical filter in the imaging channel. This filter can be used to reject light scattered by the particles of the same wavelength and higher wavelengths can be selectively passed. By fluorescent labelling methods, it is possible to selectively identify and measure the size and concentration of certain particles.

Intensity It is also possible to obtain information about nanoparticles by considering the 'brightness' of particles. Particles are brighter if they are larger as the intensity of the scattered light is proportional to the sixth power of the diameter for the homogeneous, spherical particles [10], [39]. Also if the particles have a higher refractive index then the particles also have larger intensity values. For example, Gold nanoparticles can have higher intensity values than polystyrene although they have the same diameter [34]. Therefore, it is possible to differentiate

between two sample populations by monitoring their intensity. Although this is possible for homogeneous, spherical particles, apparent intensity fluctuates with the particle orientation. For example, for the dielectric particles, the scattered intensity will be proportional to the square of the optical thickness [40].

2.2.3 Notable limitations and the efforts done to minimize them

1. The main challenge in the optical setup is to illuminate the sample with a high-intensity light so that the low levels of the scattered light from the individual particles can be detected and to minimize the stray light scattered by the interfaces such as the interface between liquid and cuvette wall. This stray light adds to the optical noise of the system and detection of the weakly scattering particle becomes difficult. This also results in the difficulty of doing low concentration measurements. Due to this issue, the currently available commercial NTA systems can only be able to offer the concentration range of around $10^7\ particles/mL$ to $10^9\ particles/mL$ [10]. The NTA system developed at the University of York tried to overcome this issue by improving the field of view for the particles and fixing the long working distance so that lens will be able to focus the laser beam through the cuvette wall. This system can offer the concentration range of $10^5\ particles/mL$ to $10^8\ particles/mL$ [17].

2. The Einstein-Stokes equation 1.1 gives valid results only if the shape of the particle is spherical. However, there are different shaped nanoparticle available and if the particle shape is not sphere the diffusion shape become invalid. There has already been some work done around this by studying the shape of the intensity distributions for different particles and determining

the shape of the particle whether it is a rod, disc or a sphere [17].

3. Brownian motion is a random process. Therefore, the Einstein-Stokes equation is valid only when a particle is tracked for infinite number of frames. This is practically not possible and therefore, there are errors in the estimation of particle size distribution e.g. Distribution may appear broader than the actual particle size distribution [19], [20]. This issue is addressed in this thesis.

2.2.4 Particle Tracking Softwares

There is a range of freely available particle tracking software packages such as Trackpy [41], TrackMate [42], HybTrack [43] etc. which have their own approach of making particle tracking as accurate as possible. The main program used for the analysis of the particle videos for the research is 'Sam's Particle Tracking Program' [17]. This program is MATLAB based and developed by Samuel Thompson, a former PhD student at the University of York. This is also the main program in use for the Nanoparticle Tracking Analysis system developed for low-concentration operation. This program has several notable features that can be used to remove noises getting added into particle size distribution profiles due to the shortcomings of the experimental setup. Details of some of these notable features are mentioned below-

- **Global Size Parameter -** This parameter determines the particle neighbourhood. By defining this parameter we are defining the minimum distance between particles for software to determine them to be different particles rather than just part of the same particle. Global size parameter and associated algorithm can be used to remove the effect of 'out-of-focus' images or the finite size of particles, which makes identification of particles by peak

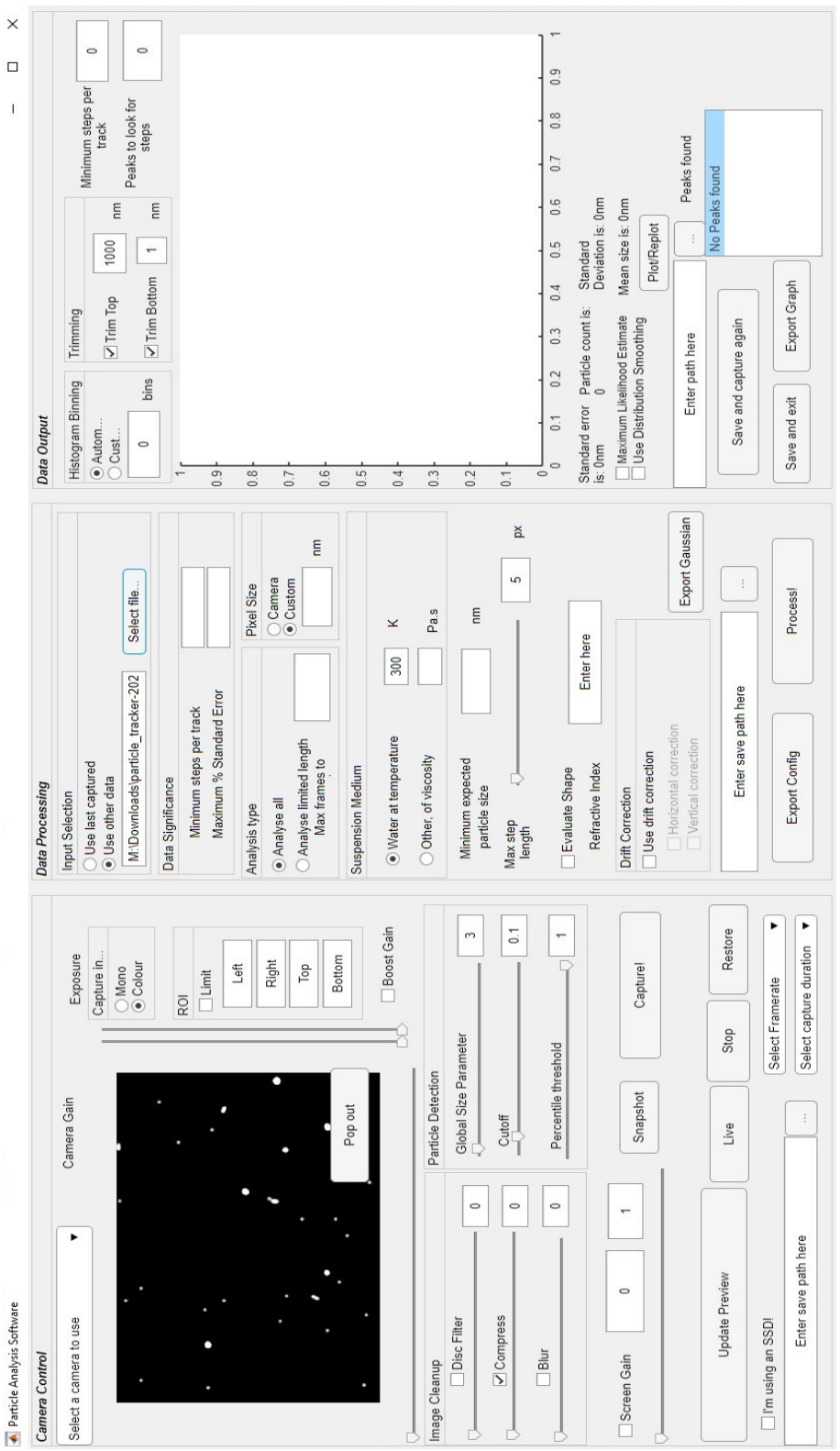


Figure 2.4: Particle Analysis software GUI

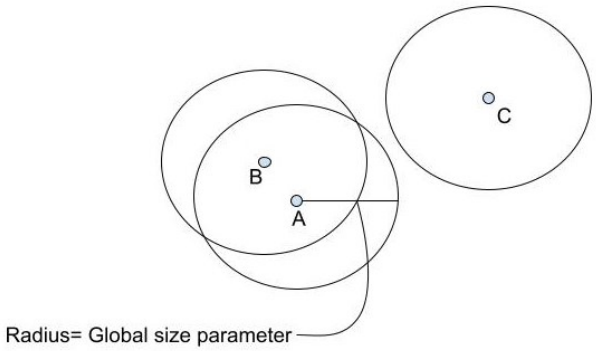


Figure 2.5: A representation of the Global Size Parameter where A,B and C's are detected particle-like structures and the radius of the circles represents the Global Size Parameter.

location algorithm unreliable. For example, extended images of large particles may have multiple local maxima that can be counted as two separate particles. Alternatively, the out-of-focus images of nanoparticles may have ripples or rings associated with the central maximum.

In the figure 2.5, A, B and C are the detected particles in the analysis. When we define the global size parameter (in pixels), we are essentially defining the circle which has a radius of the given value of the global size parameter (in pixels) with a particle at the centre of the circle. Consider area of the circles are A', B' and C'. In the figure 2.5, particles A and B are considered as a single particle as their positions are within the area of the circles A' and B'and C will be considered as different particles from A and B as their position is outside of the circle A' and B'. If the software detects a centroid larger than the defined global size parameter, it will neglect to track that particle.

- **Minimum steps per track-** These are the smallest number of steps any particle must take before it can be included in the list of particles that is used for the particle size distribution plots. Consider a particle is at a certain

position having the coordinates (x_0, y_0) in the initial frame and then if we detect the position of that particular particle in a successive frame which is let's say (x_1, y_1). This displacement of particle from (x_0, y_0) to (x_1, y_1) will be called as one jump or step.

Therefore, we can only determine the number of steps that particle has taken after we detect the position of that particle in the next frame. In conclusion, if a particle is tracked for l number of frames, then it has taken $l - 1$ number of jumps.

- **Max step length-** This parameter determines the maximum number of steps the particle can jump between the frames (in pixels) before we can classify it as a newly detected particle. For example, if we set the value of this as 3 pixels then the particles that jumped more than 3 pixels from their original positions in the next frame are considered as new particles. This value must be kept the same or smaller than the global size parameter to prevent the oscillation between two detected particles.

- **Minimum expected particle size-** This feature is used to remove particles smaller than the value of the input. This feature can be used to remove the noise (sub-nanometer particles).

- **Minimum Percent Standard Error-** This feature is used for the discrimination of particles based on the step length distribution. Standard Error is a similar term to the standard deviation which uses sample data and gives the spread of the distribution. Therefore, by using this feature, the uncertainty in the measurement of sphere-equivalent hydrodynamic radius greater than the inputted value can be discarded. But care should be taken while inputting the value for this because if the value is very small (around 1 or 2) then an almost negligible number of particles get detected and then it is not

possible to get the particle size distribution. So usually it needs to be set very largely.

- **Drift Correction-** This option examines the mean trajectory of all the particles to detect and remove the drift which can appear in the suspending liquid due to the intentional flow-through pumping or intentional/unintentional convection. This feature works on the drift-correction algorithm which determines the time-dependent group velocity of the detected particles and then generates an apparent mean flow vector across the Field of View. This calculated mean flow vector is then subtracted from the trajectories on a frame by frame basis. However, if the drift is spatially heterogeneous this subtraction is not appropriate. In this case, the subtraction can be done by limiting the analysis to motion orthogonal to the suspected drift direction [44], [45].

  In Chapter 5, where the drift correction is applied for the TiO_2 videos, every video is carefully checked for the variable drifts and no variability is found visibly. But there still can be errors in the drift correction calculations which can result in erroneous size distribution.

2.3 Maximum Likelihood Estimation in Statistics

2.3.1 Definition and Intuitive Example

"An estimator is a procedure applied to the data sample which gives a numerical value for a property of the parent population or as appropriate a property or parameter of the parent distribution function." [46]

The density estimation involves selecting a probability distribution function and the parameters that best fits the given data. This can be done mainly in two ways

[47]:

1. MLE or Maximum Likelihood Estimation [21] which does not use any prior information and the convergence of MLE solutions are possible equally likely for all the data points.

2. MAP or Maximum A Posteriori which takes prior information into account such as prior belief or guess about the possible distribution.

Let's consider a set of independent observations $\alpha_1, \alpha_2, ..., \alpha_N$, the joint probability distribution function of α or let's call it data, by independence, can be given as

$$P(data|\theta) = P(\alpha_1, ..., \alpha_N; \theta) = \prod_{i=1}^{N} P(\alpha_i|\theta) \tag{2.17}$$

(Product rule in Probability) [48] where, θ is a measurement of quantity or in case of PDF can be parameters of PDF.

The above product of probabilities in equation 2.17, gives us the conditional probability or the likelihood of observing the 'data' given the condition 'theta (θ)'.

$$L(\theta) = P(data|\theta) \tag{2.18}$$

where, $L(\theta)$ is called as a likelihood function for θ. [46], [48]

The principle of Maximum Likelihood estimation is to maximize the likelihood function given above given the set of parameters (θ) to predict the model that will fit best to the given data. So essentially the maximum likelihood estimation can be given as

$$\theta_{MLE} = arg\max P(data|\theta) \tag{2.19}$$

For example, if we have the probability density of observing a single data point

α, that is generated from a Gaussian distribution

$$P(\alpha; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(\alpha - \mu)^2}{2\sigma^2}\right) \tag{2.20}$$

and suppose if we have two single data point values of α let's say 1 and 2, we will just multiply the individual probability densities with substituting values of α to get the joint probability distribution which will be given as

$$P(1, 2; \mu, \sigma) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(1 - \mu)^2}{2\sigma^2}\right) * \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{(2 - \mu)^2}{2\sigma^2}\right)$$

We will figure out the values of parameters μ and σ which will maximize the likelihood of observing these parameters.

$$L(\mu, \sigma) = P(data|\mu, \sigma) \tag{2.21}$$

We can do this by differentiation with respect to μ and σ respectively. But if we consider the above example, we can see that it is a quite tedious job to differentiate this equation. It can get simplified to work with if we take the natural logarithm of the expression. This is possible and does not change the original form because the natural logarithm is a monotonically increasing function. This means that if the value on the x-axis increases, the value on the y-axis also increases. So even if we take the natural logarithm, we are still ensuring that the maximum value of the log of the probability occurs at the same point as the original probability function. Therefore it's better to work with the simpler log-likelihood instead of the original likelihood. Therefore,

$$\theta_{MLE} = arg\max\left[logP(data|\theta)\right] \tag{2.22}$$

For our example,

$$\ln P(1,2;\mu,\sigma) = \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(1-\mu)^2}{2\sigma^2} + \ln\left(\frac{1}{\sigma\sqrt{2\pi}}\right) - \frac{(2-\mu)^2}{2\sigma^2}$$

Again simplifying this equation using logarithmic properties,

$$\ln P(1,2;\mu,\sigma) = -2ln\sigma - ln(2\pi) - \frac{1}{2\sigma^2}\left[(1-\mu)^2 + (2-\mu)^2\right]$$

So if we differentiate the above equation with μ and σ we will get the best estimate for the distribution which can be fit for the given data points 1 and 2 with these parameters. For example, for mean μ for this example,

$$\frac{\partial \ln P(1,2;\mu,\sigma)}{\partial\mu} = \frac{1}{\sigma^2}\left[(1+2-2\mu\right]$$

Setting left hand side to zero and rearranging gives us,

$$\mu = \frac{1+2}{2} = 1.5$$

So this will be the maximum likelihood estimate for μ or mean of the Gaussian distribution. Similarly, it can be applied to σ and then we can find two parameters of the distribution which gives the best estimate or fit for the given data points.

The example given illustrates how MLE can be used to fit a Gaussian model to the observable variables $\alpha_1, \alpha_2, ...\alpha$ by finding out the model parameters μ and σ. Alternatively, the same optimization algorithm can be used to find the so-called 'hidden variables' of the model. The 'hidden variables' can mean α_j's that belong to that observation α but were missing or not detected. The hidden variables can also mean other variables γ that determines the observables $\alpha(\gamma)$ [49].

2.3.2 Application of MLE to the Particle Size Distribution obtained by NTA

As described earlier, MLE is used for the density estimation so, this method can be used to get the accurate particle size distribution for Nanoparticle Tracking Analysis by determining the frequencies of the particle sizes obtained from the NTA and by assuming that the true particle sizes are distributed over a finite set of values [18], [19]. For the application of MLE, first, the apparent sizes of the individual particles are determined using the MSD values and the Einstein-Stokes relation. (This can be done by using actual experimental data or the Monte-Carlo simulation explained in Chapter 3). Then, the sets of bins are established for the true sizes and for the virtual particles, initial values of frequencies with suitably fixed sizes are chosen. Here, virtual particles meaning we are not considering the actual data for particles, but considering some initial distribution of particles (uniform distribution in our case of MLE as we are considering convergence is equally likely possible for all the radii values) as an initial estimate or as a starting point to run the Maximum likelihood Estimation. Further, the probabilities $(p_1, p_2, p_3, ..)$ of obtaining the individual particles are calculated from this size distribution of the virtual particles. These probabilities are calculated from the values of MSD and the number of steps for each particle. After having all these data, first like explained in equation 2.17 the joint probability$(p_1 \times p_2 \times p_3 \times .....)$ is established where data will be apparent sizes of actual individual particles and θ will be frequencies of the virtual particles. And finally, to get the size distribution the frequencies of virtual particles are changed till the joint probability of obtaining this set of apparent sizes $P(data|\theta)$ or in our case $(p_1, p_2, p_3, ..)$ becomes maximum. The flowchart of the whole process has been given in the figure and the detailed algorithm for this process is explained in chapter 3.

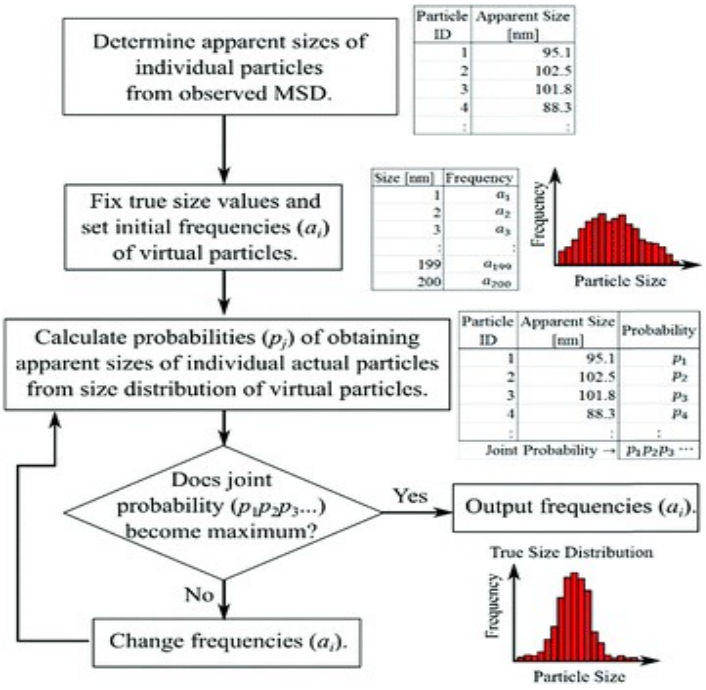


Figure 2.6: The flowchart of the MLE method for determining the improved particle size distribution obtained from the NTA[19].

2.3.3 Advantages and Disadvantages of using MLE

The principle of Maximum Likelihood (ML) estimator is to estimate the most likely value and it has some advantages such as no loss of information and all the experimental data are used and it is very suitable for the problems where multiple variables need to be estimated [46]. ML estimators are usually consistent meaning as sample sizes grow larger, ML estimate approaches quickly to the original population parameter. This has been shown in the section 4.2 where, as we increase the mean step size gradually the ML estimator goes quickly to the original population parameters. So for the large number of samples MLE can be the best estimator but for small samples, this is not necessarily true which has also been demonstrated in the section 4.2 where for the small number of step sizes ML estimator was unable to estimate the actual size distribution.

MLE can also be biased sometimes. Bias is a term used in statistics to define the tendency of an estimator to overestimate or underestimate the parameter. This has also been shown further in the results section 4.3 where MLE shows the bias towards the radius value which has the highest occurrence probability and therefore overfits the size distribution.

2.4 EM Algorithm

Maximum Likelihood Estimation involves density estimation or an optimization and search problem, where we search for a set of parameters that will be the best fit for the joint probability for the given data sample. But there is one weakness involved with this MLE approach that it assumes that the data set is complete. This is necessarily not the case every time as sometimes only some of the relevant variables are observed or some variables can remain hidden although they influence other random variables [50], [51]. These unobserved or hidden variables are called latent variables. Apart from the actual missing or unobserved variables or data, the latent variable term also applies to the situation where incompleteness is not evident. For example, a statistical model involving random effects [49]. The Expectation-Maximization algorithm (EM algorithm) is an iterative method to find an MLE or MAP estimate for the models with latent variables.

The method consists of two steps:

1. **Expectation Step -** Computes the expected value for the Likelihood $L(\theta; \alpha, \gamma)$ where, θ is the unknown parameter vector and α and γ are observed and missing dataset respectively.

2. **Maximization Step -** Maximizes the parameters of the model with the present data. The parameters of the model can include the unknown par-

ticle radius if the means or variance of the model is known.

2.4.1 Basic Idea

Let α_i be the observed variables and γ_i be the missing or hidden variables for the case i. The log-likelihood of the observed data is given as

$$L(\theta) = \sum_i logP(\alpha_i|\theta) \tag{2.23}$$

By considering the hidden variable γ_i above equation will turn into following by marginalizing over γ,

$$L(\theta) = \sum_i logP(\alpha_i, \gamma_i|\theta) \tag{2.24}$$

Now for the EM algorithm,

1. **E-Step-** The E-step computes the expected value of $L(\theta)$ given the observed data α_i and the current parameter estimate θ_{old}. For this step, we will first assume the conditional probability $P(\gamma_i|\alpha_i, \theta_{old})$ and the auxiliary Q function is defined which is formed by the expectation value of the log-likelihood,

$$Q(\theta, \theta_{old}) = E_{\gamma_i|\alpha_i,\theta_{old}}\left[L(\theta)\right] \tag{2.25}$$

Considering the conditional probability, the above equation can also be writ-

ten as

$$Q(\theta, \theta_{old}) = \sum_i P(\gamma_i|\alpha_i, \theta_{old}) P(\alpha_i, \gamma_i|\theta) \tag{2.26}$$

2. **M-Step-** The M step consists of maximizing over θ the expectation computed in the previous step which will be given as

$$\theta_{new} = \max Q(\theta, \theta_{old}) \tag{2.27}$$

We then set $\theta_{old} = \theta_{new}$ and these two steps were repeated until the sequence of θ_{new} converges [52].

In summary, the EM algorithm maintains an estimate of the parameter of θ that is updated on each iteration. If 'j' is the current iteration number, for the E-step, a function of θ called the Q-function $Q_j(\theta)$ is first formulated and then, on the next M-step, the algorithm assigns θ_{j+1} that will maximize the current $Q_j(\theta)$.

In our case of the MLE for particle size distribution, the E step is for the calculation of the expectation value(or the means) of the current estimate of the particle size distribution and the M step is for maximization which generates a new estimate of the particle size distribution.

