Analyses of Nucleation Rates from Molecular Dynamics Simulations. II. Weight Functions, Generation of Stochastic Times, and Realistic Uncertainties

E. Jean Jacob and Lawrence S. Bartell*

Department of Chemistry, University of Michigan, Ann Arbor, Michigan 48109

Received: October 16, 2002; In Final Form: December 20, 2002

In paper I of this series it was shown how to take transient nucleation into account in the spontaneous freezing of large clusters when deriving nucleation rates and time lags from sets of nucleation times. This required an estimate of the "reduced moment" characterizing the period of transient nucleation. Also, a procedure was sketched for constructing sets of stochastic times simulating nucleation times, for purposes of determining statistical uncertainties in the derived kinetic parameters of nucleation rates. In the present paper, a considerably more precise method for generating stochastic nucleation times is presented and an optimum weighting scheme for least squares analyses of nucleation rates and time lags is formulated. In the prior literature no suitable means had been established for estimating the reduced moment. Alternative ways to estimate this moment from nucleation data are discussed. It is found that the true expectation values of uncertainties, σ^{ϵ}, in rates and time lags are significantly larger than the uncertainties, σ^{ls}, derived from residuals in least squares analyses of individual sets of nucleation times. Although the elements of the least squares error matrix are lower for the optimum weight function than for the unit weights and arctangent weights used in prior analyses, the actual uncertainties do not depend strongly upon which weighting scheme is employed. The derived kinetic parameters do, however, depend appreciably upon the weighting, and results of the optimum weighting are preferred. A virtue of the analysis of simulated stochastic nucleation times is that it provides a valid measure of the *actual* uncertainties in derived nucleation parameters as well as the smaller, and therefore misleading, uncertainties inferred from a conventional error matrix. The analysis presented leads to guidelines conveying how large a set of nucleation times must be in order to provide meaningful determinations of nucleation rates and time lags. The new procedure also provides the first estimates of the uncertainty in reduced moments derived from sets of nucleation times, including their dependence on sample size.

Introduction

Molecular dynamics (MD) simulations with realistic potential functions are providing a fruitful new way to study the phenomenon of homogeneous nucleation in the freezing of liquids in the realm of very deep supercooling. Nevertheless, a completely satisfactory procedure to analyze such MD simulations has not yet appeared. The aim of the present paper is to outline an approach for such an analysis. In a previous paper[1] (paper I) a procedure was developed for deriving nucleation rates and nucleation time lags from simulations of freezing in sets of large supercooled clusters, provided the transient nucleation regime had already been characterized. How to deal with the parameter characterizing this regime was not addressed satisfactorily, however. To date, to our knowledge, no reliable information establishing this parameter exists. Paper I also analyzed the statistical errors to be expected, although the errors were somewhat distorted because they were not based upon the optimum weighting of data.

In the following we sketch aspects that still need to be considered in the analysis of kinetic data from MD simulations. Three parameters are involved: the rate, J_s, the time lag, t_o, indicating the duration of the transient regime, and what we have chosen to call the "reduced moment," M_R, to characterize nucleation kinetics in the transient regime.

As before, we adapt Wu's method of moments[2] for bulk systems to systems of clusters and characterize the transient kinetic regime in terms of the reduced moment, M_R, related to Wu's moment M as outlined below. Because MD sample sizes for clusters tend to be comparatively small from a statistical perspective, a scheme was developed in paper I for constructing realistic data sets to model stochastic nucleation times (SNTs). Statistics gathered for J_s and t_o values derived from least squares analyses of such sets of SNTs were used to assess the relative merits of two basic approaches to MD data analysis. The first (option 1 in paper I) was based on expressions, summarized below, which explicitly include the effects of transient nucleation. The other (option 2 in paper I) was the conventional model of sudden onset of nucleation that ignores the gradual build up of pre-critical embryos. It was found that the sudden onset model produced smaller apparent statistical errors, and that corrections could be made for the associated systematic errors. Recent advances in computer technology have greatly increased the practical limit of MD sample sizes for molecular systems as complex as 700-molecule clusters of MF_6 from about 20 to many hundreds of nucleation events. Uncertainties corresponding to these larger sample sizes are sufficiently reduced that the explicit treatment of transient nucleation (option 1) is now preferred.