2.4.2 Convergence of EM algorithm

As mentioned earlier, EM algorithm maximizes $L(\theta)$ by the iterative procedure. We are computing an updated estimate of θ such that,

$$L(\theta) > L(\theta_j)$$

where, θ_j is the current estimate for j^{th} iteration. Essentially we are maximizing the difference $L(\theta) - L(\theta_j)$. Let's say,

$$L(\theta) - L(\theta_j) \triangleq \Delta(\theta|\theta_j) \tag{2.28}$$

Equivalently we can write,

$$L(\theta) \geq L(\theta_j) + \Delta(\theta|\theta_j) \tag{2.29}$$

θ_{j+1} is the estimate for θ which maximizes $\Delta(\theta|\theta_j)$ and for the current estimate θ_j we had $\Delta(\theta_j|\theta_j) = 0$ [53]. As θ_{j+1} has chosen to maximize $\Delta(\theta|\theta_j)$ we will have,

$$\Delta(\theta_{j+1}|\theta_j) \geq \Delta(\theta_j|\theta_j) = 0 \tag{2.30}$$

Therefore, from the above equation likelihood $L(\theta)$ is non-decreasing. For convenience we will define $l(\theta|\theta_j)$ such that,

$$l(\theta|\theta_j) \triangleq L(\theta_j) + \Delta(\theta|\theta_j) \tag{2.31}$$

Therefore, from equation 2.29,

$$L(\theta) \geq l(\theta|\theta_j) \tag{2.32}$$

Therefore, $l(\theta|\theta_j)$ is bounded by the likelihood $L(\theta)$. Also for the current iteration

$$l(\theta_j|\theta_j) = L(\theta_j) + \Delta(\theta_j|\theta_j) = L(\theta_j) \tag{2.33}$$

Therefore, from equation 2.32 and 2.33, it can be concluded that any θ which

increases $l(\theta|\theta_j)$, increases $L(\theta)$.

Now, if L and l are equal at θ_j and also differentiable, then we can safely say that θ_j must be the stationary point of L which proves the general convergence of the EM algorithm. But the stationary point is not always the local maximum and also in exceptional cases sometimes can converge to local minima or saddle points [54]. This is often not the problem if the likelihood function is convex, but for the non-convex likelihood function, it can be an issue. This can be checked by reinitializing the EM starting point multiple times and choosing the convergent point that has the highest likelihood value [49], [55].

2.5 Hypothesis Testing with Chi-Squared Test

In the experimental studies, there are generally two groups where researchers do some process on one group such as giving a drug to the group of animals and comparing the results with the other group where this process is not applied. The results are compared and if the results are different then the researcher checks if it is possible to say with certainty if the difference is large enough to say there is a systematic difference in the group as opposed to the group where this process is not applied. This is called Hypothesis Testing [56].

In statistics, the goodness-of-fit test is used to compare if your sample data is comparable to the actual population. There are mainly 3 goodness of fit tests i.e. The chi-square which is used for the discrete distributions, Kolmogorov-Smirnov and Anderson-Darling.

The Chi-Squared test is the test that can be only used for the data that can be put into classes or bins of Histogram. The other two tests are for continuous distributions. The Chi-squared test is defined as-

$\mathbf{H_0}$- Data that's known to follow a specific distribution,

$\mathbf{H_1}$- Data that need to be tested to see if it follows the same distribution

To test the goodness-of-fit, if the data is divided into certain b number of bins then,

$$\chi^2 = \sum_{i=1}^{b} (O_i - E_i)^2/(E_i) \tag{2.34}$$

[57] where, O_i is the observed frequency for the bin i and E_i is the expected frequency for bin i. The main disadvantages of these tests are that it depends on how the data is binned and it is very sensitive to the sample size [58], [59].

3 | Methodology

This chapter provides the research strategy used to deal with the research problem and further provides the data processing method used for the Maximum Likelihood Estimation of the particle size distribution. The results obtained from this approach are presented in the next chapter.

3.1 Research Approach

As mentioned earlier, the present strategy of forming particle size distribution is to get the values of the individual particle radius from the Diffusion constant values and form histogram or size distribution of different particle sizes and this method lacks accuracy. To make the improvements in the particle size distribution obtained by NTA, it was essential to study how the present NTA software obtains particle size data and the present approach of obtaining the particle size distribution from the NTA system. To get the particle size distribution from the videos obtained by the low-concentration NTA system, Sam's Particle Analysis program, the software developed in MATLAB was chosen for ease of accessibility and availability of the different analysis parameters to make the particle size distribution as accurate as possible by the present histogram method. As the particle tracking software was made in the University and not yet fully tested for the commercial application, the first steps were to find out the sources of errors or bugs in the software such as tidying up the Graphical User Interface part of the software by removing distortions so that the user interface will look clean, improving on the Data Visualisation part of the software by adding position coordinates of the identified particles in the particle display window and making changes in the App environment from 'Guide' to 'App Designer' so that the software will be

future proof. Further analysis of the software was done by studying the Crocker and Grier algorithm which is used for the detection of particles from the video.

After testing out the bugs in the software, the matrices produced by the software were studied which contains the particle size data. It was observed that some particle data were giving infinity values which were producing errors for the particle size histogram. These infinity values were appearing because some particles were giving zero values for the Diffusion constant as they were just tracked for just one frame as particles can go out of the frame of the video after some time. These errors were removed by removing these zero values from the particle size matrix which is used to plot the final particle size distribution profile.

A systematic investigation was carried out further on a sample video to analyze the impact of the different analysis parameters on the particle size distribution profile. To study this impact, basic parameters such as temperature, the viscosity of the liquid, pixel size and minimum expected particle size were kept constant throughout the analysis. The impact of the other analysis parameters such as max step length, global size parameter and minimum steps per track was studied by varying values of these parameters and obtaining particle size distribution along with obtaining mean size, standard deviation and standard error for the distribution. The description of these parameters is already given in subsection 2.2.4. Different statistical methods for the improvement in the particle size distribution were reviewed further such as Walker's iterative maximum likelihood algorithm for the NTA system [18], Bayesian inference method for Dynamic Light Scattering data given by Naimm et al. [60] and Matsuura and co-workers [19]. Further Maximum A Posteriori Nanoparticle Tracking Analysis or MApNTA is given by Silmore et al. [20] was studied extensively.

Before applying any algorithm to the experimental data it was necessary to test the

algorithm with different conditions as very little experimental data was available and with the current experimental set-up, any particular particle was tracked only for a few number of frames or steps. Therefore, the simulation for the Brownian motion of particles was developed to simulate the similar experimental conditions of NTA. To make the improvements in the size distribution, Walker's iterative Maximum Likelihood Algorithm [18] or EM-type algorithm was chosen to convert the current version of the NTA program into the Maximum Likelihood Estimation program. Further, the application of this program is tested on the actual experimental data. The distribution obtained from this type of algorithm can be used for further improvement with the Maximum A posteriori Estimation but currently, only the MLE algorithm is applied in the program.

3.2 Brownian Motion Simulation

For the testing of the modified size distribution approach and the conventional approach used in the current version of the NTA program, the Brownian Motion of particles was simulated in MATLAB and particle tracks are made according to the NTA particle detection principle to generate the well-defined dataset. By using this simulation, it was possible to control and simulate different experimental conditions such as monodispersed/polydispersed samples, a solvent of different viscosities or different temperature conditions. It was also possible to control the number of particles present or the mean and standard deviation for the number of tracks for different particles (To mimic the actual experimental conditions, the program generates a random number of values around a mean step number, with the probability of the distribution adjustable with a step number variance parameter for each particle).

The simulation takes the values of input to form an original particle size distri-

bution for the sample where you first need to define whether you expect uniform sized particles (monodispersed) to be present in the solvent or the particles of varied sizes are expected (bidispersed or polydispersed). You also need to input the value for the mean number of tracks and expected uncertainty in the number of tracks along with the expected number of particles present in the solvent. The simulation automatically considers that the particles are present in the water and particle size measurements are taking place at room temperature but this can also be changed if we change the viscosity of the solvent and temperature in the Einstein-Stokes formula (equation 1.1).

After giving the input values for the above parameters, the simulation then selects random sizes from the allowed particle sizes defined in the originally defined distribution for the given number of particles. Further, the Diffusion constant is calculated from the Einstein-Stokes equation from each particle. To get the random mean squared displacement values from this data, the normally distributed random number of 2-D coordinates of particles are generated and scaled with the variance of the Gaussian PDF (as mentioned in section 2.1 equation 2.8) of $\sqrt{2D_i\tau}$ where D_i represents calculated diffusion constant for each particle and τ represents the time interval between each displacement or captured frames which is set at 100 milliseconds which is an average value for the time interval between the captured frames of the CCD camera.

After getting these coordinate values of the particles along the tracks simulated for each particle, the simulation calculates mean-squared displacement values from these coordinates for each particle. These randomly generated scaled coordinates can mimic the data obtained from the actual experimental videos as this part is similar to the NTA set-up where NTA determines the particle coordinates from each captured frame and then particle tracks are generated from these coordinates to obtain the mean-squared displacement values. MSD is calculated by using the

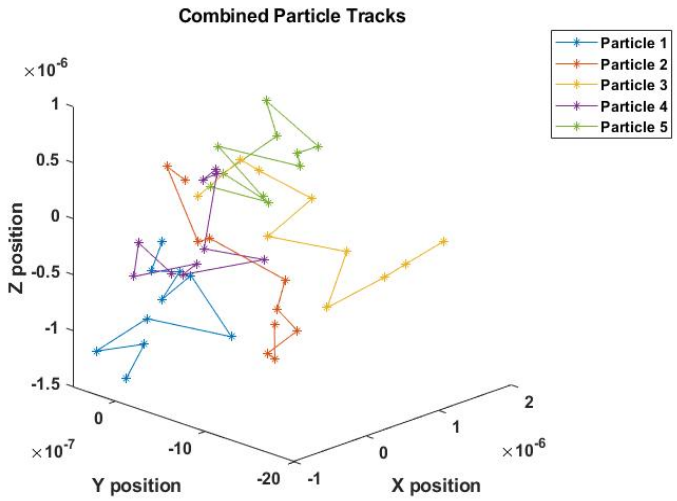


Figure 3.1: The figure shows the 3D random walk for the five Brownian particles with the mean number of steps 10 and with no variance in the count of steps for any particle meaning each particle has the same number of steps.

formula:

$$M.S.D. = z =< |(r(t) - r(0))^2| > \tag{3.1}$$

where, $r(t)$ is the position of the particle at time t and $r(0)$ is the initial position.

These mean-squared values are already scaled with the diffusion constant values so converting back these values into the diffusion constant and then the particle radii will be similar to the conventional approach of the particle size distribution of NTA. The MSD values are converted into the particle radii values by using the equation 2.16 to get the Einstein-Stokes recovered distribution. This is the direct derivation of particle radius from the estimate of the means from a finite number of steps and is subject to statistical uncertainty. The smaller the number of steps, the statistical uncertainty would be larger. To overcome this difference, one can get the Maximum Likelihood estimate to take the statistical uncertainty into account. The data for the mean number of steps and mean squared displacement for

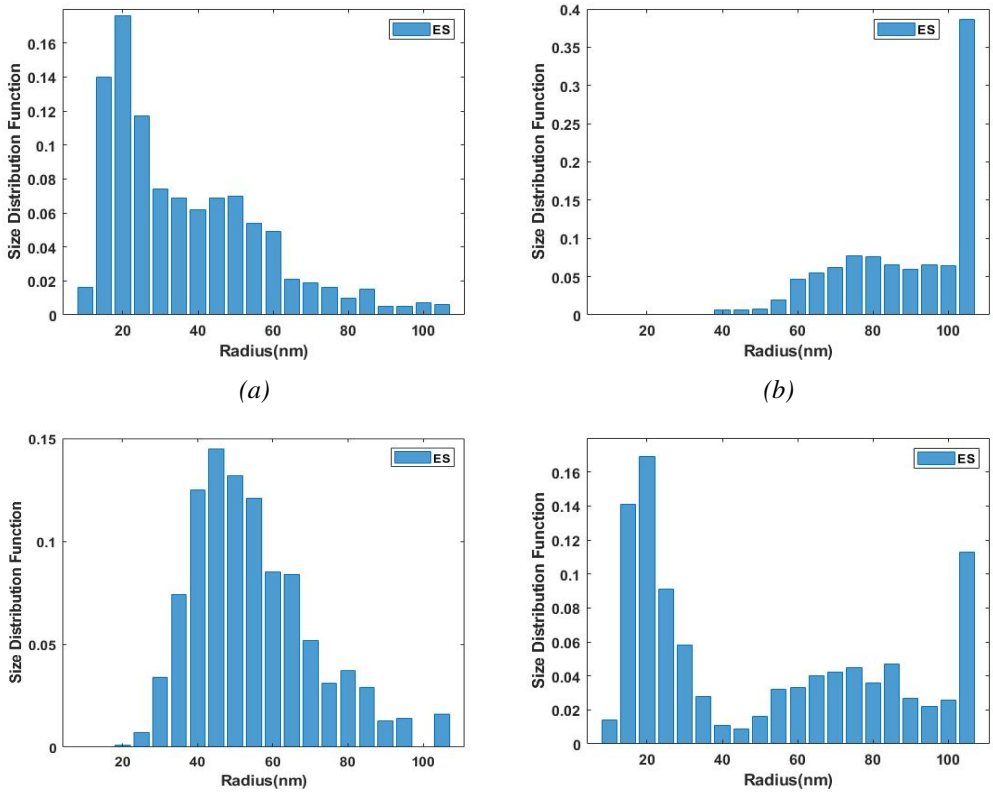


Figure 3.2: The figure shows the different types of Einstein-Stokes distributions obtained by the Brownian motion simulation for the input parameters of 1000 number of particles, 10 mean number of steps and the diffusion constant is scaled to obtain the particle sizes of around 20 nm for subfigure a, around 100 nm for subfigure b, around 50 nm for subfigure c and around 20 and 100 nm for subfigure d.

each particle is saved in a separate file to use further for the Maximum Likelihood Estimation. The different types of distributions that can be obtained by the simulation are presented in the figure 3.2. The simulation does not take into account the effect of drift in the solution or the experimental measurement errors which get added while determining the position of the particle. The simulation code along with the Maximum Likelihood Estimation program is added in the Appendix C for the reference.

3.3 MLE Based Size Distribution Determination

Due to the statistical nature of the random walk, the mean-square displacement measured from the finite number of the steps tracked would not be an accurate estimate of the exact values. This estimate would be different even for particles from a monodispersed population. In the conventional NTA approach adopted by Sam's program, the particle size is worked out directly from these estimated means. As these estimated means would form a broad distribution about the true mean, the radii deduced would also form a broad distribution even for the monodispersed particles. This shows that the conventional NTA distribution can not be very accurate for particles with a very narrow size distribution. MLE is developed to remove this statistical broadening. The new improved approach of forming particle size distribution which uses the iteration based Maximum Likelihood Estimation is explained in this section.

Consider the total number of detected particle-like structures to be N which are tracked for a certain number of frames. For the n^{th} particle tracked there will be a finite number of 2-D *x,y* coordinates let's say that number is l_n which also will be the number of frames for which this particular particle is detected. Further, we can calculate the total number of tracks for the particle which will be one unit less than the frame number which will be denoted as $k_n = l_n - 1$. Then the mean square displacement will be given as

$$z_n = \frac{1}{k_n} \sum_{i=1}^{k_n} d_{n,i}^2 \tag{3.2}$$

where, z_n is the mean squared or averaged squared displacement over entire tracks and $d_{n,i}^2$ will be the squared displacement for the particular track i.

So here rather than using the z_n as an accurate estimate for the $< d^2 >$ and then

calculating the radius to plot the particle size distribution, we will be finding the particle size distribution which takes into account the number of tracks for each particle or k_n along with the z_n values by using MLE.

3.3.1 Gamma PDF for Mean Squared Displacement

For the arbitrary pair of z_n and k_n, the maximum likelihood solution is obtained by evaluating the probability that a value z_n is obtained, for the parameters k_n and the temperature and viscosity of the liquid from which the particle radius r is calculated.

Consider the Gaussian nature of the Brownian motion as discussed in section 2.1. Then, each squared displacement i.e. $d^2_{n,i}$ has a random value drawn from the negative exponential distribution as Brownian motion is a random process and exponential distribution is used to predict the wait time until further event or time interval between events between random processes. The PDF of exponential distribution is given by

$$f(t;\lambda) = \begin{cases} \lambda e^{-\lambda t}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.3}$$

where, λ is a rate parameter or the average number of events per interval.

So if we consider our case, the rate of the occurrence of events will be $\lambda = \dfrac{3\pi\eta r}{2K_bT\tau}$ as the number of random displacements is the event happening here for the given interval of time.

Therefore, the mean (θ_r) of this distribution is given by

$$\theta_r = \frac{1}{\lambda} = \frac{2K_bT\tau}{3\pi\eta r} + 2\sigma_e^2 \tag{3.4}$$

where, σ^2 is the variance of the experimental measurement errors which occurs due to the uncertainty of the track length determination. The particle track is obtained by determining the difference in the positions of the same particles between the successive frames. So an estimate of this uncertainty is the equivalent physical dimension of one pixel in the CCD camera. The width of the pixel is determined by the pixel pitch i.e. the physical dimension of the CCD camera and the magnification of the optical system.

Now if we consider the equation 3.2, the product $k_n z_n$ will be a summation over 1 to k_n exponentially distributed squared displacements $d^2_{n,i}$ or random numbers each with a mean value of θ_r. The exponential distribution is a special case of Gamma distribution as the sum of exponentially distributed variables has a gamma distribution, the PDF of $j_n = k_n z_n$ which is the sum of all squared displacements will have a gamma PDF.

Let's consider the PDF for a gamma distribution

$$f(x; k\theta) = \begin{cases} \dfrac{1}{\theta^k \Gamma(k)} x^{(k-1)} exp\left(\dfrac{-x}{\theta}\right), & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{3.5}$$

putting the mean value θ as θ_r given in equation 3.4, x as j_n where $j_n = k_n z_n$ and k as k_n equation 3.5 becomes,

$$P_j(j_n; k_n, r) = \frac{j_n^{k_n - 1} \exp(-j_n/\theta_r)}{\theta_r^{k_n} \Gamma(k_n)} \tag{3.6}$$

As we are interested in M.S.D. or z_n and not j_n So we will use the change of

variables method to obtain the PDF for z_n.

Let's consider the P_j (PDF for j_n) as $f_a(a)$ and z_n as b. z_n can be written as $(1/k_n)j_n$ which is of the form $b = Ka$ where, $K = 1/k_n$ and we will be finding the PDF for z_n i.e. $f_b(b)$. According to the scale transformation formula for change of variables in PDF [61],

$$f_b(b) = f_a(b/K)/K \tag{3.7}$$

Putting the respective values for a, b, and K we will get,

$$P_d(z_n; k_n, r) = \frac{k_n(k_n z_n)^{(k_n-1)} \exp(-k_n z_n/\theta_r)}{\theta_r^{k_n}\Gamma(k_n)} \tag{3.8}$$

which is the Gamma PDF for mean squared displacement value or z_n.

3.3.2 Likelihood for Mean Squared Displacement

To find the likelihood of observing all MSD values first, we need to obtain the probability of obtaining a single MSD value. So, we will first consider that the particle radii has un-normalized PDF $P_r(r)$. According to the property of PDF, The probability of obtaining the single MSD value is then can be obtained by integrating the above PDF within the limits of 0 to ∞ as any positive values of MSD's are possible. Also, as we are interested in the particle radius distribution, we will be finding the probability of obtaining a single MSD value weighted with the different radius or 'r' values which are given as

$$P_{z_n} = \frac{\int_0^\infty P_d(z_n; k_n, r)\delta P_r(r)dr}{\int_0^\infty P_r(r)dr} \tag{3.9}$$

where, δ is the resolution to which z_n is measured and $\int_0^\infty P_r(r)$ is a normalization constant.

Now as we represent the particle size distribution in the form of a histogram, and as the histogram will always have the discrete number of bins we can write the above equation by replacing the integral with the summation over the number of bins as

$$P_{z_n} = \frac{\sum_{m=1}^{M} P_d(z_n; k_n, r_m)\delta P_m}{\sum_{m=1}^{M} P_m} \tag{3.10}$$

In the above equation, M is the total number of bins in the Histogram and $P_m = P_r(r_m)$ will be the overall particle size distribution.

The above equation gives the final form of the probability for obtaining the single value for z_n. Our ultimate aim is to find the likelihood of obtaining this value. As explained in section 2.3, the likelihood can be obtained by the product of individual probabilities. Hence, the likelihood for obtaining the different probabilities for the values of $z_1, z_2, z_3, ..., z_n$ can be obtained by multiplying these individual probabilities which will be given as,

$$L = \prod_{n=1}^{N} \frac{\sum_{m=1}^{M} P_d(z_n; k_n, r_m)\delta P_m}{\sum_{m=1}^{M} P_m} \tag{3.11}$$

Using the log-likelihood form for the above equation,

$$LL = log\left[\prod_{n=1}^{N} \frac{\sum_{m=1}^{M} P_d(z_n; k_n, r_m)\delta P_m}{\sum_{m=1}^{M} P_m}\right] \tag{3.12}$$

Therefore using the logarithmic rules,

$$LL = \sum_{n=1}^{N} log \left[\sum_{m=1}^{M} P_d(z_n; k_n, r_m) \delta P_m \right] - \sum_{n=1}^{N} log \left[\sum_{m=1}^{M} P_m \right]$$

As $\sum_{m=1}^{M} P_m$ is constant for all the 'n' values, the summation will be replaced by total number of n values i.e. N. Then, the final form of the above equation will be given as

$$LL = \sum_{n=1}^{N} log \left[\sum_{m=1}^{M} P_d(z_n; k_n, r_m) \delta P_m \right] - N log \left[\sum_{m=1}^{M} P_m \right] \tag{3.13}$$

To maximize this likelihood, we will first differentiate the above equation with respect to P_m,

$$\frac{\partial LL}{\partial P_m} = \sum_{n=1}^{N} \frac{P_d(z_n; k_n, r_m) \delta}{\sum_{m=1}^{M} P_d(z_n; k_n, r_m) \delta P_m} - \frac{N}{\sum_{m=1}^{M} P_m} \tag{3.14}$$

As δ is present in both the denominator and numerator of the first term, it will get cancelled and it also tells us that the resolution by which the mean squared displacement is measured is irrelevant to estimate the likelihood. So the equation will be

$$\frac{\partial LL}{\partial P_m} = \sum_{n=1}^{N} \frac{P_d(z_n; k_n, r_m)}{\sum_{m=1}^{M} P_d(z_n; k_n, r_m) P_m} - \frac{N}{\sum_{m=1}^{M} P_m} \tag{3.15}$$

The differential will be zero for the maximum point so to obtain the MLE, we will equalise the above equation with zero to get,

$$\sum_{n=1}^{N} \frac{P_d(z_n; k_n, r_m)}{\sum_{m=1}^{M} P_d(z_n; k_n, r_m) P_m} = \frac{N}{\sum_{m=1}^{M} P_m} \tag{3.16}$$

3.3.3 EM Algorithm for the Iterative MLE solution

As explained in section 2.4, the EM-algorithm is an iterative procedure to approximate the maximum likelihood estimator. So, to get the final estimate of the size distribution and to ensure that there are no latent variables, we will be using the form of an iterative equation which can be derived as follows using an EM algorithm. From the above equation 3.16, the following relationship can be established

$$P_m = \Psi[P_m] \tag{3.17}$$

where, Ψ is an operator and defined as

$$\Psi[P_m] = P_m . \frac{1}{N} \sum_{n=1}^{N} \left[\frac{P_d(z_n; k_n, r)}{\sum_{m=1}^{M} \frac{P_d(z_n; k_n, r) P_m}{\sum_{m=1}^{M} P_m}} \right] \tag{3.18}$$

From the observation of Ψ, it can be concluded that required solution is a fixed point for operator Ψ. So we will apply EM algorithm [18], [62], [63] where,

- The E step is where we will start with the initial estimate $P_m^{(j)}$. Here, the algorithm is always started with the uniform function for the initial particle size distribution estimate ($P_m^{(1)}$).
- For the M step, if $P_m^{(j)}$ denotes current estimate, the new estimate $P_m^{(j+1)}$ will be defined by

$$P_m^{(j+1)} = \Psi[P_m^{(j)}] = P_m^{(j)} . \frac{1}{N} \sum_{n=1}^{N} \left[\frac{P_d(z_n; k_n, r)}{\sum_{m=1}^{M} \frac{P_d(z_n; k_n, r) P_m^{(j)}}{\sum_{m=1}^{M} P_m^{(j)}}} \right] \tag{3.19}$$

This form of the equation 3.19 should converge to the fixed point and give us the Maximum Likelihood Estimate.

3.3.4 Stopping Criterion for Iterative Algorithm

As this MLE algorithm is based on the iterative solutions, defining a stopping criterion is necessary as, without the stopping criteria, the algorithm will keep repeating the iterations which could result in over smoothing or overfitting data. Therefore, the chi-squared goodness of fit formula is used where iterations are stopped when the change in χ^2 becomes smaller than 1 % of the previous value.

$$\chi^2 = \sum_{b=1}^{b_{max}} \frac{\left[H_{d^{2(*)}}(b) - H_{ML}^{(j)}(b)\right]^2}{H_{ML}^{(j)}(b)} \tag{3.20}$$

In this equation, $H_{d^{2(*)}}$ is the histogram formed by the displacement data and the H_{ML} is the calculated form of histogram obtained by the current iterative solution j which is given below.

$$H_{ML}^{(j)}(b) = \sum_{k=k_{min}}^{k_{max}} N_k \frac{P_d(b\Delta_b; k, r_m)\Delta_b P_m^{(j)}}{\sum_{m=1}^{M} P_m^{(j)}} \tag{3.21}$$

During the iterations, the χ^2 values initially decrease very rapidly but then stabilize and remain constant for each iteration.

4 | Results and Discussion: Simulated Data

4.1 Initial Checks with the Existing Program-

As mentioned earlier, the Particle Analysis software developed for the low-concentration NTA system has several characteristic features that can be used to weed out the noise so the effects of these different parameters on the particle size distribution were studied initially. To study the effect of these parameters, the available simulated video of particles *test.avi* was chosen used for all the analyses and then particle size distributions were observed by varying all the other parameters. The simulation was obtained to get the particle size range of around 150 nm to 400 nm. After some initial tests, it was observed that the global size parameter value affects the number of particles as it decides the minimum distance between particle centres or particle neighbourhoods. Therefore, the effect of different values of global size parameter is studied on the particle size distribution profile while values of other parameters were suitably chosen and kept constant e.g. Max Step Length value is chosen as 3 pixels as the value of Max Step Length should be equal to or less than global size parameter to prevent oscillation of particles or pixel size value is chosen according to the camera that was used to capture the video of particles. The effect of the different values of global size parameter on the particle count, mean particle size, standard deviation and standard error is summarized in table 4.1.

Constant Parameters for all the checks-

- Video - *test.avi*
- Maximum Percentage Standard Error - 30
- Pixel Size - 10 μm

Parameters	**Values**				
Global Size Parameter (Pixels)	3	4	5	6	7
Mean Particle Size (nm)	365.4	319.8	241.4	278.2	186.1
Particle Counts	104	36	24	23	22
Standard Deviation (nm)	157.3	188.5	163.4	166.6	120.7
Standard Error (nm)	15.43	31.43	33.37	34.74	25.74

Table 4.1: Comparison of Mean Particle Sizes, Particle Counts, Standard Deviation and Standard error with varying Global Size Parameter.

- Water Temperature - 300 K
- Minimum Expected Particle Size - 4 nm
- Max Step Length- 3 Pixels

Discussion-

It can be observed from the above table that, as we increase the global size parameter i.e. the minimum distance between the particles (in pixels) to consider them as separate particles or particle neighbourhoods from 3 pixels to 7 pixels, the number of valid tracked particles gets reduced and almost remains constant after the value 5. The Mean particle size values also follow a similar trend to the global size parameter value till 5, but the values do not remain constant after 5. The reason behind both of these trends could be that besides grouping particles together, the global size parameter also neglects the centroid larger than the defined value. So it is possible that with the increase in the global size parameter, new particles that are out of focus or have rings around them are detected and therefore, the mean particle size value changes. It can also be seen that the width of the distribution is slightly reduced after the value 5, even though the particle count is the same. But any specific conclusion cannot be drawn from these analyses as the software does not allow to check and compare the width of the Gaussian peaks fitted to this

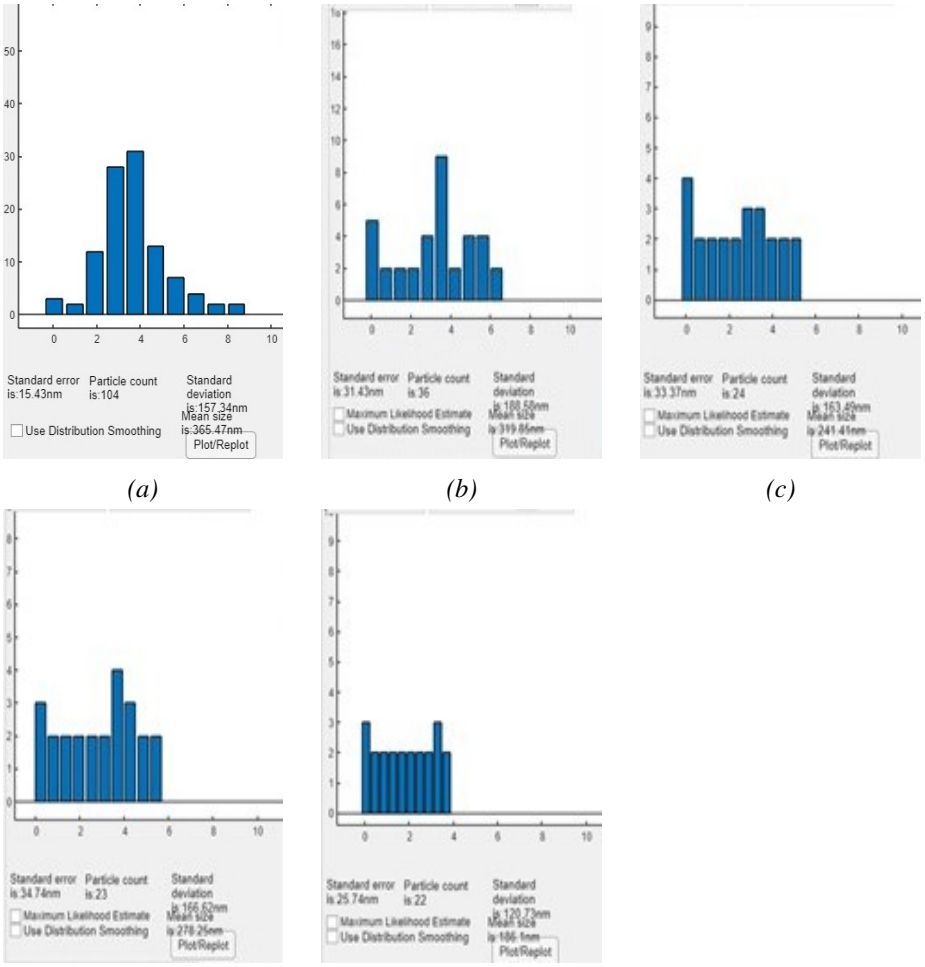


Figure 4.1: The particle size distribution obtained from video 'test.avi' with the parameter set described in the list, except for different values of 'Global Size Parameter' ranging from (a) 3 pixels, (b) 4 pixels and (c) 5 pixels (d) 6 pixels and (e) 7 pixels. As the 'Global Size Parameter' increases, the number of the 'valid' particles decreases.

distribution.

4.2 Application of MLE to the Simulated Data

As the number of particles detected from the video is very small, a good statistical estimate is not possible to check the application of MLE on this video. Therefore, Brownian Motion Simulation is used to check the application of the MLE program on several possible experimental scenarios which are presented in the following subsections. Also, as mentioned before, as MLE is an iterative program, a stopping criterion is necessary to prevent over smoothing. The stopping criterion is defined where the χ^2 value becomes smaller than 1 % of the previous value. The graph of χ^2 values for each iteration versus the number of iterations is given below where it can be clearly seen that the χ^2 value initially drops very rapidly but stabilizes after a certain number of iterations.

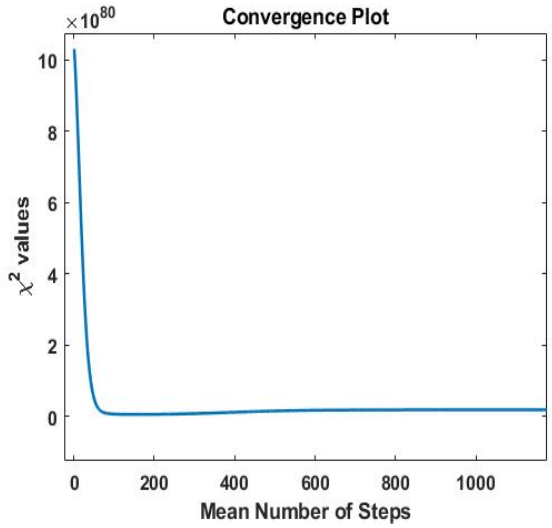


Figure 4.2: Convergence plot showing the relationship between χ^2 value and number of mean steps. χ^2 value initially decreases rapidly but then remains constant throughout the plot.

4.2.1 Monodispersed Solution

To study the effect of the Maximum Likelihood Estimation program on the simulated Brownian motion and therefore the obtained particle radii values from the Einstein-Stokes formula with the random number of generated steps and mean-squared displacement values, first monodispersed case is considered. First, the size distributions obtained by the different mean number of steps are compared. Further, for the same allowed particle value, Gaussian fitting is done to compare the widths of these distributions and to verify if MLE is actually able to give the narrower distribution when compared to the conventional approach.

4.2.1.1 Comparison of ES and ES+MLE Distributions with the Original Distribution

User Defined Parameters-

- Radius value - 50 ± 5 nm
- Mean number of steps - 10, 20, 30, 40, and 50.
- Uncertainty in the track length - 3 (Meaning random generated values of the mean number of steps will be in the range of 50 to 53)
- Number of particles - 1000

Discussion-

Figure 4.3 shows the comparison of five particle size distributions obtained for different numbers of mean particle steps. The number of particles is kept to 1000 so that better statistical estimates can be obtained. The blue bars in all the figures above is the original size distribution or defined size distribution meaning it's the starting point for the simulation to run and also the ideal distribution to recover.

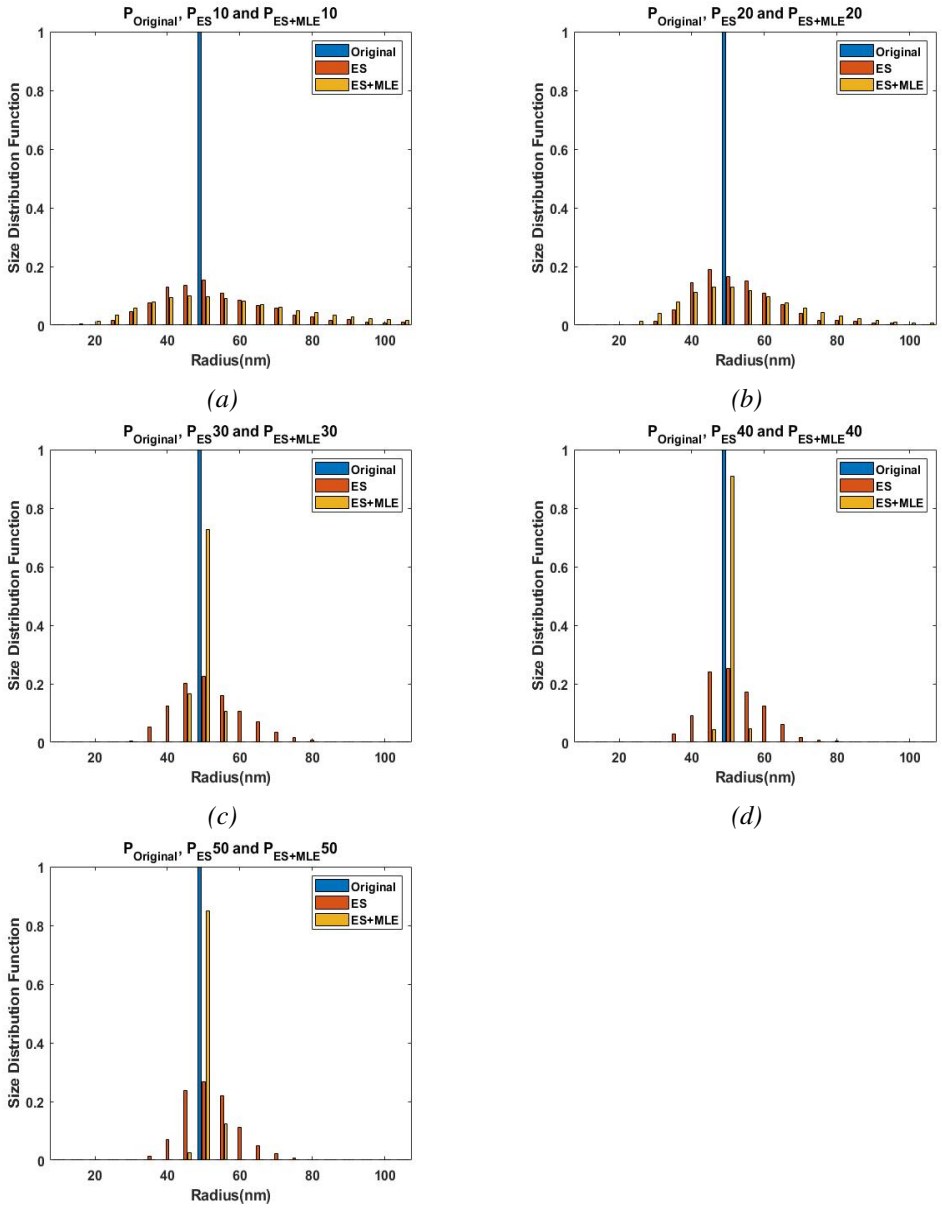


Figure 4.3: Comparison of the original particle size distribution used for molecular simulation of random walks of 1000 particles which has just one peak on 50 ± 5 *nm(Blue bar) with the recovered particle size distribution by ES method (red bars) and ES+MLE method (red bars) for different number of particle steps followed. ES+MLE size distribution is plotted with an iterative algorithm and iterations are stopped when the current* χ^2 *value is less than 1 % smaller than the value for the previous iteration. The number of iterations for the ES+MLE plots are 2, 2, 75, 95 and 111 respectively with an increase in the number of steps.*

This size distribution is obtained by only allowing the radius value of 50 ± 5 nm in the input. The red bars in all the figures are the Einstein-Stokes bars which are obtained by first generating random values(within the limit of the track length uncertainity value) of a mean number of steps, generating random mean squared displacement values scaled with the original distribution and with this added randomness again by converting these values again into the radii values to recover the original distribution. The yellow bars are the ES+MLE bars which are obtained with the combination of the Einstein-Stokes method and Maximum Likelihood Estimation method. For this distribution also first a random number of steps and scaled mean squared displacement values are generated like the ES method but rather than directly converting it back into radii values, the ES+MLE method further uses the Maximum Likelihood Estimation method and uses equation 3.19 to estimate the final distribution using iterative algorithm method. The iterative algorithm is terminated when the χ^2 value of the final estimate is less than one percent smaller than the value of the previous iterations. The values for both the ES and ES+MLE method are normalized before plotting.

In the figure 4.3, the original distribution (blue bars) has the same values and has the same peak positions in all the figures as it is the input. For the distribution where the mean number of steps is 10, the bar values for both the ES and ES+MLE are scattered all around the target radii values and not necessarily concentrated around the original distribution which is expected and simulates experimental condition also as the Brownian motion is stochastic and particles are needed to be tracked for a large number of steps (ideally infinite) to get the accurate estimate of the radii values and therefore, a small number of steps like around 10 won't give the better estimate for the particle radii values. For the next subplot, where the mean number of steps is 20, the result for both the ES and ES+MLE is better than the previous subplot in the sense that, the bars or peak positions are

slightly concentrated near the original distribution but still pretty much scattered around all the available radii values and not at all sufficient to estimate what the original distribution is. For the subplot where the mean number of steps is 30, it can be observed that the ES+MLE distribution is converging to the original distribution and the central yellow bar of the ES+MLE distribution has size distribution function values of around 0.7 nearly reaching the original distribution which has the value of 1. The ES distribution is still not converged to the original size distribution but now it's not scattered to all over the available radii values and slowly coming close to obtaining the original distribution. For the next cases of 40 and 50 steps, the ES+MLE distribution further converges nearly to the original distribution. ES distribution also starts to converge further and further to the original distribution, but still not close enough to recover the original distribution.

It is also worthy to note that, the number of iterations for the ES+MLE distribution, is constant (2 iterations) for the first two cases of 10 and 20 mean number of steps but as the mean number of steps increases the number of iterations also increases before termination. The number of iterations for 30, 40 and 50 number of steps are 75, 95 and 111 respectively. The execution time for these iterations also increases slightly with increase in number of steps but it is still in order of seconds for all the cases of 10, 20, 30, 40 and 50 mean number of steps.

4.2.1.2 Gaussian Fitting

Einstein-Stokes Distribution

Mean Number of Steps	**Width of the Peak (W) (nm)**	**Height of the Peak (unit) (H)**	**Peak Centre(nm)**	**Approx. Area under the Peak W*H** $(unit)^2$
10	19.7	0.1375	51.14	2.708
20	14.95	0.1863	50.04	2.785
30	11.9	0.2342	50.06	2.786
40	10.6	0.2619	49.3	2.776
50	10.06	0.2781	50.19	2.797
75	8.026	0.3448	49.8	2.767
100	7.435	0.3802	49.98	2.826

Table 4.2: Gaussian fitting Data for the monodispersed ES Distribution

ES+MLE Distribution-

Mean Number of Steps	**Width of the Peak (W) (nm)**	**Height of the Peak (unit) (H)**	**Peak Centre(nm)**	**Approx. Area under the Peak W*H** $(unit)^2$
10	29.91	0.0952	54.17	2.847
20	20.94	0.1321	51.01	2.766
30	4.929	0.5719	49.75	2.817
40	4.881	0.578	49.47	2.821
50	2.786	0.9435	49.84	2.628
75	1.051	0.962	50	1.011
100	0.763	0.9722	50	0.741

Table 4.3: Gaussian fitting Data for the monodispersed ES+MLE Distribution

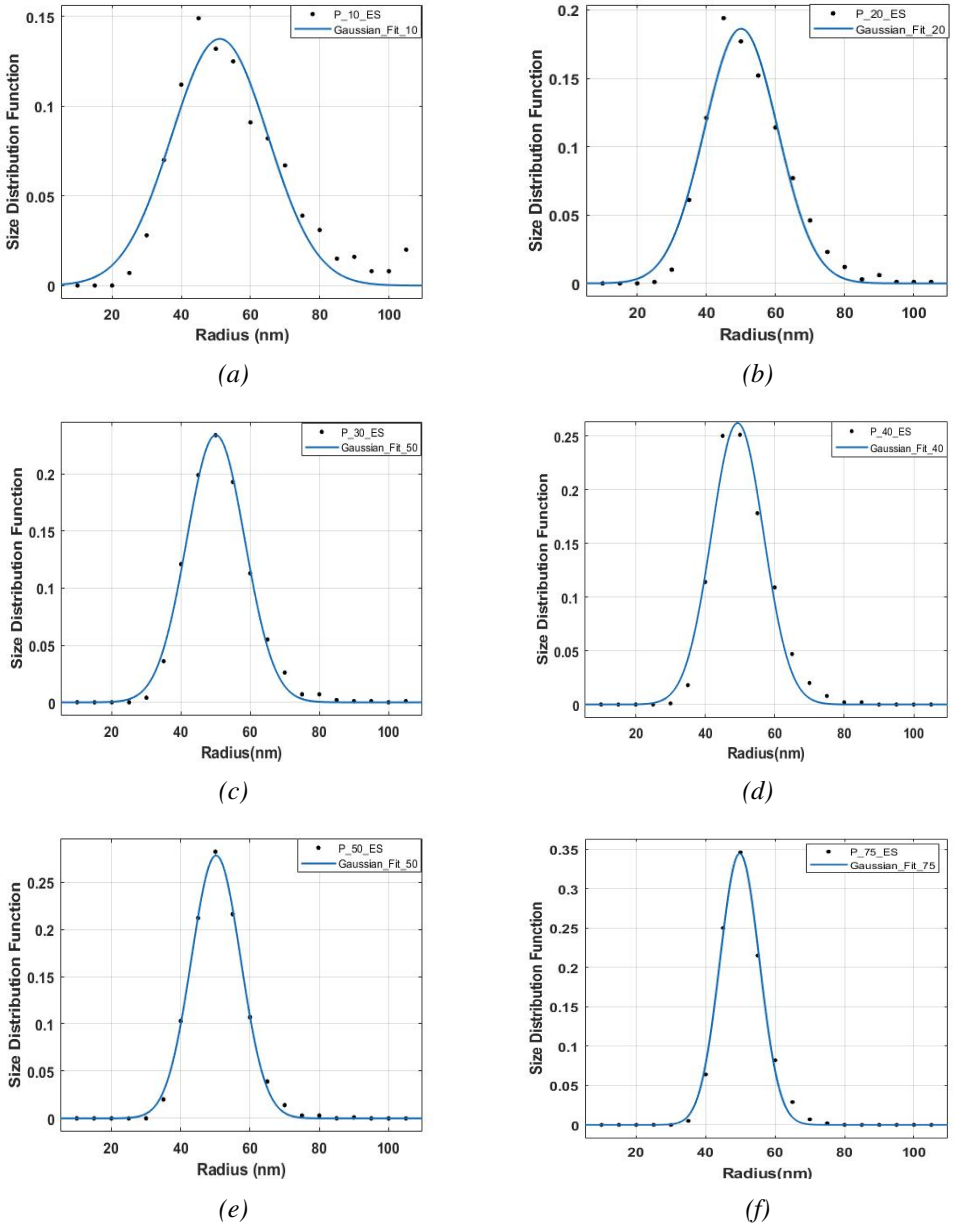


Figure 4.4: The figure represents the Gaussian fitting for the normalized distribution obtained by Einstein-Stokes method for 10, 20, 30, 40, 50 and 75 mean number of steps respectively for a.b,c,d and e subplots. The variance of the noise in measuring step size in the experimental setup is assumed to be zero.

Discussion-

Figure 4.4 shows the Gaussian fitting for the particle size distribution obtained from the original distribution by Einstein-Stokes method and figure 4.5 shows the Gaussian fitting for the size distribution obtained from the Einstein-Stokes plus Maximum Likelihood Estimation. The Gaussian fitting is done for the mean number of particle steps of 10, 20, 30, 40, 50, 75 and 100 to compare the widths of the peaks obtained from both methods. From comparing the widths of peaks it is possible to determine if the Maximum Likelihood Estimation method is quicker in recovering the original distribution and if it is quicker, then how quickly it can reach the original distribution.

The original distribution is defined on the 50 ± 5 nm value for a monodispersed case as only one size is present. In figure 4.4, which represents the ES case it can be observed that for the case of 10 mean number of steps distribution width is 19.7 nm which is expected as a small number of steps cannot give accurate size estimation. As we increase the mean number of steps to 20, the peak width gets reduced to 14.95 nm. For the 30 mean number of steps, the peak width further decreases as it is known that the increase in the number of steps improves the estimate. Further for the 40 and 50 number of steps, peak width further reduces to 10.6 and 10.06 and now almost started to remain constant.

In the case of 'ES+MLE' distribution, the peak width for the 10 mean number of steps is 29.91 nm which is broader than the 'ES' method and for the mean number of steps of 20, the peak width is 20.94 nm which is still slightly broader than the ES method but as the number of steps goes to 30, the peak width suddenly drops to 4.929 nm and further reduces to the 4.881 and 2.786 nm respectively for 40 and 50 number of steps. So in comparison to the ES method, MLE is unable to estimate the original size distribution at first but as the number of steps increases

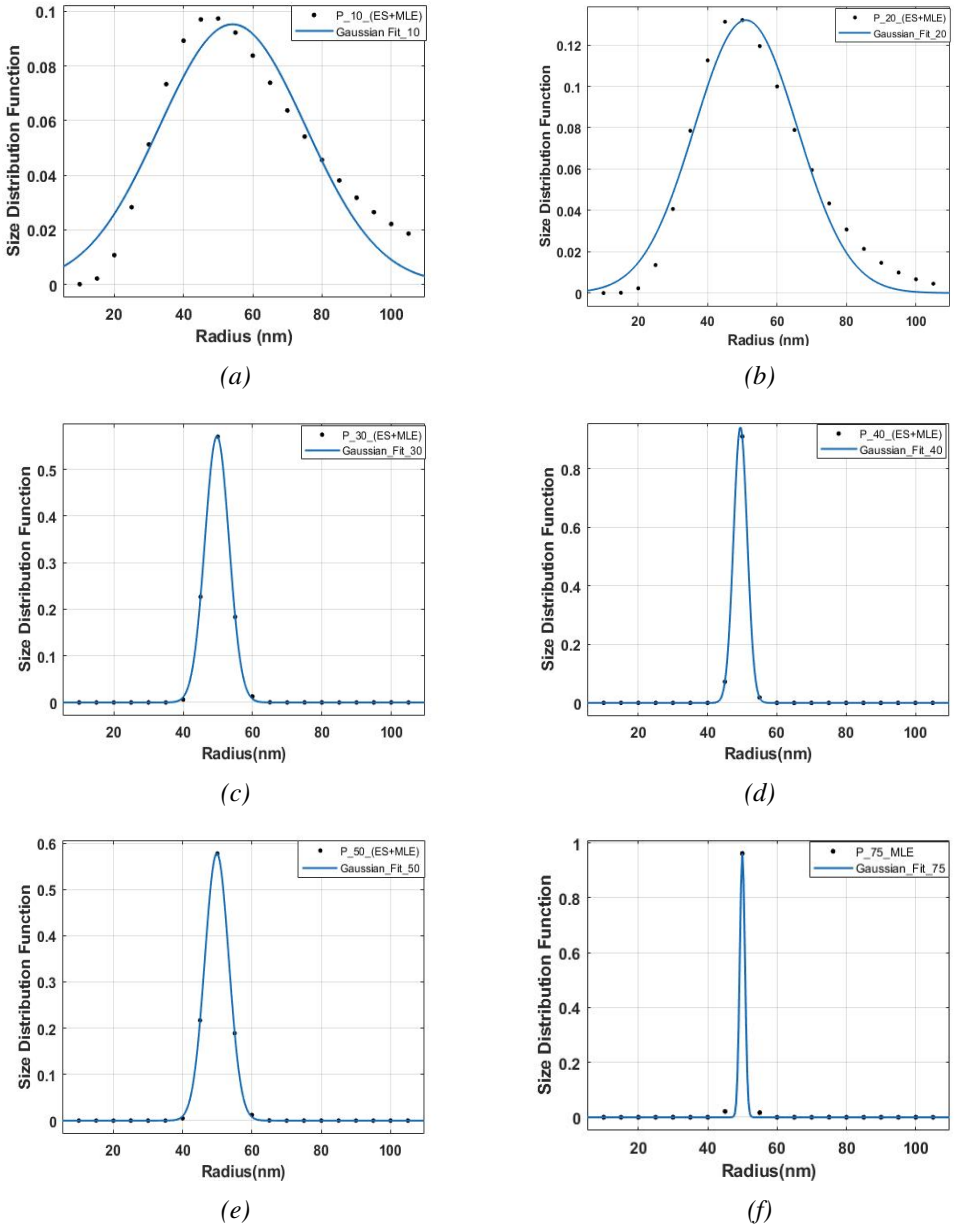


Figure 4.5: The figure represents the Gaussian Fitting for the normalized distribution obtained from the original distribution by Einstein-Stokes plus Maximum Likelihood Estimation method for 10, 20, 30, 40, 50 and 75 number of steps respectively for a, b, c, d, e, and f subplots. The original distribution is a monodispersed case where the peak is defined on 50 ± 5 *nm. The variance of the noise in measuring step size in the experimental setup is assumed to be zero. ES+MLE size distribution is obtained with an iterative algorithm and iterations are stopped when the current* χ^2 *value is less than 1% smaller than the value for the previous iteration.*

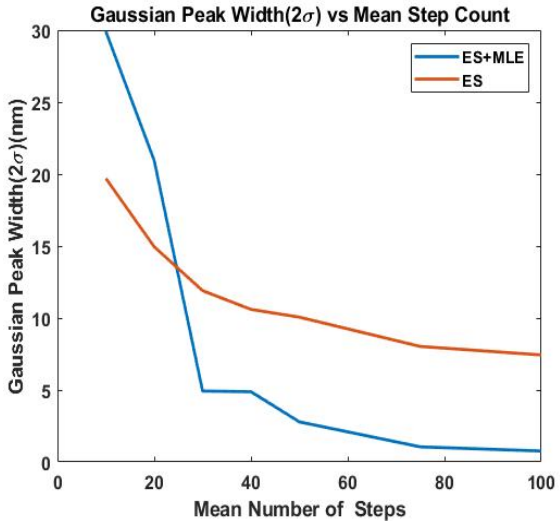


Figure 4.6: Figure shows the comparison between the Gaussian peak width or (2σ) and the different number of mean step counts for both the 'ES' and 'ES+MLE' method. The original distribution is defined to be 50 ± 5 and the ES+MLE iterations are stopped when the current χ^2 value is less than 1 % of the value for the previous iteration. The variance of the noise in measuring step size in the experimental setup is set to zero.

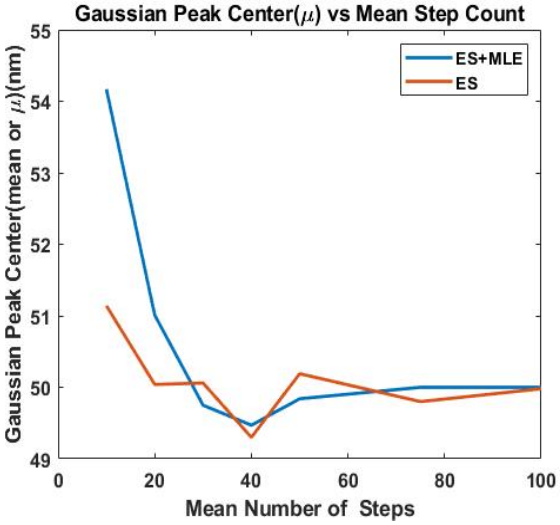


Figure 4.7: Figure shows the comparison between the Gaussian peak centre or (μ) and the different number of mean step counts for both the 'ES' and 'ES+MLE' method. The original distribution is defined to be 50 ± 5 and the ES+MLE iterations are stopped when the current χ^2 value is less than 1 % of the value for the previous iteration. The variance of the noise in measuring step size in the experimental setup is set to zero.

slightly it is quickly able to estimate the original distribution and the width of the Gaussian peak reduces rapidly after a certain number of steps while in ES method, the decrease in the width of the peak is steady and if we need to obtain the original distribution from this method, we need to track the particle for more number of steps as compared to MLE method.

Figure 4.6 gives this comparison with the graph of Gaussian peak width vs the mean number of steps and it is evident from the slope of the line which is 0.2363 units for ES method and 0.7031 units for the ES+MLE method that the ES+MLE method is able to obtain original distribution more quickly as the slope of the line is greater for the ES+MLE method than the ES method.

The height of the peaks are also increasing with the decrease in the width of the peak for ES and ES+MLE method and the area under the peak (width*height) is almost remaining constant which suggests that the fitting and therefore MLE is not giving any false result and the decrease in peak width is not random but according to increase in the peak height with respect to the peak position. Also, for the ES method, the peak position is always within the ± 1 nm of the defined distribution of 50 nm but in the case of the ES+MLE method, the peak position is not accurate for the mean number of steps of 10 but quickly goes to almost at the accurate position of the peak to the 50 nm with the increase in the number of steps and therefore, the number of iterations.

4.2.2 Bi-dispersed solution-

4.2.2.1 Comparison of ES and ES+MLE Distributions with the Original Distribution

User Defined Parameters-

- Original radii values - 25 ± 5 nm, 70 ± 5 nm with occurrence probability of 50% each
- Mean number of steps - 10, 20, 30, 40 and 50.
- Uncertainty in the track length - 3 (Meaning random generated values of the mean number of steps will be in the range of 50 to 53)
- Number of particles - 1000

Discussion-

Figure 4.8 shows the comparison of five particle size distributions obtained for different numbers of mean particle steps. Like the previous case, the number of particles is kept to 1000 so that better statistical estimates can be obtained. The blue bars in all the figures above are the original size distribution. This size distribution is obtained by only allowing the radii values of 25 ± 5 nm and 70 ± 5 nm in the input and normalized further. The red bars in all the figures belong to the Einstein-Stokes radius histogram which is obtained by the method described in the previous case. The yellow bars belong to the ES+MLE size distribution which is obtained with the combination of the Einstein-Stokes method and Maximum Likelihood Estimation method. The iterative algorithm is terminated when the χ^2 value of the final estimate is less than one percent smaller than the value of the previous iterations. The values for both the ES and ES+MLE method are normalized before plotting.

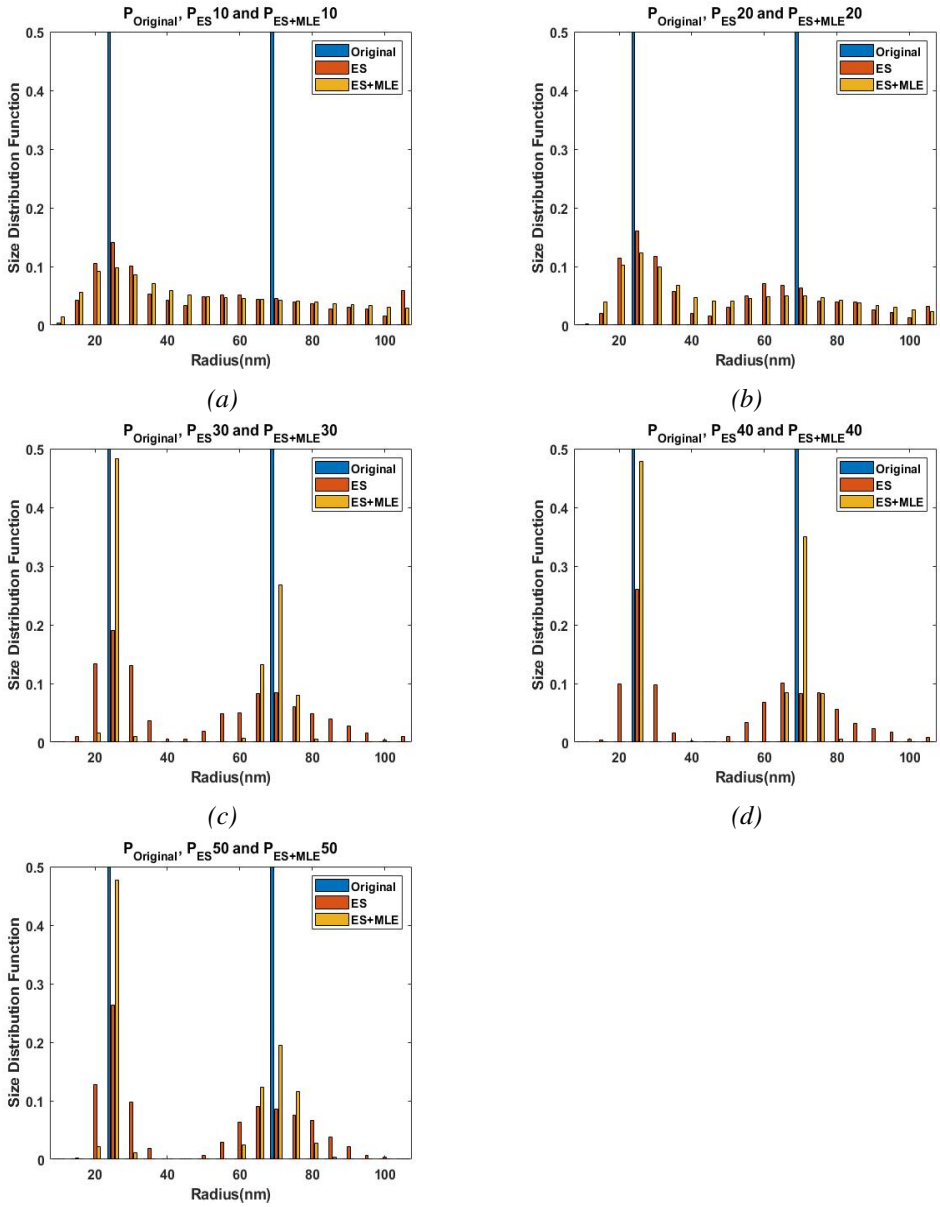


Figure 4.8: Comparison of the original particle size distribution used for molecular simulation of random walks of 1000 particles, has two defined radii values on $25 \pm 5nm$ *and* $70 \pm 5nm$*(Blue bars) with the recovered particle size distribution by ES method (red bars) and ES+MLE method (red bars) for different number of particle steps followed (10, 20, 30, 40 and 50, as indicated in the subplot titles). ES+MLE size distribution is plotted with an iterative algorithm and iterations are stopped when the current* χ^2 *value is less than 1 % smaller than the value for the previous iteration. The number of iterations for the ES+MLE plots are 2, 2, 79, 98 and 119 respectively with an increase in the number of steps.*

In the figure, the original distribution(blue bars) has the same values and has the same peak positions in all the figures as it is the input and like the previous case, for the distribution where the mean number of steps of 10 and 20, the bar values for both the ES and ES+MLE are scattered all around the available radii values. Again, for the subplot where the mean number of steps is 30, it can be observed that the ES+MLE distribution is converging to the original distribution and yellow bars of ES+MLE distribution has size distribution function values of around 0.48 on 25 nm radius value and around 0.25 on 70 nm radius value as compared to 0.5 and 0.5 for the original blue bars. For the next cases of 40 nm and 50 nm, the ES as well as the ES+MLE distribution further converges nearly to the original distribution. The number of iterations for the ES+MLE distribution is constant (2 iterations) for the first two cases of 10 and 20 mean number of steps but as the mean number of steps increases the number of iterations also increases before termination. The number of iterations for 30, 40 and 50 nm cases are 79, 98 and 119 respectively. So it can be observed that for the bidispersed case, the number of iterations required are greater than the monodispersed case in the aspect of the mean steps value of 30,40 and 50.

4.2.2.2 Gaussian Fitting

Einstein-Stokes Distribution

Discussion-

Figure 4.9 shows the Gaussian fitting for the particle size distribution obtained from the original distribution by Einstein-Stokes method and figure 4.10 shows the Gaussian fitting for the size distribution obtained from the Einstein-Stokes plus Maximum Likelihood Estimation. The Gaussian fitting is done with respect to the mean number of particle steps of 10, 20, 30, 40, 50, 75 and 100 to compare

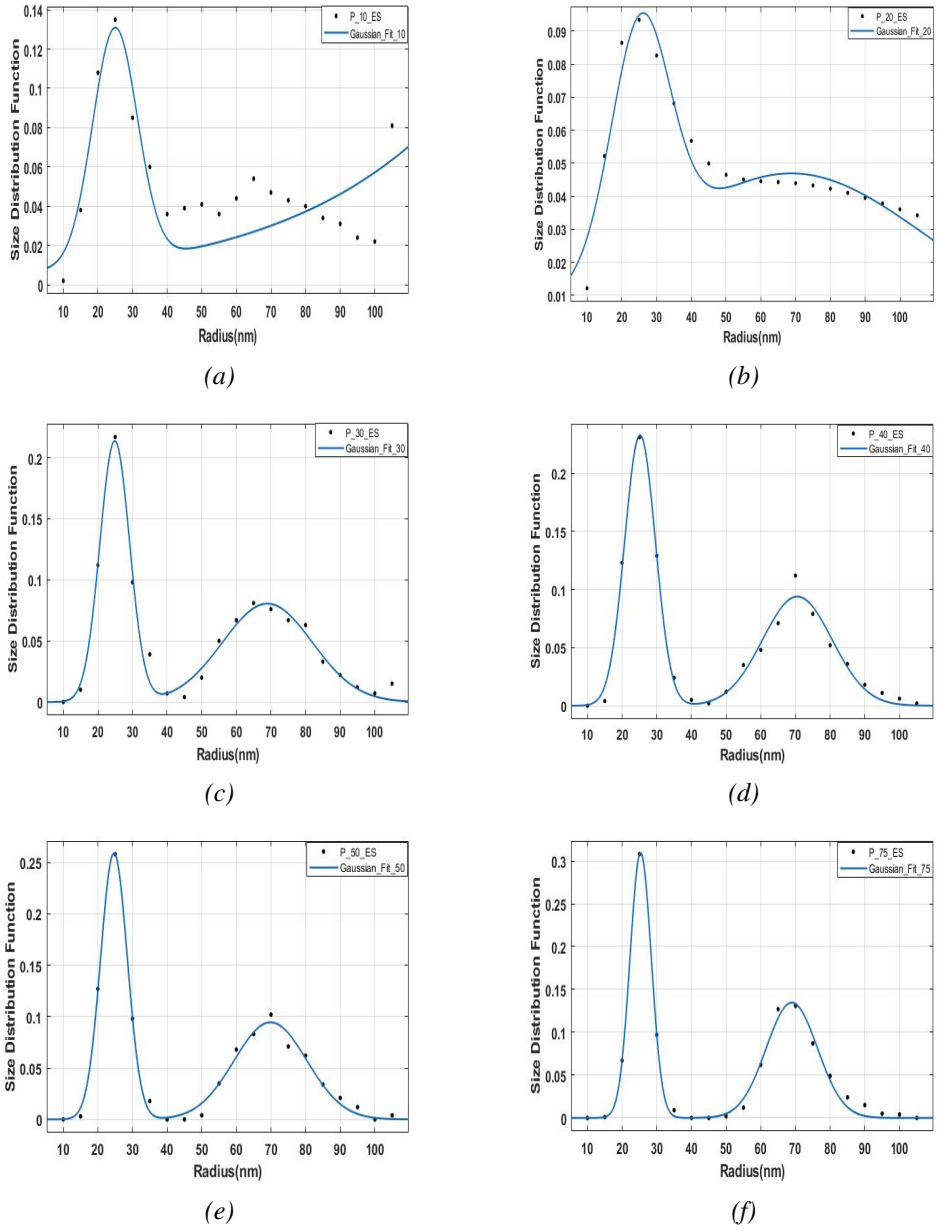


Figure 4.9: The figure represents the Gaussian Fitting for the normalized distribution obtained by Einstein-Stokes method for 10, 20, 30, 40, 50 and 75 mean number of steps respectively for a.b,c,d and e subplots where original distribution has radii values of 25 ± 5 *nm and* 70 ± 5 *nm. The variance of the noise in measuring step size in the experimental setup is assumed to be zero.*

Mean Number of Steps	**Width of Peaks (W) (nm)**	**Height of Peaks (Size Distribution units) (H)**	**Peak Centres (nm)**
10	24.93, NA	0.1197, NA	24.93, NA
20	11.79, 54.45	0.0703, 0.0469	25.41, 68.7
30	6.085, 18.36	0.2137, 0.0805	24.87, 68.98
40	6.281, 14.15	0.2329, 0.0939	25.18, 70.55
50	5.456, 14.85	0.2593, 0.0945	24.65, 69.86
75	4.326, 10.43	0.3099, 0.1347	25.35, 68.95
100	3.93,9.684	0.3582,0.1434	25.09,69.54

Table 4.4: Gaussian fitting Data for the bidispersed ES Distribution where, the value before comma(,) in any particular cell gives the value for the 25 ± 5 *nm radius value and the value after comma(,) gives the value for* 70 ± 5 *nm radius value.*

the widths of the peaks obtained from both methods.

The original distribution is defined on the 25 ± 5 nm and 70 ± 5 nm radii values. In figure 4.9, which represents the ES case it can be observed that for the case of 10 mean number of steps the two-peak Gaussian fitting is not possible as the recovered ES distribution has distributed radii values and therefore, is not well suited for the two-peaks Gaussian fitting model whereas, for the same mean number of steps, in the case of ES+MLE distribution, it is possible to fit the two-fit model but both the peaks have very large width. From the mean number of steps of 20, it is possible to fit the two-peak model to all the ES distributions and a similar trend is observed as the monodispersed case.

It is interesting to note one more thing that although there is an equal chance of obtaining the particles of the values of 25 ± 5 nm and 70 ± 5 nm in the simulation, the peak width and height is not equal for ES and ES+MLE distribution and the peak at 25 nm has a narrower peak width for 10, 20, 30, 40, 50 and 75 mean number of steps. As the value of the mean number of steps reaches 100, ES+MLE can give almost equivalent peak width and height but in the case of ES distribution,

there is still the difference of almost 6 nm in the peak width.

ES+MLE Distribution

Mean Number of Steps	**Width of Peaks (W) (nm)**	**Height of Peaks (Size Distribution units) (H)**	**Peak Centres (nm)**
10	11.79, 54.45	0.0703, 0.4692	25.41, 68.7
20	9.116, 38.51	0.1242, 0.0490	24.8, 69.95
30	2.393, 5.663	0.4715, 0.2585	24.85, 68.88
40	2.5, 4.206	0.4965, 0.3236	25.1, 70.01
50	2.245, 4.393	0.6008, 0.3185	24.91, 69.96
75	2.062, 3.851	0.4794, 0.3744	25.03, 69.68
100	1.739, 1.056	0.504, 0.4928	25, 70

Table 4.5: Gaussian fitting Data for the bidispersed ES+MLE Distribution where, the value before comma(,) in any particular cell gives the value for the 25 ± 5 *nm peak and the value after comma(,) gives the value for* 70 ± 5 *nm peak.*

Figure 4.11 shows the comparison for the Gaussian peak widths (2σ) with the increase in the number of mean step sizes from 10 to 100 for the bi-dispersed solution i.e. the original distribution is defined with the two peaks at 25 ± 5 nm and 70 ± 5 nm and then the ES and ES+MLE methods are used to recover the original size distribution. For both the peaks the analysis shows almost similar results for both the ES and ES+MLE method except for the mean count of steps of 10. For both the peaks, with the ES and ES+MLE method the peak width reduces rapidly till the mean number of steps of 30 and then almost remains constant throughout till 100. The ES+MLE has narrower peak width than the ES method throughout the analysis except in the case of the mean number of steps of 10 for 70 nm peak where the peak width is almost equal for both the methods. Also, it is interesting to note that even if it is expected to get the equal peak width as chances of getting the particles of both sizes are 50-50% (defined in the simulation), both ES and ES+MLE method fails to get the same peak width for two peaks throughout the

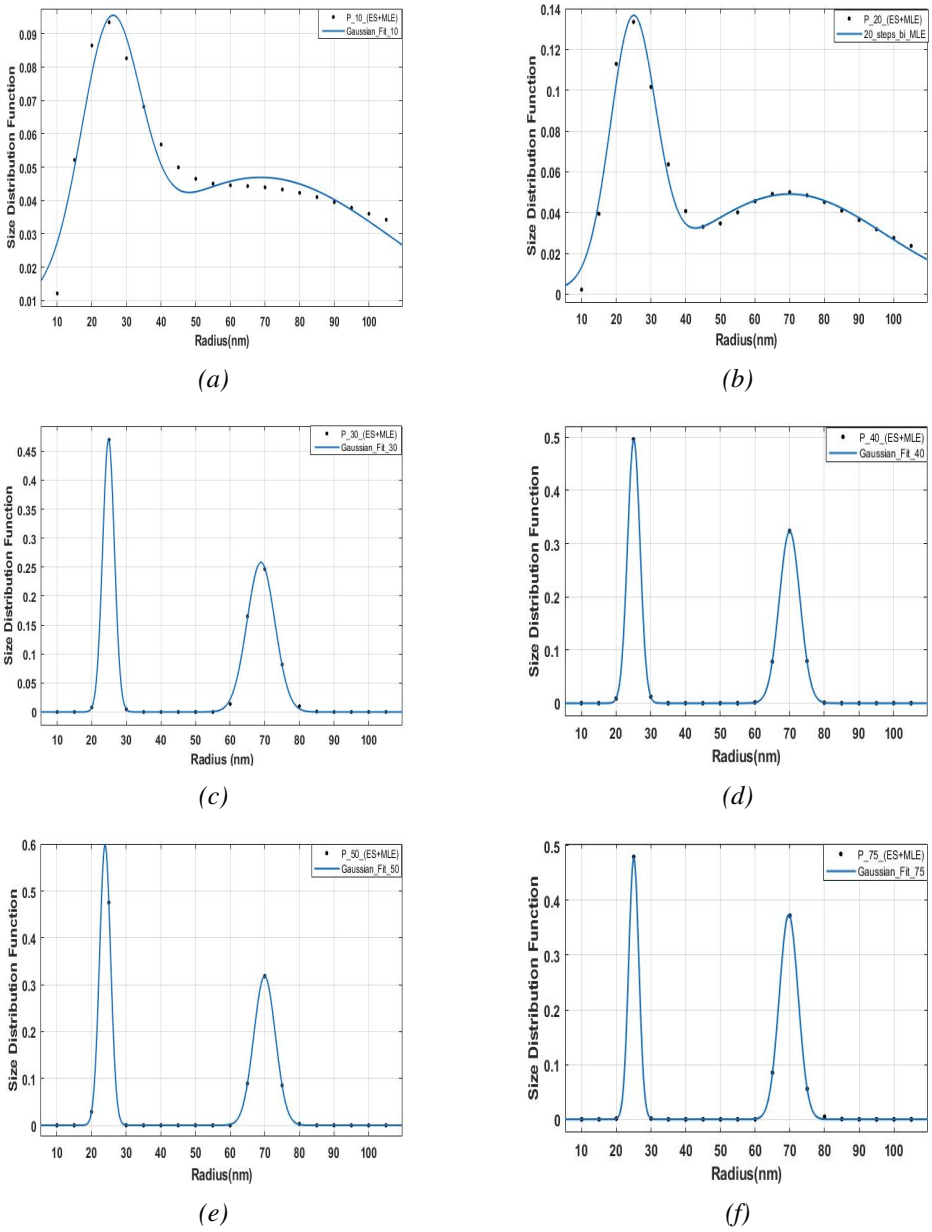
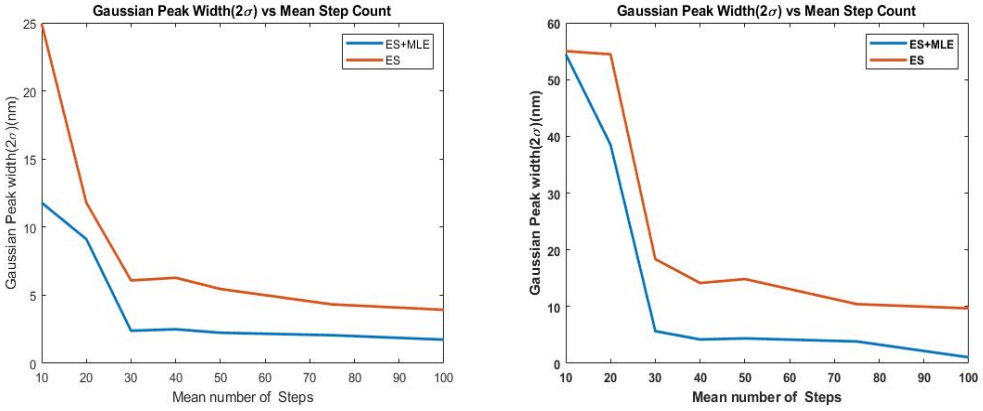


Figure 4.10: The figure represents the Gaussian Fitting for the normalized distribution obtained from the original distribution by Einstein-Stokes plus Maximum Likelihood Estimation method for 10,20,30,40, 50 and 75 number of steps respectively for a, b, c, d, e, and f subplots. The original distribution is a bidispersed case where the peak is defined on 25 ± 5 *nm and* 70 ± 5 *nm. The variance of the noise in measuring step size in the experimental setup is assumed to be zero. ES+MLE size distribution is obtained with an iterative algorithm and iterations are stopped when the current* χ^2 *value is less than 1% smaller than the value for the previous iteration.*

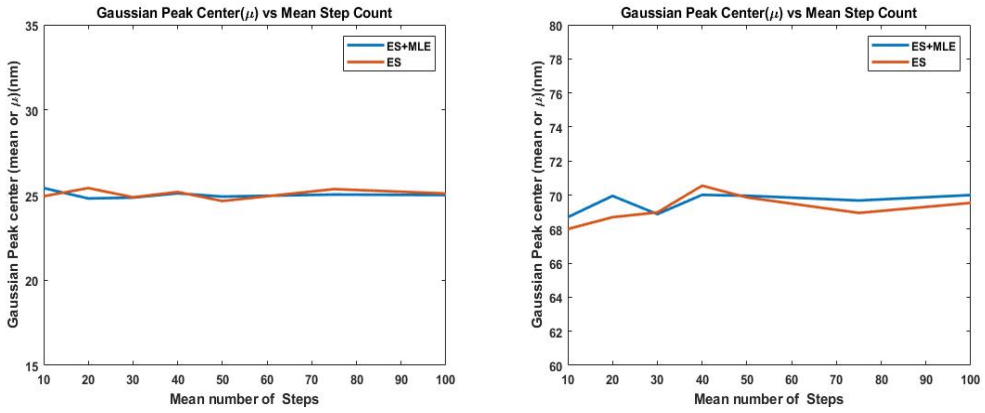


(a) *(b)*

Figure 4.11: Figure shows the comparison between the Gaussian peak width or (2σ) with the mean step count for both the 'ES' and 'ES+MLE' method. The original distribution is defined by the two peaks on 25 ± 5 *nm and* 70 ± 5 *nm and the ES+MLE iterations are stopped when the current value is less than 1% of the value for the previous iteration. The sub figure a compares the ES and ES+MLE Gaussian peak width with step count for peak defined at* 25 ± 5 *nm in original distribution while the sub figure b compares the ES and ES+MLE Gaussian peak center with the step counts for peak defined at* 70 ± 5 *nm in the original distribution. The variance of the noise in measuring step size in the experimental setup is set to zero.*

comparison. The ES+MLE method almost shows the equal-sized peaks for the mean steps of 100 but this method also fails to show the equal peak width for a lesser number of mean steps as the ES method.

Figure 4.12 compares the Gaussian peak centres or μ with the increase in the number of step sizes for the ES and ES+MLE methods wherein subplot 'a' the peak centre is defined in the original distribution at 25 ± 5 nm and then comparison with the peak centres obtained by both the method with the increase in the number of steps is shown and in subplot 'b' the peak centre is defined in the original distribution at 70 ± 5 nm and then the comparison with the peak centres obtained by both the method with the increase in the number of steps is shown. For both the methods there are small oscillations of peak centre values around the defined peak centres, But for the ES+MLE method, for 100 mean number of steps peak



(a) *(b)*

Figure 4.12: Figure shows the comparison between the Gaussian peak center or mean(μ) with the mean step count for both the 'ES' and 'ES+MLE' method. The original distribution is defined by the two peaks on 25 ± 5 *nm and* 70 ± 5 *nm and the ES+MLE iterations are stopped when the current value is less than 1% of the value for the previous iteration. The sub figure a compares the ES and ES+MLE Gaussian peak width with step count for peak defined at* 25 ± 5 *nm in original distribution while the sub figure b compares the ES and ES+MLE Gaussian peak center with the step counts for peak defined at* 70 ± 5 *nm in original distribution. The variance of the noise in measuring step size in the experimental setup is set to zero.*

centre reaches at exact 25 and 70 nm up to significant numbers of 4.

4.3 Weakness of the MLE Approach

Although MLE is useful to get the accurate size distribution obtained from the Nanoparticle Tracking Analysis system, there is one inherent problem of overfitting associated with the MLE approach if the stopping criterion is not applied. The figure 4.13 illustrates the comparison of the original size distribution with the ES+MLE distribution when a large number of iterations are taken for the analysis. The radii values are sampled randomly to get the original distribution. The original distribution is defined in such a way where 35 ± 5 has the highest occurrence probability and the probability of occurrence for the values 30 ± 5 nm and 40 ± 5

nm are equal but higher than the 25 ± 5 nm and 45 ± 5 nm. The 25 ± 5 nm and 45 ± 5 nm have equal but lowest occurrence probability than any other value. For the iterations of 4 and 8, MLE shows a very wide distribution than the original distribution and therefore lower size distribution function values than the original distribution. For 20 iterations MLE almost goes to the original distribution with size distribution function values almost similar to the original distribution. But it can be observed that for the 400 iterations, the ES+MLE goes quite close to the original distribution but the size distribution function value is slightly larger for the 35 ± 5 nm peak. For 4000 and 40000 number of iterations, it can be observed from the figures that MLE distribution favours the radius value which has the highest occurrence probability. This is because MLE selects the best possible parameters and is biased towards the highest occurrence probability. As the probability of occurrence of 35 ± 5 is higher, it converges to that value as compared to the other available values.

From the above discussion it can be concluded that the 20-40 iterations produce the best fit and even if the self convergence criteria are not met for this number of iterations, it is recommended that the user should stop the iterations within this range.

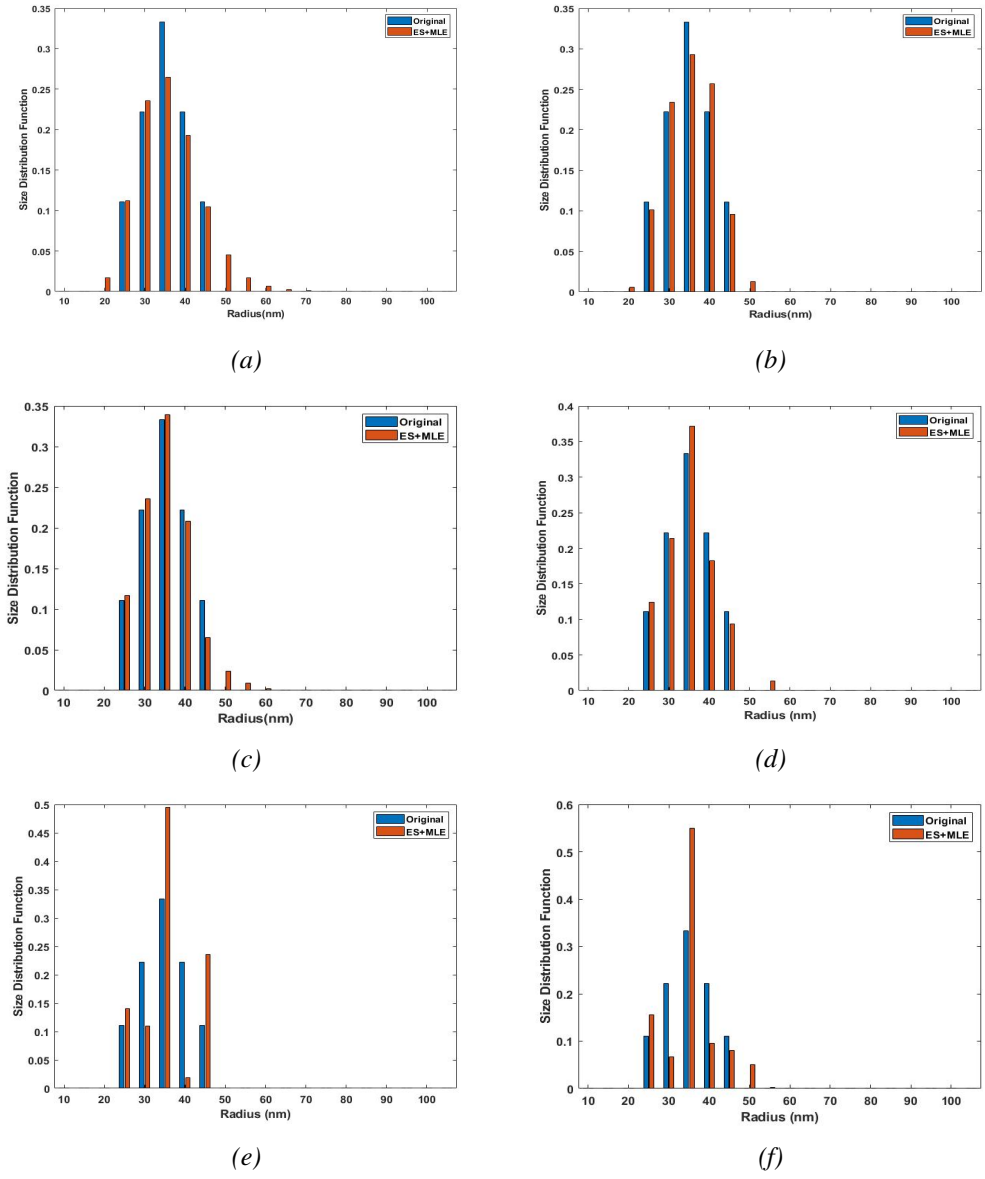


Figure 4.13: Figure shows the comparison of the ES+MLE distribution to the original distribution. The original distribution is defined on the 25 ± 5 *nm,* 30 ± 5 *nm,* 35 ± 5 *nm,* 40 ± 5 *nm and* 45 ± 5 *nm where, the probability of occurrence for the value* 35 ± 5 *nm is higher than any other value. The probability of occurrence for the values* 30 ± 5 *nm and* 40 ± 5 *nm are equal but higher than the* 25 ± 5 *nm and* 45 ± 5 *nm. The* 25 ± 5 *nm and* 45 ± 5 *nm have the equal but lowest occurrence probability. The ES+MLE distribution is obtained for the 4, 8, 20, 400, 4000 and 40,000 number of iterations respectively for the subfigures a, b, c, d, e and f. The variance of the noise in measuring step size in the experimental setup is set to zero.*

5 | Results and Discussion: Experimental Data

For further testing, the application of the MLE algorithm on experimental data, the size distributions obtained from the video samples of the food-grade TiO_2 are used. Food grade TiO_2 (E171) has been used as a food additive for its whitening effect [64]. There is still a doubt about health risks associated with the consumption of TiO_2 as it contains nanoparticles that can persist in a body even after a long time after consumption [65]. The characterisation of this food-grade TiO_2 has been done as a part of the research project at the University of Leeds to understand the surface chemistry properties [66]. To check the change in surface chemistry and surface properties, thermal treatment is done on the TiO_2 or E171 sample where E171 powder was heated in air to temperatures of 500^oC and 1000^oC. TiO_2 nanoparticles have a phosphate coating around them and by heating the sample, it can be possible to alter their phosphate coating and thereby to change the dispersion [66], [67]. The videos of TiO_2 samples obtained by the low-concentration NTA system of the University of York are used to study the difference between particle size distribution profiles obtained by the conventional NTA approach and the MLE approach.

5.1 Standard Analysis

For the NTA analysis, 3 readily available videos for each untreated, 500^0C and 1000^0C samples of food-grade TiO_2 dispersed in water (E171) are used. Samples were heated in the air before dispersing in the water. The videos were taken for different durations with the first video being 30 seconds long, the second with 60 seconds and the third video with the 90 seconds duration. These videos are taken

by the BSc students at the University of York as part of their BSc project. All three samples were 100 $\mu g/ml$.

Constant Parameters for all the checks-

- Global Size Parameter - 6 Pixels
- Minimum Steps per Track - 5
- Maximum Percentage Standard Error - 50
- Pixel Size - 0.6 μm
- Water Temperature - 300 K
- Minimum Expected Particle Size - 5 nm
- Max Step Length- 5 Pixels
- Horizontal Drift Correction - ON
- Vertical Drift Correction - ON

Discussion-

Figure 5.1 shows the conventional approach of the NTA particle size distribution for the untreated TiO_2 sample. The number of particles found is 1499, 841 and 619 respectively for the first, second and the third video and also the mean particle sizes obtained are 472 ± 138 nm, 413 ± 174 nm and 358 ± 117 nm.

Figure 5.2 shows the conventional approach of the NTA particle size distribution for the TiO_2 sample heated to 500^0C. The number of particles found is 3154, 548 and 572 respectively for the first, second and the third video and also the mean particle sizes obtained are 382 ± 109 nm, 323 ± 97 nm and 401 ± 125 nm.

Figure 5.3 shows the conventional approach of the NTA particle size distribution for the TiO_2 sample heated to 1000^0C. The number of particles found is 818, 698

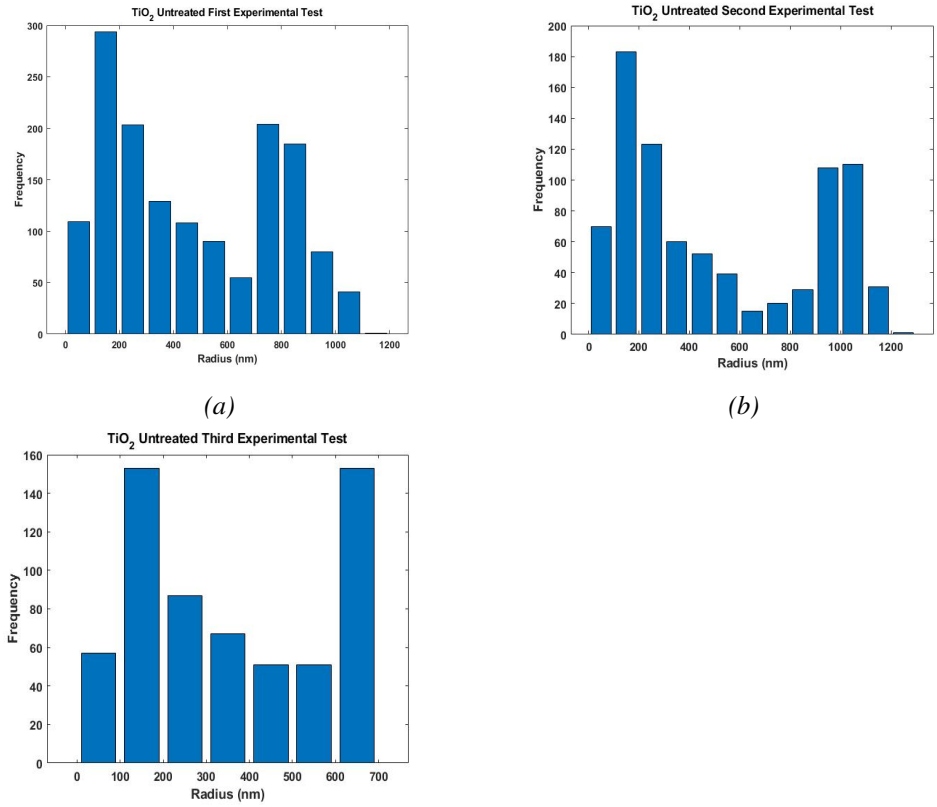


Figure 5.1: Figure shows the conventional NTA analysis for the three videos of the same untreated TiO_2 sample where the duration of the videos for the samples are taken for 30, 60 and 90 seconds respectively for a,b and c. The number of particles found for each analysis were 1499, 841 and 619 respectively.

and 2345 respectively for the first, second and the third video and also the mean particle sizes obtained are 463 ± 114 nm, 231 ± 50 nm and 293 ± 63 nm.

From the above observations, it can be seen that there is inconsistency in the particle size distribution profile even if the different videos for the same samples are used. This is maybe due to the different number of particles detected for each of the videos even if the different video of the same sample is used for the analysis. But for all of the graphs, it can be observed that there are always two peaks or maxima and a wide particle size distribution profile over multiple radii

values. This can be associated with the shortcomings of the current NTA approach of the particle size distribution which gives a broader particle size distribution than the expected particle size distribution.

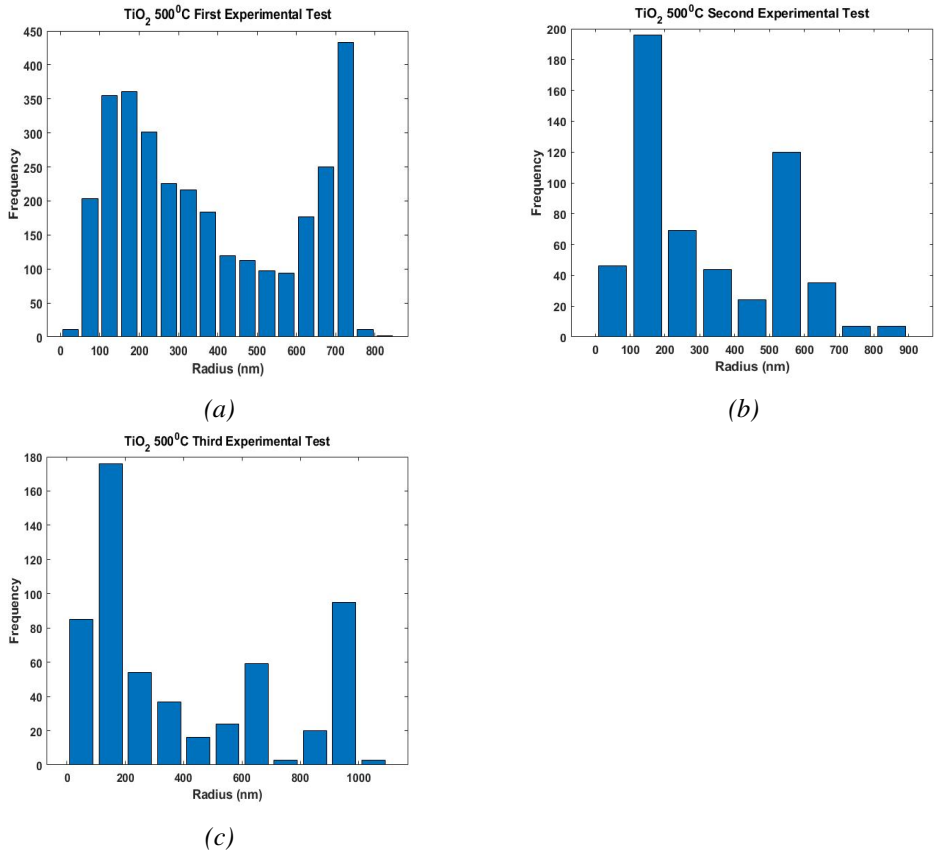


Figure 5.2: Figure shows the conventional NTA analysis for the three videos of the same TiO_2 sample heated in air to 500^0C where the duration of the videos for the samples are taken for 30, 60 and 90 seconds respectively for a,b and c. The number of particles found for each analysis were 3154, 548 and 572 respectively.

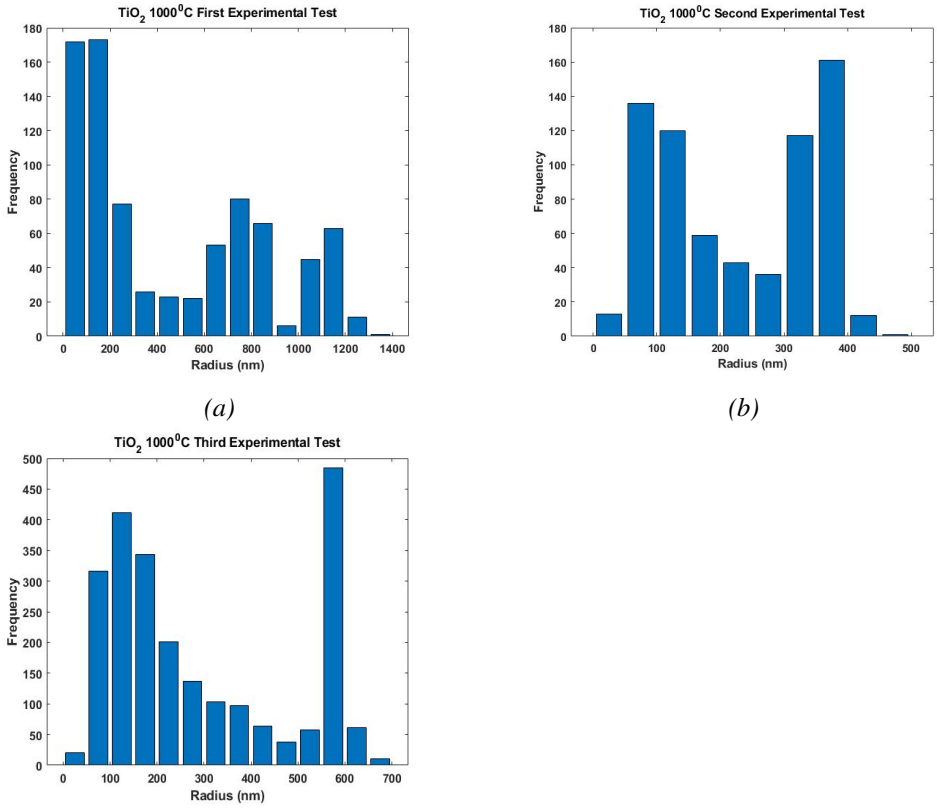


Figure 5.3: Figure shows the conventional NTA analysis for the three videos of the same TiO_2 sample heated in air to 1000^0C where the duration of the videos for the samples are taken for 30, 60 and 90 seconds respectively for a,b and c. The number of particles found for each analysis were 818, 698 and 2345 respectively.

5.2 Comparison of MLE Analysis with the standard analysis

Discussion-

Results show the comparison of the MLE approach to obtaining the particle size distribution with the conventional NTA approach. For this analysis, the variance of the noise in measuring step size in the experimental setup is set to zero and MLE size distribution is obtained with an iterative algorithm. Iterations are stopped

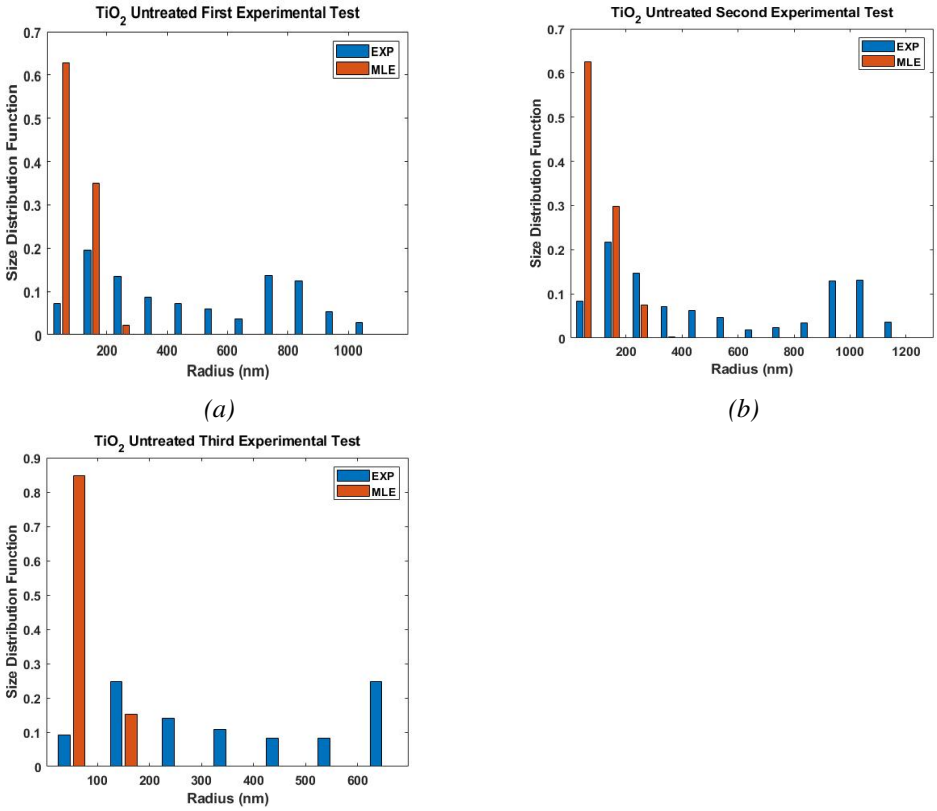


Figure 5.4: Figure shows the conventional and MLE NTA analysis for the three videos of untreated TiO_2 sample where the duration of the videos are 30,60 and 90 seconds respectively for a,b and c. Conventional NTA analysis is shown in Blue and the MLE NTA is shown in Orange. MLE size distribution is obtained with an iterative algorithm and iterations are stopped when the current χ^2 value is less than 1% smaller than the value for the previous iteration. MLE analysis shows the sample distribution lies between 50 ± 10 nm to 250 ± 10 nm for the first and second videos and between 50 ± 10 nm to 150 ± 10 nm for the third video. The number of iterations for MLE analyses are 6, 5 and 4 respectively for a, b and c. The variance of the noise in measuring step size in the experimental setup is set to zero.

when the current χ^2 value is less than 1% smaller than the value for the previous iteration for all the analyses.

For all the MLE analyses, it can be observed that for all the video samples MLE is giving narrow distribution with only two or three bins. For the untreated sample of TiO_2 it can be observed that, for the conventional NTA approach, particle size distribution is very wide and ranges from 5 to 1000 nm for the first video, 5 to 1200 nm for the second video and 5 to 650 nm for the third video. MLE shows a very narrow distribution with just three bins for the sample around 50 ± 10 nm, 150 ± 10 nm and 250 ± 10 nm for the first and second videos and two bins around 50 ± 10 nm and 150 ± 10 nm for the third video. It takes just 6, 5 and 4 iterations respectively for the first, second and third video.

Similarly for the sample heated to 500^0C, the particle size range is 0 to 850 nm for the first and second video and 0 to 1100 nm for the third video. Although MLE is showing the size distribution ranging from 25 ± 10 nm to 200 ± 10 nm for the first video, the size distribution still lies between 50 ± 10 nm and 250 ± 10 nm for the second and third video and does not show any increase or decrease in the size with the increase in the temperature. The number of iterations is 5, 4 and 3 respectively for the first, second and third video.

For the sample heated to the 1000^0C, conventional NTA similarly shows a broad size distribution profile for all the videos but, MLE shows the size distribution ranging from 25 ± 10 nm to 250 ± 10 nm. The number of iterations is 5, 4 and 4 respectively for the first, second and third video.

From the above analysis, it can be concluded that the MLE is giving narrow size distribution for all the TiO_2 samples. But according to the findings of Shah [66] and Talamini et al [65], the different sources of E171 TiO_2 particles dispersed in water at 100 $\mu g/ml$ give an average hydrodynamic particle size by DLS of

around 350 nm with sizes ranging from 50 nm to 500 nm and the average primary particle size is around 100 nm with sizes ranging from 50 nm to 200 nm for the TEM analysis. For the untreated samples the sizes obtained from the MLE results are near to those findings but with a slightly narrow size distribution for the hydrodynamic sizes with the sizes ranging from 50 nm to 250 nm. This could be the outcome of the overfitting or the effect of the stopping algorithm used. Further as shown by Shah [66] the hydrodynamic particle size for E171 heated to 500^0C and 1000^0 C and then dispersed in water at 100 $\mu g/ml$ increases to 395 nm and 554 nm respectively for the DLS analysis with TEM showing the average particle size around 187 nm. MLE analysis fails to observe this increase in size and still gives the same results with the bias towards smaller particle sizes. This bias can be the effect of the accuracy by which the mean of the exponential distribution θ_r (subsection 3.3.1) is calculated as this value is used to further obtain the Gamma PDF in the algorithm. Therefore, further investigation is required to make any conclusion about the accuracy of these results with the more improved methods such as Maximum A Posteriori or MAP estimation [20] which uses the Bayesian probability and the regularization approach.

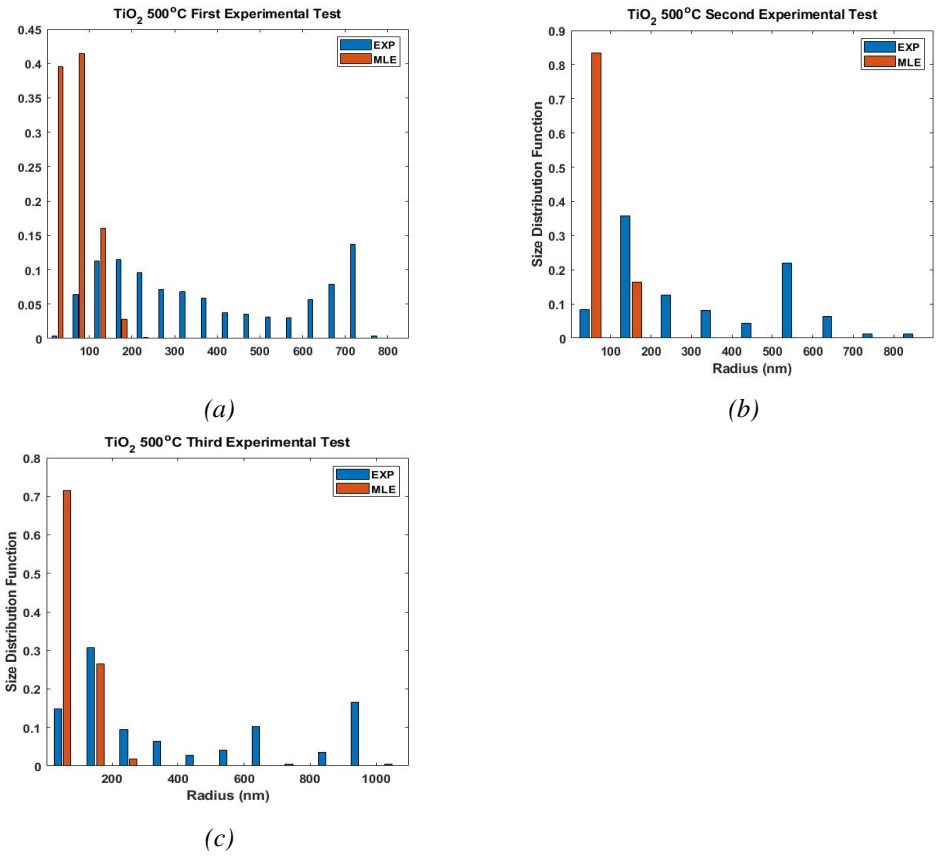


Figure 5.5: Figure shows the conventional and MLE NTA analysis for the the three videos of TiO_2 sample heated to 500^0C where the duration of the videos are 30,60 and 90 seconds respectively for a,b and c. Conventional NTA analysis is shown in Blue and the MLE NTA is shown in Orange. MLE size distribution is obtained with an iterative algorithm and iterations are stopped when the current χ^2 value is less than 1% smaller than the value for the previous iteration. MLE analysis shows the sample distribution lies between 25 ± 10 nm to 200 ± 10 nm for the first video and between 50 ± 10 nm to 250 ± 10 nm for the second and third video. The number of iterations for MLE analyses are 5, 4 and 3 respectively for a, b and c. The variance of the noise in measuring step size in the experimental setup is set to zero.

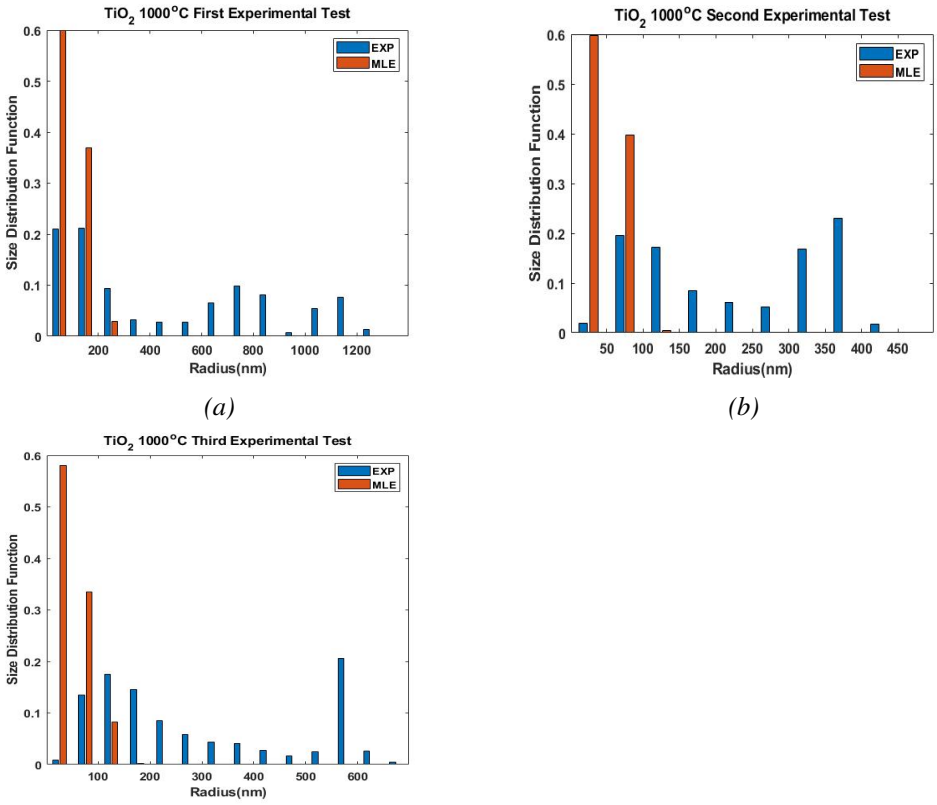


Figure 5.6: Figure shows the conventional and MLE NTA analysis for the three videos of TiO_2 sample heated to 1000^0C where the duration of the videos are 30,60 and 90 seconds respectively for a,b and c. Conventional NTA analysis is shown in Blue and the MLE NTA is shown in Orange. MLE size distribution is obtained with an iterative algorithm and iterations are stopped when the current χ^2 value is less than 1% smaller than the value for the previous iteration. MLE analysis shows the sample distribution lies between 25 ± 10 nm to 250 ± 10 nm for the first video and between 50 ± 10 nm to 200 ± 10 nm for the second and third video. The number of iterations for MLE analyses are 5, 4 and 4 respectively for a, b and c. The variance of the noise in measuring step size in the experimental setup is set to zero.

6 | Conclusion and Future Scope

In this research work, an improvement in the particle size distribution obtained by NTA has been presented with the help of the Brownian motion simulation and the probabilistic MLE approach.

In Chapter 3, the method of Brownian motion simulation is presented along with the data processing method for MLE. It can be observed that Brownian motion simulation is very useful in mimicking random walks of the Brownian particle and therefore, it is a very useful tool to mimic the experimental conditions of NTA. It can also be seen from some of the figures in 3.2 that it is easily possible to obtain different types of particle distributions by this simulation. Also, as the conventional approach of NTA is subject to statistical uncertainty, the new approach, Maximum Likelihood Estimate is presented in detail further in the chapter. Unlike the conventional approach, the Maximum Likelihood Estimate approach takes into account the finite accuracy of the mean squared displacement along with the average number of steps for each track to solve this statistical uncertainty and the apparent broadened size distribution obtained by the conventional method due to this uncertainty.

Chapter 4 and Chapter 5 present the results of this MLE approach applied to the particle size distributions obtained by the simulation and by the actual experimental videos. In Chapter 4 the comparison is shown between the distributions obtained by the conventional 'Einstein-Stokes' and 'Einstein-Stokes+MLE' approaches. This comparison is done for the monodispersed and bidispersed distributions and it has been observed that 'Einstein-Stokes+MLE' was quickly able to converge to the original distribution and for the lesser value of the mean number of steps than the 'Einstein-Stokes' approach for both the cases. Further, to compare

the broadness of the distribution obtained by both the methods, Gaussian fitting is done and it can also be observed that the widths of the 'Einstein-Stokes+MLE' distributions are narrower even for the distributions obtained with the mean number of steps as small as 10 and was able to quickly converge to the Gaussian centre or mean value which is defined by the original distribution. In chapter 5, the MLE approach was able to give a narrower size distribution for TiO_2 samples but was slightly inaccurate in determining the actual particle sizes. Therefore, it can be concluded that although the MLE method can be very useful to determine the accurate size distribution obtained by NTA and can be used to remove the inherent uncertainties in the conventional approach but has some weaknesses such as bias or overfitting if the stopping criterion is not applied which was also shown at the end of Chapter 4. All the existing publications on MLE such as Walker [18] truncated the iterations without giving any proper justifications of why they truncated these iterations. So based on the results obtained with the simulations, iterations should be truncated at 40 to get results similar to the actual distribution. Results found in chapter 5 are also subject to additional experimental and numerical uncertainties such as misidentification of particles and/or possible errors in drift corrections. Thus the accuracy of the MLE result shown for the experimental data is also affected by these uncertainties. It can also be observed that the MLE distribution is biased towards the peaks appearing at the smaller sized particles in the size distribution. This bias is mostly dependent on how accurately the mean of the initial exponential distribution θ_r (refer to subsection 3.3.1) is calculated. Future work should focus on the investigation of the possible effect of the known experimental artefacts on the MLE accuracy by simulation and how accurately the initial parameters need to be calculated for the unbiased distribution.

MLE is a useful technique for the determination of particle size distribution but as mentioned earlier there is an inherent disadvantage of overfitting or oversmooth-

ing and a stopping criterion is necessary to stop the iterations. Also, as evidenced in Chapter 5, the accuracy of the size distribution can also be dependent upon how accurately the initial parameters such as the mean of the initial exponential distribution (θ_r) are calculated. The alternate approach Maximum A posteriori Estimation [20], an approach based on the Bayesian probability that takes into account the prior information can be used and tested to overcome these disadvantages. This approach also takes into account the data like mean trajectory length or the mean squared displacement but unlike MLE, has better regularization approaches such as cross-validation and won't be dependent on the statistical stopping choices for the iterations.

A | Solution of the Diffusion Equation

To find the solution of the diffusion equation given by equation 2.7, we will place the assembly of the particle at the origin for convenience i.e. at $x = 0$. Then our problem will be to solve this equation with the delta function $x = 0$ as our initial condition

$$f(x, 0) = \delta(x) \tag{A.1}$$

In general terms, Fourier Transform is given as

$$f(x, t) = \int_{-\infty}^{\infty} F(u, t)e^{-iux}du \tag{A.2}$$

Also the Inverse Fourier Transform is given as,

$$F(u, t) = \int_{-\infty}^{\infty} f(x, t)e^{iux}dx \tag{A.3}$$

From equation 2.7,

$$\frac{\partial}{\partial t}\left[\int_{-\infty}^{\infty} F(u, t)e^{iux}du\right] = D\frac{\partial^2}{\partial x^2}\left[\int_{-\infty}^{\infty} F(u, t)e^{iux}du\right] \tag{A.4}$$

Therefore,

$$\int_{-\infty}^{\infty} \frac{\partial}{\partial t}\left[F(u, t)e^{iux}du\right] = D\int_{-\infty}^{\infty} \frac{\partial^2}{\partial x^2}\left[F(u, t)e^{iux}du\right] \tag{A.5}$$

Now we can remove the integration sign from the equation A.5 because we are taking partial differentiation with respect to t and x, Therefore,

$$\frac{\partial}{\partial t}\left[F(u, t)\right]e^{iux} = D\frac{\partial^2}{\partial x^2}\left[F(u, t)e^{iux}\right] \tag{A.6}$$

Therefore,

$$\frac{\partial}{\partial t}\left[F(u,t)\right]e^{iux} = DF(u,t)\frac{\partial^2}{\partial x^2}\left[e^{iux}\right]$$

$$\frac{\partial}{\partial t}\left[F(u,t)\right]e^{iux} = DF(u,t)(i)(i)u^2e^{iux}$$

So we will get,

$$\frac{\partial}{\partial t}\left[F(u,t)\right] = -Du^2F(u,t) \tag{A.7}$$

Dividing both sides by $F(u,t)$ on both sides of the equation A.7,

$$\frac{1}{F(u,t)}\frac{\partial}{\partial t}\left[F(u,t)\right] = -Du^2 \tag{A.8}$$

Rearranging the equation to remove the partial dependence and integrating both sides,

$$\int \frac{1}{F(u,t)}dF(u,t) = -Du^2\int dt$$

The solution of above equation can be given as,

$$lnF(u,t) = -Du^2t + C \tag{A.9}$$

At $t = 0$, equation A.9 becomes,

$$lnF(u,0) = C \tag{A.10}$$

From equation A.3,

$$F(u,0) = \int_{-\infty}^{\infty} f(x,0)e^{iux}dx$$

And from equation A.1,

$$F(u,0) = \int_{-\infty}^{\infty} \delta(x) e^{iux} dx \tag{A.11}$$

We know the delta function property,

$$\int_{-\infty}^{\infty} \delta(x) f(x) dx = f(0) \tag{A.12}$$

Therefore,

$$\int_{-\infty}^{\infty} \delta(x) e^{iux} dx = e^0 = 1 \tag{A.13}$$

Therefore, equation A.11 becomes,

$$F(u,0) = \int_{-\infty}^{\infty} \delta(x) e^{iux} dx = 1 \tag{A.14}$$

Hence, from equation A.10,

$$C = lnF(u,0) = ln1 = 0 \tag{A.15}$$

And therefore, equation A.8 becomes,

$$lnF(u,t) = -Du^2 t$$

So $F(u,t)$ is given as,

$$F(u,t) = e^{-Du^2 t} \tag{A.16}$$

Now considering the general alternate form of the Fourier transformation,

$$f(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} F(u,t)e^{-iux}du \tag{A.17}$$

Substituting the value of $F(u,t)$ from equation A.16 to above equation,

$$f(x,t) = \frac{1}{2\pi}\int_{-\infty}^{\infty} e^{-Du^2t}e^{-iux}du \tag{A.18}$$

Let's consider,

$$\int_{-\infty}^{\infty} e^{-Du^2t}e^{-iux}du = \int_{-\infty}^{\infty} e^{-(au^2+bu)}du \tag{A.19}$$

where, $a = Dt$ and $b = ix$

To solve the integral in the equation A.17, we will first solve the integral $\int_{-\infty}^{\infty} e^{-u^2}du$ Let's consider,

$$I = \int_{-\infty}^{\infty} e^{-u^2}du \tag{A.20}$$

By squaring the above integral we get,

$$I^2 = \int_{-\infty}^{\infty} e^{-u^2}du \int_{-\infty}^{\infty} e^{-u^2}du$$

or we can also write it as,

$$I^2 = \int_{-\infty}^{\infty} e^{-v^2}dv \int_{-\infty}^{\infty} e^{-u^2}du$$

Therefore,

$$I^2 = \int_{-\infty}^{\infty} e^{-(u^2+v^2)}dv \int_{-\infty}^{\infty} du$$

Changing into polar coordinates,

$$I^2 = \int_0^\infty re^{-r^2}dr \int_0^{2\pi} d\theta$$

$$I^2 = 2\pi \int_0^\infty \frac{1}{2}d(r^2)e^{-r^2}dr = \pi$$

Therefore,

$$I = \sqrt{\pi} \tag{A.21}$$

Now, for a real constant $a > 0$, the simple change of variable gives us,

$$I(a) = \int e^{-au^2}du = \frac{1}{\sqrt{a}} \int e^{-(\sqrt{a}u)^2}d(\sqrt{a}u) = \sqrt{\frac{\pi}{a}} \tag{A.22}$$

To solve the equation of the required form we will complete the square and solve the integration using similar change of variables,

$$\int_{-\infty}^{\infty} e^{-(au^2+bu)}du = \int_{-\infty}^{\infty} e^{[-a(u^2-2u\frac{b}{2a}+\frac{b^2}{4a^2})+\frac{b^2}{4a}]}du$$

$$\int_{-\infty}^{\infty} e^{-(au^2+bu)}du = \int_{-\infty}^{\infty} e^{-a(u-\frac{b}{2a})^2}e^{\frac{b^2}{4a}}du \tag{A.23}$$

From equation A.22, we can write equation A.23 as,

$$\int_{-\infty}^{\infty} e^{-(au^2+bu)}du = \sqrt{\frac{\pi}{a}}\exp{(\frac{b^2}{4a})} \tag{A.24}$$

Therefore, from equation A.19,

$$\int_{-\infty}^{\infty} e^{-Du^2t}e^{-iux}du = \sqrt{\frac{\pi}{Dt}}\exp{(\frac{i^2x^2}{4Dt})} \tag{A.25}$$

B | Probability Distributions

B.1 Random variable

A random event is an event where more than one outcomes are possible. As the outcome of the event itself is not predictable, we can only predict the probabilities of all the possible outcomes. Variables associated with this event is called Random variables. Let's denote the random variable by α then this random variable can take different possible numerical values such as $\alpha_1, \alpha_2, \alpha_3$ etc. corresponding to different possible events. The corresponding probabilities $P(\alpha_1), P(\alpha_2), P(\alpha_3)$ then form a probability distribution. There are two types of random variables: one is Discrete Random variables which are associated with a countable number of values like the number of throws of dice and corresponding probability distribution for these variables are called as Probability Mass Function. Some examples of discrete distributions are: Binomial Distribution, Poisson Distribution, Bernoulli Distribution etc.

The other types of random variables are called Continuous Random variable which takes all values between the given intervals such as age, height, weight, temperature etc. In Principle, these variables are continuous, only the limitations of our measuring instruments make it discrete or sometimes very finely divided [68]. The probability distribution of continuous random variables is described by Probability Density Function or PDF. Some examples of continuous distributions and their details are given in subsection B.3.

B.2 Probability Density Functions

Probability Density Function (PDF) or density of a continuous variable is the basic building block of statistical estimations.

Formal definition: If α is a continuous random variable, then a probability distribution or probability density function (PDF) of α is a function f(α) such that for any two numbers a and b with

$$P(a \leq \alpha \leq b) = \int_a^b f(\alpha) d\alpha \tag{B.1}$$

[68]. The probability of α taking value in the interval between a and b is the area above that interval as shown in Figure B.1.

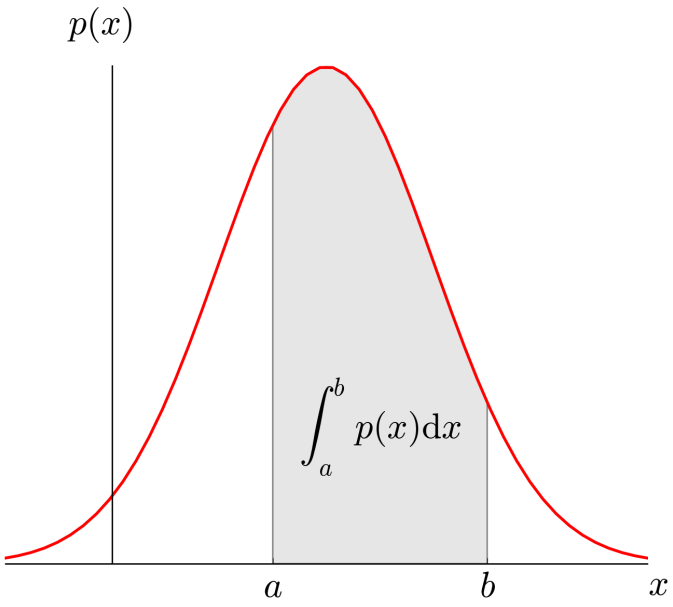


Figure B.1: A representation of the sample probability distribution or density curve where area under the curve between intervals a and b is $P(a \leq x \leq b)$. Here, x is considered as a random variable. Diagram taken from [69].

It is the relationship between the outcomes of a random variable and their prob-

ability meaning it can be used to specify the probability of the random variable falling within a particular range of values. The PDF is the density of probability which is very similar to the concept of mass density. As PDF gives the counts of the number of occurrences of a value in a particular range it is basically a normalized histogram. (A normalized histogram is a histogram obtained by dividing the number of counts in a particular bin by the total number of observations and multiplying it by bin width.)

The shape of this probability density function across the domain for the random variable is designated as a probability distribution. Knowing which kind of distribution is we can determine the metrics or the parameters of the distribution. Parameters are the descriptive measures of the entire sample and can be used as the inputs in the PDF to generate the distributions. Parameters are usually denoted by Greek letters. For example, for the Normal or Gaussian distribution population mean is denoted by μ and the population variance is denoted by σ. Parameters are fixed constants and do not vary with a sample. The main problem is usually for the random variable, the parameters of the probability distribution are unknown as we normally do not know all the possible outcomes and just have the sample of observations. But we can estimate the probability distribution from the sample of observations we have which is referred to as 'density estimation'.

B.3 Some important Probability Distributions and their significance

B.3.1 Continuous Uniform Distribution

The Continuous Uniform Distribution is the simplest distribution from all the distributions. The Uniform distribution is used where the probabilities of all the

event are equally likely possible. The continuous random variable α is uniformly distributed if PDF of the distribution is given by

$$f(\alpha) = \begin{cases} \frac{1}{b-a}, & \text{if } a \leq \alpha \leq b \\ 0, & \text{otherwise} \end{cases} \tag{B.2}$$

Parameters- a, b where $a < b$

Expected value- $E(\alpha) = \frac{a+b}{2}$

Variance- $V(\alpha) = \frac{(b-a)^2}{12}$

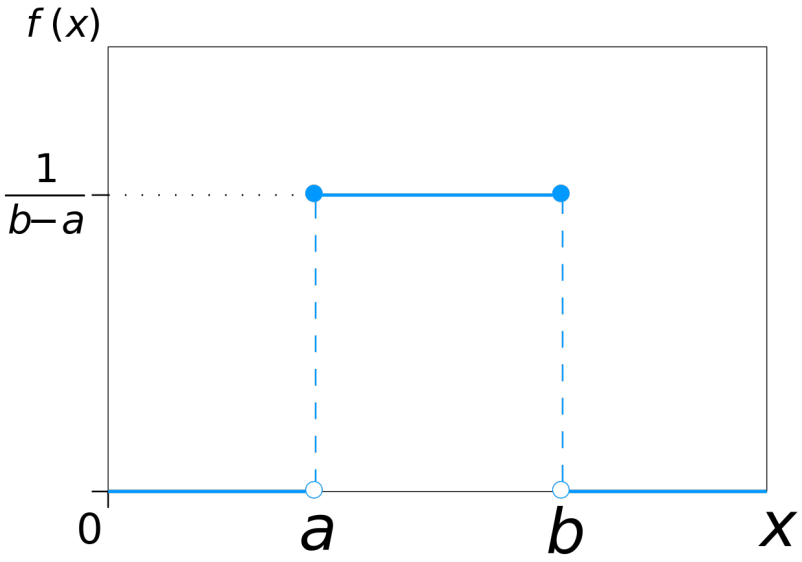


Figure B.2: Diagram showing the Continuous Uniform distribution where f(x) is the probability density and a and b are the parameters of the PDF. Here, X is considered as a random variable Diagram taken from [70]

B.3.2 Normal or Gaussian Distribution

The normal distribution is the most important distribution in statistics. Many distributions for numerical populations can be fitted very closely with Gaussian or

Normal distribution. Sometimes even if individual random variables are not normally distributed, under suitable conditions sums and averages of the variables can be normally distributed which is the basic postulate of the Central Limit Theorem. The normal distribution is defined by two parameters mean (μ), which gives the central tendency or the location of the peak and standard deviation (σ) which gives the width or spread of the distribution from its mean. The PDF of this distribution is given as

$$f(\alpha) = N(\mu, \sigma^2) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(\frac{-(\alpha-\mu)^2}{2\sigma^2}\right), -\infty < x < \infty \tag{B.3}$$

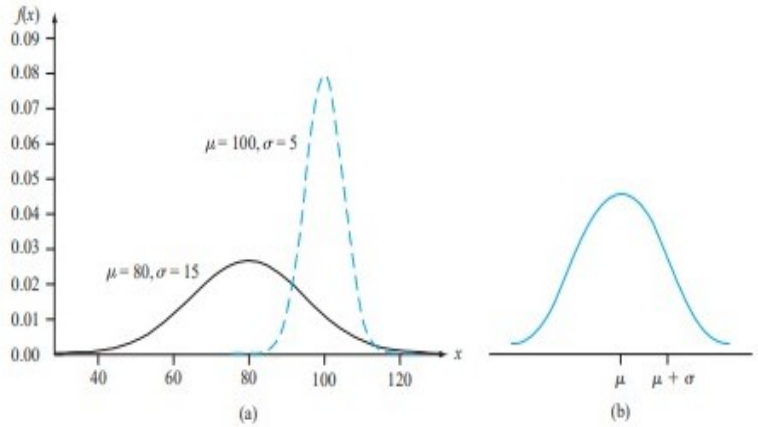


Figure B.3: Figure a shows two different Gaussian distribution curves one for the mean value of 80 and standard deviation of 15 and another for mean value of 100 and standard deviation of 5. Figure b helps to visualize the mean and standard deviation for the typical Gaussian or Normal distribution. Diagrams taken from [68]

Parameters- μ- real, σ- real positive number

Expected Value- $E(\alpha) = \mu$

Variance- $V(\alpha) = \sigma^2$

B.3.3 Exponential Distribution

The exponential distribution is used in the probability models to determine the waiting time for the next event to occur if events happen independently and randomly with a constant rate over time. For example, exponential distribution can be used to predict the time until next bus arrives or to predict the failure of certain hardware. The random variable αhas an exponential distribution when,

$$f(\alpha;\lambda) = \begin{cases} \lambda e^{-\lambda\alpha}, & \text{if } x \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{B.4}$$

where, λ gives rate of events. The exponential distribution has a memoryless property [71] meaning even if you have waited for some time for the successful event to happen, the mean waiting time for the next successful event is the same as when you started. Because all the events are independent of each other. For example, if you know the average waiting time for the customer to arrive in a certain store is 15 minutes which is given by exponential distribution. Some customers can arrive in 10 minutes or some customers can arrive in 20 minutes. But memory-less property means average customer waiting time will always remain 15 minutes independent of when the last customer arrived as the event of a customer arriving in the store is random and independent.

Parameters- λ Positive real number

Expected Value-$E(\alpha) = \dfrac{1}{\lambda}$

Variance-$V(\alpha) = \dfrac{1}{\lambda^2}$

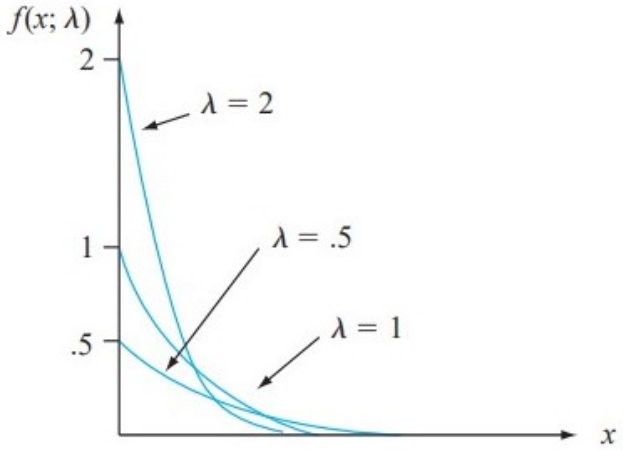


Figure B.4: Figure shows different exponential distributions for the parameters $\lambda = 0.5$, 1 and 2. $f(x;\lambda)$ is the probability density. Here, x is considered as a random variable Diagram taken from [68].

B.3.4 Gamma Distribution

A continuous random variable α will have a Gamma distribution if the PDF of α is given as

$$f(\alpha; vu) = \begin{cases} \dfrac{1}{v^u \Gamma(u)} \alpha^{(u-1)} e^{\frac{-\alpha}{v}}, & \text{if } \alpha \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{B.5}$$

Gamma distribution is also used to predict wait time until the future event but unlike exponential distribution which is used to predict the wait time of the very first event, Gamma distribution gives us the wait time until the k^{th} event occurs.

Gamma distribution has two different parametrization sets- u -shape, v -rate and k-shape and θ-scale. u, v parametrization is same as the k, λ parametrization where, k is the number of events and λ is the rate of events but for k, θ parametrization θ is given by $\dfrac{1}{\lambda}$ which is the mean wait time or the average time between the

intervals. The PDF for the k, θ parametrization is given as

$$f(\alpha; k\theta) = \begin{cases} \frac{1}{\theta^k \Gamma(k)} \alpha^{(k-1)} e^{\left(\frac{-\alpha}{\theta}\right)}, & \text{if } \alpha \geq 0 \\ 0, & \text{otherwise} \end{cases} \tag{B.6}$$

Parameters- $u\ v$, $k\ \theta$

Expected Value- $E(\alpha) = k\theta$ or $E(\alpha) = \frac{u}{v}$

Variance- $V(\alpha) = k\theta^2$ or $V(\alpha) = \frac{u}{v^2}$

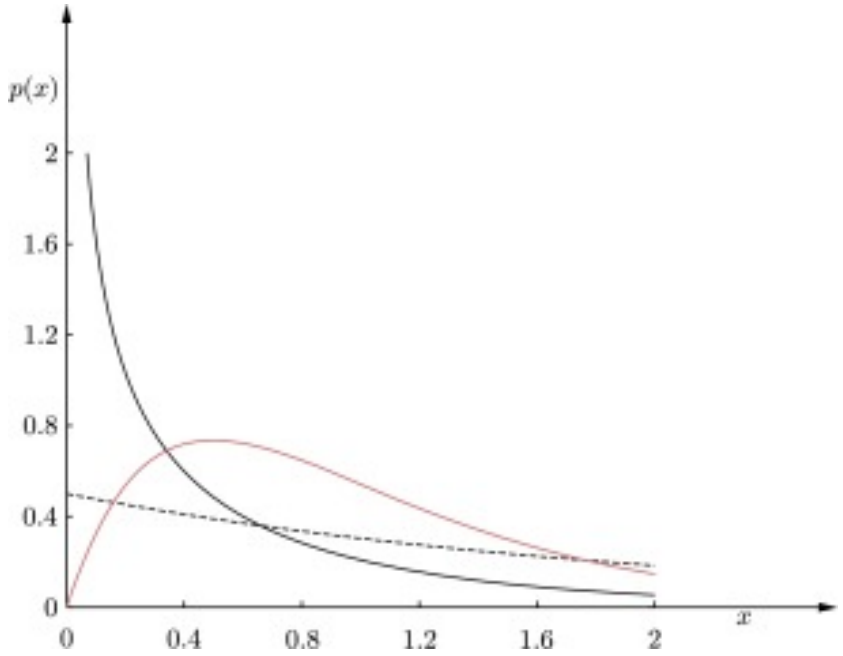


Figure B.5: The diagram shows the PDF for the gamma distribution with different parameters: u=0.5, v=1 (full line gray), u=2, v=0.5 (red), u=1,v=2 (dotted). Here x is considered as a random variable. Diagram taken from [72]

C | MATLAB code for the MLE program

```
1
2  %
3  % This program first defines an initial probability distribution
4  % for the sizes of the particles whose random walk is simulated
5  % in 2D projection and the mean displacement calculated together
6  % with the total number of steps tracked par particles.
7  % The mean displacement is used to recover an estimate of the
8  % particle radius using Einstein-Stokes relation.
9  %
10 % Alternatively, the means and number of steps tracked are used
11 % together and over a large number of particles to obtain a
12 % maximum entropy estimate of the particle radius distribution
13 % that has the same means as the first estimate, but maximize
14 % the entropy with the knowledge of the number of steps tracked
15 %  for each estimate of the mean radius of the
16 % particls.
17 %
18 clear;
19 %
20 % The default setting is that particles are tracked on average
21 % for 50 consecutive frames (steps) with a step variance of 3.
22 %
23 step.mean=50;
24 step.variance=3;
25 %
26 % The particle size distribution(the probability distribution)
27 % is defined, by default, to be bimodel centered at 25 and 70
28 %
29 pm.r=
30 [10,15,20,25,30,35,40,45,50,55,60,65,70,75,80,85,90,95,100,105];
```

```
pm.f=[0,0,0,1,0,0,0,0,0,0,0,0,1,0,0,0,0,0,0,0];
%
%The next block of the code is to get the allowed values of
%radius from the defined radius histogram above.
%
pm.allowed_r1=zeros(length(pm.r),1);
pm.allowed_f1=zeros(length(pm.r),1);
for j=1:length(pm.r)
if((pm.r(j)).*(pm.f(j))~=0)
    pm.allowed_r1(j)=pm.r(j);
    pm.allowed_f1(j)=pm.f(j);
end
end
pm.zero_values_r=find(any(pm.allowed_r1==0,2));
pm.zero_values_f=find(any(pm.allowed_f1==0,2));
pm.allowed_r1(pm.zero_values_r,:)=[];
pm.allowed_f1(pm.zero_values_f,:)=[];
pm.allowed_r_init=cat(2,pm.allowed_r1,pm.allowed_f1);


   for u=1:length(pm.allowed_r_init(:,2))
     if pm.allowed_r_init(u,2)~=1
     p(u).allowed_r = repelem(pm.allowed_r_init(u,1),pm.
    allowed_r_init(u,2));
     else
     p(u).allowed_r=pm.allowed_r_init(u,1);
     end
   end
  %repelem- repeat copies of array elements

allowed_r=cat(2,p.allowed_r);
pm.radius=datasample(allowed_r,1000); % random sampling
pm.radius=pm.radius';
```

```
%
% Now we have a random set of (1000) particles radius
% (pm.radius(1:1000))in accorance of the size distribution
% defined in [pm.r, pm.f].
%
%
for j=1:length(pm.r)
    theta_r(j)=(2*1.38*10^-23*0.1*300)/(3*pi*0.000853*pm.r(j));
end

D=stokes_einstein_D(300,pm.radius,0.00085);
%Define D as a function given temperature,radius,viscosity.
%particle.mean_d=sqrt(2*D*0.1);
%k=sqrt(2*mean_D*0.1);%scaling factor
%
% In the following, the random walk is simulated for a
% normalized distribution of unit variance. To convert
% this to one for a particle of radius r, the variance
% is multiplied by sqrt(mean displacement estimated
% by the Einstein relationship as calculated in the
% above few lines of the code.
%
particlecount=length(D);
delta_t=0.1;
dummy=zeros(particlecount,1);
rng('shuffle');
%meandisp=0.0;

particle = struct('steps',dummy,'meandisp',dummy);
for i=1:particlecount
particle.steps=round(step.mean + step.variance.*rand(
   particlecount,1));
%This command is used to generate random no. of steps for
```

```
95 %particle with given mean and variance of steps.
96 particle.dx= sqrt(2*D(i)*delta_t)*randn(1,particle.steps(i));%
       Scaling
97 particle.dy= sqrt(2*D(i)*delta_t)*randn(1,particle.steps(i));%
       Scaling
98 particle.rsquared = (particle.dx) .^2 + (particle.dy) .^ 2;
99 particle.meandisp(i)=mean(particle.rsquared);
100 particle.radius(i)=(2*1.38*10^-23*0.1*300)/(3*pi*0.000853*
       particle.meandisp(i));
101 end
102 %
103 % The key data (the number of steps, the mean square
104 % displacement) for each particles are collected in
105 % the data 'sim_data', saved to a csv file and
106 % plotted in a histogram with a bin width set to
107 % 0.6*10^-21.
108 %
109 sim_data=cat(2,particle.steps,particle.meandisp);
110 csvwrite("Sim_Data.csv",sim_data);
111
112 [N,edges] = histcounts(particle.meandisp,'BinWidth',0.6*10^-21);
113
114 Num_Par=unique(sim_data(:,1),'rows');
115 counts_1 = hist(sim_data(:,1),Num_Par);
116 %
117 %  Initiation of the 1st guess of the true particle
118 %  size distribution as P_m_1. Here the guess is a
119 %  uniform distribution.
120 %
121 b=max(pm.r);
122 a=min(pm.r);
123 P_m_1=pdf('Uniform',pm.r,a,b);
124 %
```

```
% Calculation of the Gamma distribution for probability of n
% steps of random motions.
%
M=cell(particlecount,1);%To store values in a sample cell
for i=1:particlecount
x(i)=sim_data(i,1)*sim_data(i,2);%k_n*z_n
M{i}=gampdf(x(i),sim_data(i,1),theta_r);%Gammapdf
end
%
% Declare matrix to store successive estimate of particle size
   distribution.
%
First_cell=cell(particlecount,1);
last_cell=cell(particlecount,1);
P_m_=cell(5000,1);
P_m_{1}=P_m_1;%Defined uniform pd
Sum(1)=sum(P_m_{1}, 'all');
for i=1:particlecount
First_cell{i}=M{i}.*P_m_{1}/Sum(1);%P_d*P_m(k)/Sum of P_m(k)
Denominator(i)=sum(First_cell{i},'all');
%Sum over all bins for the term P_d*P_m(k)/Sum of P_m(k)
last_cell{i}=M{i}/Denominator(i);
%P_d divided by the term in the denominator
end
A_1 = cell2mat(last_cell);%Conversion of cell into single matrix
final_sum=sum(A_1,1);
%Sum of the individual columns as the sum is over all the
   particles
%i.e.N(Sum over all the particles)

final_term=final_sum/particlecount;%Sum divided by no. of
   particles(N)
P_m_{2}=P_m_{1}.*final_term;
```

```
155
156 for j=1:length(N)
157   for l=1:length(Num_Par)
158   MLE_hist{l}=gampdf((0.6*10^-21)*j,Num_Par(l,1),theta_r);
159   MLE_Hist_Num{l}=MLE_hist{l}.*P_m_{1}*(0.6*10^-21);
160   Sum_HIST(l)=sum(MLE_Hist_Num{l},'all');
161   Fin_HIST(l)=Sum_HIST(l)/Sum(1);
162   end
163 Sum_step=counts_1.*Fin_HIST;
164 H_ML(j,1)=sum(Sum_step,'all');
165 Chi_init(j,1)=((0.6*10^-21)-H_ML(j,1))^2/H_ML(j,1);
166 end
167 Chi_FIN(1,1)=sum(Chi_init,'all');
168
169     for k=2:5000
170     Sum(k)=sum(P_m_{k}, 'all');%Sum of P_m(k)
171 for i=1:particlecount
172 First_cell{i}=M{i}.*P_m_{k}/Sum(k);%P_d*P_m(k)/Sum of P_m(k)
173 Denominator(i)=sum(First_cell{i},'all');
174 %Sum over all bins for the term P_d*P_m(k)/Sum of P_m(k)
175 last_cell{i}=M{i}/Denominator(i);%P_d divided by the term in the
      denominator
176 end
177 A_1 = cell2mat(last_cell);%Conversion of cell into single matrix
178 final_sum=sum(A_1,1);
179 %Sum of the individual columns as the sum is over all the
    particles
180 %i.e.N(Sum over all the particles)
181
182 final_term=final_sum/particlecount;%Sum divided by no. of
    particles(N)
183 P_m_{k+1}=P_m_{k}.*final_term;%Iterations
184
```

```
%Convergence calculation
for j=1:length(N)
  for l=1:length(Num_Par)
  MLE_hist{l}=gampdf((0.6*10^-21)*j,Num_Par(l,1),theta_r);
  MLE_Hist_Num{l}=MLE_hist{l}.*P_m_{k}*(0.6*10^-21);
  Sum_HIST(l)=sum(MLE_Hist_Num{l},'all');
  Fin_HIST(l)=Sum_HIST(l)/Sum(k);
  end
Sum_step=counts_1.*Fin_HIST;
H_ML(j,1)=sum(Sum_step,'all');
Chi_init(j,1)=((0.6*10^-21)-H_ML(j,1))^2/H_ML(j,1);
end
Chi_FIN(k,1)=sum(Chi_init,'all');
if (Chi_FIN(k,1)-Chi_FIN(k-1,1))<(0.01*Chi_FIN(k-1,1))
    break
end
    end
%
% Normalization of the updated particle size distribution
hold on
pm.f = pm.f / sum(pm.f(:));%normalize data
P_m_{k}= P_m_{k}/ sum(P_m_{k}(:));%normalize data
[N1,centers] = hist(particle.radius,pm.r);
N1=N1/sum(N1(:));
pm.r=pm.r';
pm.f=pm.f';
N1=N1';
P_m_{k}=P_m_{k}';
plot_data_y=cat(2,pm.f,N1,P_m_{k});
fin_plot_data=cat(2,pm.r,pm.f,N1,P_m_{k});
csvwrite("plot_Data50_1000.csv",fin_plot_data);%Create CSV file
   for plot data.
%
```

```
% Plotting of three curves within one plots.
bar(pm.r,plot_data_y)
legend('Original','ES','ES+MLE')
xlabel('Radius(nm)');
ylabel('Size Distribution Function');
box on;


function output = stokes_einstein_D(temp,radius,viscosity)
%applies stokes-einstein relationship to determine mean
   hydrodynamic radius
boltzmann = 1.38*(10^(-23));
a=boltzmann*temp;
b=6*pi*viscosity.*radius;
output = a./b;
end
```

References

[1] *Definition - Nanomaterials - Environment - European Commission*. [Online]. Available: https://ec.europa.eu/environment/chemicals/nanotech/faq/definition_en.htm.

[2] I. Khan, K. Saeed, and I. Khan, "Nanoparticles: Properties, applications and toxicities," *Arabian Journal of Chemistry*, vol. 12, no. 7, pp. 908–931, Nov. 2019, ISSN: 18785352. DOI: 10.1016/j.arabjc.2017.05.011.

[3] X. Huang and M. A. El-Sayed, "Gold nanoparticles: Optical properties and implementations in cancer diagnosis and photothermal therapy," *Journal of Advanced Research*, vol. 1, no. 1, pp. 13–28, Jan. 2010, ISSN: 2090-1232. DOI: 10.1016/J.JARE.2010.02.002.

[4] J. K. Patra, G. Das, L. F. Fraceto, *et al.*, "Nano based drug delivery systems: recent developments and future prospects," *Journal of Nanobiotechnology 2018 16:1*, vol. 16, no. 1, pp. 1–33, Sep. 2018, ISSN: 1477-3155. DOI: 10.1186/S12951-018-0392-8.

[5] J. M. Zook, R. I. Maccuspie, L. E. Locascio, M. D. Halter, and J. T. Elliott, "Stable nanoparticle aggregates/agglomerates of different sizes and the effect of their size on hemolytic cytotoxicity," *Nanotoxicology*, vol. 5, no. 4, pp. 517–530, 2011. DOI: 10.3109/17435390.2010.536615.

[6] A. Kim, "Size distribution measurement of polydisperse macromolecular samples using nanoparticle tracking analysis," Ph.D. dissertation, Nanyang Technological University, 2019, pp. 18–18. DOI: 10.32657/10356/137145.

[7] D. Williams, "Measuring & Characterizing Nanoparticle Size – TEM vs SEM," *AZO NANO*, 2015. [Online]. Available: https://www.azonano.com/article.aspx?ArticleID=4118.

[8] *Dynamic Light Scattering - Particle Technology Labs*. [Online]. Available: https://www.particletechlabs.com/analytical-testing/particle-size-distribution-analyses/dynamic-light-scattering.

[9] J. Stetefeld, S. A. McKenna, and T. R. Patel, "Dynamic light scattering: a practical guide and applications in biomedical sciences," *Biophysical Reviews*, vol. 8, no. 4, p. 409, Dec. 2016, ISSN: 18672469. DOI: 10.1007/S12551-016-0218-6.

[10] V. Filipe, A. Hawe, and W. Jiskoot, "Critical Evaluation of Nanoparticle Tracking Analysis (NTA) by NanoSight for the Measurement of Nanoparticles and Protein Aggregates," *Pharmaceutical Research*, vol. 27, no. 5, p. 796, May 2010, ISSN: 07248741. DOI: 10.1007/S11095-010-0073-2.

[11] C. M. Maguire, M. Rösslein, P. Wick, and A. Prina-Mello, "Characterisation of particles in solution – a perspective on light scattering and comparative technologies," *Science and Technology of Advanced Materials*, vol. 19, no. 1, p. 732, Dec. 2018, ISSN: 18785514. DOI: 10.1080/14686996.2018.1517587.

[12] N. Farkas and J. A. Kramar, "Dynamic light scattering distributions by any means," *Journal of Nanoparticle Research*, vol. 23, no. 5, pp. 1–11, May 2021, ISSN: 1572896X. DOI: 10.1007/S11051-021-05220-6/FIGURES/4.

[13] A. Kim, W. B. Ng, W. Bernt, and N. J. Cho, "Validation of Size Estimation of Nanoparticle Tracking Analysis on Polydisperse Macromolecule Assembly," *Scientific Reports 2019 9:1*, vol. 9, no. 1, pp. 1–14, Feb. 2019, ISSN: 2045-2322. DOI: 10.1038/s41598-019-38915-x.

[14] R. A. Dragovic, C. Gardiner, A. S. Brooks, *et al.*, "Sizing and phenotyping of cellular vesicles using Nanoparticle Tracking Analysis," *Nanomedicine: Nanotechnology, Biology and Medicine*, vol. 7, no. 6, pp. 780–788, Dec. 2011, ISSN: 1549-9634. DOI: 10.1016/J.NANO.2011.04.003.

[15] J. Gross, S. Sayle, A. R. Karow, U. Bakowsky, and P. Garidel, "Nanoparticle tracking analysis of particle size and concentration detection in suspensions of polymer and protein samples: Influence of experimental and data evaluation parameters," *European Journal of Pharmaceutics and Biopharmaceutics*, vol. 104, pp. 30–41, Jul. 2016, ISSN: 18733441. DOI: 10.1016/j.ejpb.2016.04.013.

[16] D. T. Yang, X. Lu, Y. Fan, and R. M. Murphy, "Evaluation of nanoparticle tracking for characterization of fibrillar protein aggregates," *AIChE Journal*, vol. 60, no. 4, pp. 1236–1244, Apr. 2014, ISSN: 1547-5905. DOI: 10.1002/AIC.14349.

[17] S. Thompson, "Optical Approaches to Characterising Engineered Nanoparticles for Size and Shape in Aquatic Systems," Ph.D. dissertation, University of York, 2017.

[18] J. G. Walker, "Improved nano-particle tracking analysis," *Measurement Science and Technology*, vol. 23, no. 6, p. 065 605, May 2012, ISSN: 13616501. DOI: 10.1088/0957-0233/23/6/065605.

[19] Y. Matsuura, N. Ouchi, A. Nakamura, and H. Kato, "Determination of an accurate size distribution of nanoparticles using particle tracking analysis corrected for the adverse effect of random Brownian motion," *Phys. Chem. Chem. Phys*, vol. 20, p. 17 839, 2018. DOI: 10.1039/c7cp08332g.

[20] K. S. Silmore, X. Gong, M. S. Strano, and J. W. Swan, "High-Resolution Nanoparticle Sizing with Maximum A Posteriori Nanoparticle Tracking

Analysis," *ACS nano*, vol. 13, no. 4, pp. 3940–3952, Apr. 2019, ISSN: 1936-086X. DOI: 10.1021/ACSNANO.8B07215.

[21] R. B. Millar, *Maximum Likelihood Estimation and Inference: With Examples in R, SAS and ADMB*. Chichester, Sussex: Wiley, Jul. 2011, pp. 1–357, ISBN: 9780470094846. DOI: 10.1002/9780470094846.

[22] R. Brown F.R.S. Hon. M.R.S.E. & R.I. Acad. V.P.L.S., " XXVII. A brief account of microscopical observations made in the months of June, July and August 1827, on the particles contained in the pollen of plants; and on the general existence of active molecules in organic and inorganic bodies," *The Philosophical Magazine*, vol. 4, no. 21, pp. 161–173, 1828. DOI: 10.1080/14786442808674769.

[23] W. T. Coffey, Y. P. Kalmykov, and J. T. Waldron, *The Langevin Equation*, 2nd ed., ser. World Scientific Series in Contemporary Chemical Physics. WORLD SCIENTIFIC, Mar. 2004, vol. 14, pp. 1–168, ISBN: 978-981-238-462-1. DOI: 10.1142/5343.

[24] Gouy, "Note sur le mouvement brownien," *Journal de Physique Théorique et Appliquée*, vol. 7, no. 1, pp. 561–564, 1888, ISSN: 0368-3893. DOI: 10.1051/JPHYSTAP:018880070056101.

[25] E. W. Woolard, A. Einstein, R. Furth, and A. D. Cowper, "Investigations on the Theory of the Brownian Movement," *The American Mathematical Monthly*, vol. 35, no. 6, p. 318, Jun. 1928, ISSN: 00029890. DOI: 10.2307/2298685.

[26] M. von Smoluchowski, "Zur kinetischen Theorie der Brownschen Molekularbewegung und der Suspensionen," *Annalen der Physik*, vol. 326, no. 14, pp. 756–780, 1906, ISSN: 15213889. DOI: 10.1002/ANDP.19063261405.

[27] D. S. Lemons and A. Gythiel, "Paul Langevin's 1908 paper "On the Theory of Brownian Motion" ["Sur la théorie du mouvement brownien," C. R. Acad. Sci. (Paris) 146, 530–533 (1908)]," *American Journal of Physics*, vol. 65, no. 11, p. 1079, Jun. 1998, ISSN: 0002-9505. DOI: 10.1119/1.18725.

[28] M. D. Haw, "Colloidal suspensions, Brownian motion, molecular reality: a short history," *Journal of Physics: Condensed Matter*, vol. 14, no. 33, p. 7769, Aug. 2002, ISSN: 0953-8984. DOI: 10.1088/0953-8984/14/33/315.

[29] J. Perrin, "Mouvement brownien et réalité moléculaire," *Annales de chimie et de physique*, vol. 18, pp. 5–114, 1909.

[30] B. Peters, *Reaction Rate Theory and Rare Events Simulations*, 1st ed. Elsevier, 2017, pp. 129–145, ISBN: 978-0-444-56349-1.

[31] W. Coffey, "Development and Application of the Theory of Brownian Motion," *Advances in Chemical Physics*, vol. 63, pp. 69–252, Dec. 2006. DOI: 10.1002/9780470142875.CH2.

[32] E. Nelson, *Dynamical Theories of Brownian Motion*. Princeton, NJ: Princeton University Press, 1967, p. 13. DOI: 10.1515/9780691219615.

[33] D. D. Vvedensky, "Brownian motion, random walks, and the diffusion equation," *Transformations of Materials*, 2019. DOI: 10.1088/2053-2571/AB191ECH2.

[34] P. Hole, "Particle Tracking Analysis (PTA)," *Characterization of Nanoparticles: Measurement Processes for Nanoparticles*, pp. 79–96, Jan. 2020. DOI: 10.1016/B978-0-12-814182-3.00007-9.

[35] M. Instruments, *NANOSIGHT NS300 OPERATING MANUAL*, 2015. [Online]. Available: https://www.malvernpanalytical.com/en/learn/knowledge-center/user-manuals/man0541en.

[36] P. Metrix, *Introduction to Nanoparticle Tracking Analysis (NTA) Measurement Principle of ZetaView®*. [Online]. Available: www.particle-metrix.de.

[37] M. Panalytical, *Nanoparticle Tracking Analysis for the Measurement of Environmental Impact of Nanomaterial Wastes and Contaminants*, 2015. [Online]. Available: https://www.azonano.com/article.aspx?ArticleID=4075.

[38] J. C. Crocker and D. G. Grier, "Methods of Digital Video Microscopy for Colloidal Studies," *Journal of Colloid and Interface Science*, vol. 179, no. 1, pp. 298–310, Apr. 1996, ISSN: 00219797. DOI: 10.1006/jcis.1996.0217.

[39] W. Jiskoot and D. J. A. Crommelin, *Methods for structural analysis of protein pharmaceuticals*. AAPS Press, 2005, p. 678, ISBN: 9780971176720.

[40] S. Klose, "The scattering of light by dielectric particles," *Astrophysics and Space Science 1986 125:1*, vol. 125, no. 1, pp. 157–167, Aug. 1986, ISSN: 1572-946X. DOI: 10.1007/BF00643980.

[41] D. B. Allan, T. Caswell, N. C. Keim, C. M. van der Wel, and R. W. Verweij, *Trackpy v0.5.0*, Apr. 2021. DOI: 10.5281/ZENODO.4682814. [Online]. Available: github.com/soft-matter/trackpy.

[42] D. Ershov, M. Phan, J. W. Pylvänäinen, *et al.*, "Bringing TrackMate into the era of machine-learning and deep-learning," *bioRxiv*, p. 2021.09.03.458852, Sep. 2021. DOI: 10.1101/2021.09.03.458852.

[43] B. H. Lee and H. Y. Park, "HybTrack: A hybrid single particle tracking software using manual and automatic detection of dim signals," *Scientific Reports 2017 8:1*, vol. 8, no. 1, pp. 1–7, Jan. 2018, ISSN: 2045-2322. DOI: 10.1038/s41598-017-18569-3.

[44] K. A. Rose, M. Molaei, M. J. Boyle, D. Lee, J. C. Crocker, and R. J. Composto, "Particle tracking of nanoparticles in soft matter," *Journal of Applied Physics*, vol. 127, no. 19, p. 191 101, May 2020, ISSN: 0021-8979. DOI: 10.1063/5.0003322.

[45] E. J. Fong, Y. Sharma, B. Fallica, D. B. Tierney, S. M. Fortune, and M. H. Zaman, "Decoupling directed and passive motion in dynamic systems: Particle tracking microrheology of sputum," *Annals of Biomedical Engineering*, vol. 41, no. 4, pp. 837–846, Apr. 2013, ISSN: 00906964. DOI: 10.1007/s10439-012-0721-2.

[46] R. J. Barlow, *Statistics : A Guide to the Use of Statistical Methods in the Physical Sciences*. Wiley, 1989, p. 204, ISBN: 978-0-471-92295-7.

[47] J. Brownlee, *A Gentle Introduction to Maximum Likelihood Estimation for Machine Learning*, Nov. 2019. [Online]. Available: https://machinelearningmastery.com/what-is-maximum-likelihood-estimation-in-machine-learning/.

[48] F. James, *Statistical methods in Experimental Physics*, 2nd ed. New Jersey: World Scientific, 2012, pp. 20–24, ISBN: 9789812567956.

[49] G. J. McLachlan and T. Krishnan, *The EM algorithm and extensions*. Wiley, 1997, p. 21, ISBN: 0-471-12358-7.

[50] K. P. Murphy, *Machine Learning: A Probabilistic Perspective*. Cambridge: MIT Press, 2012, pp. 350–353, ISBN: 9780262018029.

[51] J. Brownlee, *A Gentle Introduction to Expectation-Maximization (EM Algorithm)*, Aug. 2020. [Online]. Available: https://machinelearningmastery.com/expectation-maximization-em-algorithm/.

[52] M. N. Bernstein, *Expectation-maximization: theory and intuition*, May 2020. [Online]. Available: https://mbernste.github.io/posts/em/.

[53] S. Borman, "The Expectation Maximization Algorithm: A short tutorial," Jul. 2004. [Online]. Available: http://www.seanborman.com/publications/EM_algorithm.pdf.

[54] C. F. J. Wu, "On the Convergence Properties of the EM Algorithm," *Ann. Statist.*, vol. 11, no. 1, pp. 95–103, Mar. 1983, ISSN: 0090-5364. DOI: 10.1214/AOS/1176346060.

[55] M. Collins, "The EM Algorithm," *IN FULFILLMENT OF WRITTEN PRELIMINARY EXAM II REQUIREMENT*, 1997.

[56] S. Allua and C. B. Thompson, "Hypothesis Testing," *Air Medical Journal*, vol. 28, no. 3, pp. 108–153, May 2009, ISSN: 1067-991X. DOI: 10.1016/J.AMJ.2009.03.002.

[57] G. Snedecor and W. Cochran, *Statistical methods*, 8th ed. Iowa State University Press, 1989, p. 503, ISBN: 0813815614.

[58] *NIST/SEMATECH e-Handbook of Statistical Method, 1.3.5.15. Chi-Square Goodness-of-Fit Test*. [Online]. Available: https://www.itl.nist.gov/div898/handbook/eda/section3/eda35f.htm.

[59] *Chi-Square - Sociology 3112 - Department of Sociology - The University of utah*. [Online]. Available: https://soc.utah.edu/sociology3112/chi-square.php.

[60] M. Naiim, A. Boualem, C. Ferre, M. Jabloun, A. Jalocha, and P. Ravier, "Multiangle dynamic light scattering for the improvement of multimodal particle size distribution measurements," *Soft Matter*, vol. 11, no. 1, pp. 28–32, 2015, ISSN: 1744-6848. DOI: 10.1039/C4SM01995D.

[61] W. Greene, *Econometric analysis*, 6th ed. Upper Saddle River, N.J. : Pearson/Prentice Hall, Apr. 2008, p. 1068, ISBN: 9780273753568.

[62] Y. Vardi, L. A. Shepp, and L. Kaufman, "A Statistical Model for Positron Emission Tomography," *Journal of the American Statistical Association*, vol. 80, no. 389, p. 8, Mar. 1985, ISSN: 01621459. DOI: 10.2307/2288030.

[63] L. A. Shepp and Y. Vardi, "Maximum Likelihood Reconstruction for Emission Tomography," *IEEE Transactions on Medical Imaging*, vol. 1, no. 2, pp. 113–122, 1982, ISSN: 1558254X. DOI: 10.1109/TMI.1982.4307558.

[64] M. Ropers, H. Terrisse, M. Mercier-Bonin, and B. Humbert, "Titanium Dioxide as Food Additive," *Application of Titanium Dioxide*, Jul. 2017. DOI: 10.5772/INTECHOPEN.68883.

[65] L. Talamini, S. Gimondi, M. B. Violatto, *et al.*, "Repeated administration of the food additive E171 to mice results in accumulation in intestine and liver and promotes an inflammatory status," *Nanotoxicology*, vol. 13, no. 8, pp. 1087–1101, 2019, ISSN: 17435404. DOI: 10.1080/17435390.2019.1640910.

[66] I. Shah, "Characterisation of Food Grade TiO2 -CAPE 5010M: Research Project," University of Leeds, Tech. Rep., May 2020.

[67] Y. Yang, K. Doudrick, X. Bi, *et al.*, "Characterization of food-grade titanium dioxide: The presence of nanosized particles," *Environmental Science and Technology*, vol. 48, no. 11, pp. 6391–6400, Jun. 2014, ISSN: 15205851. DOI: https://doi.org/10.1021/es500436x.

[68] J. Devore, *Probability and Statistics for Engineering and the Sciences*, 8th ed. BROOKS/COLE CENGAGE Learning, 2009, pp. 137–192, ISBN: 0538733527.

[69] *Characterizing a Distribution — Introduction to Statistics 6.4 documentation*. [Online]. Available: http://work.thaslwanter.at/Stats/html/statsDistributions.html.

[70] *User:IkamusumeFan - Wikimedia Commons*. [Online]. Available: https://commons.wikimedia.org/wiki/User:IkamusumeFan.

[71] M. A. Pinsky and S. Karlin, *An Introduction to Stochastic Modeling*, 4th ed. Academic Press, 2010, pp. 1–42, ISBN: 9780123814166.

[72] S. Theodoridis, *Machine Learning: A Bayesian and Optimization Perspective*. Academic Press, 2015, pp. 9–51, ISBN: 9780128188033.