The motivation for writing a sequel to paper I was 3-fold. First was the recognition that neither choice of weights suggested in prior work[3] and used in paper I, namely $w_j = 1$ or $w_j = \arctan(1/J_s V_c t_j)$, with V_c the cluster volume, was the optimum choice for treating transient nucleation. In the present paper an

* Corresponding author. E-mail: lbart@umich.edu.

10.1021/jp022237i CCC: $25.00
Published on Web 02/20/2003

improved weighting function is presented that is computationally simple and appropriate for the sets of stochastic nucleation times associated with transitions in clusters. A second motivation was the development of a greatly improved method for constructing model sets of SNTs for least squares analyses. Third, a method was developed to treat the extraction of information about the reduced moment. Paper I had dealt only briefly with this problem. A more detailed examination of ways to determine M_R is outlined in the following, including examples from molecular dynamics data sets. Moreover, the first analysis to date is presented for determining the uncertainty in M_R.

The virtue of realistic constructions of sets of nucleation times is that the rates and time lags fed into the constructions are known exactly so that actual errors in rates and time lags derived by least squares analyses of sets of nucleation times can be recognized. Therefore, realistic expectation values of standard deviations in derived parameters can be determined. Such measures of error should also apply to analyses of data from MD simulations where actual errors are unknown.

The described treatment of model data makes it possible to find how parameter uncertainties depend on the number of nucleation events per data set, both for standard deviations derived from least squares residuals, σ^{ls}, and for expectation values, σ^{ϵ}, of standard deviations derived from analyses of large ensembles of sets of SNTs. Guidelines are obtained for estimating realistic uncertainties applicable to single MD data sets where the error matrix alone is shown to be entirely insufficient. Illustrative results of this approach that treat preliminary MD data[4,5] for $(RbCl)_{108}$ and $(SeF_6)_{550}$ clusters are presented.

Summary of Method of Moments for Clusters. Prior to the treatment in ref 1, the decay of a population of N_o unfrozen clusters due to nucleation had been considered to follow the first-order law[3]

$$\ln[N_l(t)/N_o] = -K(t - t_o) \tag{1}$$

for the sudden onset of nucleation at time t_o, the so-called time lag. Here $t \geq t_o$ is the time of the lth nucleation event, $N_l(t)$ is the number of clusters not yet having experienced formation of a critical nucleus before the lth nucleation, and K represents the product J_sV_c of the steady-state nucleation rate, J_s, and the cluster volume, V_c. Paper I adapted Wu's method of moments, with its explicit treatment of transient nucleation in a large bulk volume to one suitable for use with finite sets of clusters, resulting in an expression of the form

$$\ln[N_l(t)/N_o] = -KS(t) \tag{2a}$$

The derivation and meaning of $S(t)$ are given below in eqs 3,4. $S(t)$ differs from $(t - t_o)$ in the region of transient nucleation, but approaches that expression in the limit of large t.

For convenience in later equations, we adopt the notation

$$g_l(t) \equiv -\ln[N_l(t)/N_o] \tag{2b}$$

Wu's method of moments[2] yields an explicit expression for the ratio $R(t)$ in the development of nucleation rate $J(t)$, where

$$J(t)/J_s \equiv R(t) \tag{3a}$$

with the ratio expressed in Wu's notation as

$$R(t) = 1 - \frac{1}{2}\mathrm{erfc}\left[\frac{\ln(t/t_o) - a}{\sqrt{2b^2}}\right] \tag{3b}$$

This ratio differs from unity during the time it takes for the buildup of precursors that ultimately leads to a steady-state rate J_s of production of critical nuclei. Integration of eq 3b yields a relation for the accumulated number of critical nuclei, $N(t)$, in the freezing of a fixed volume, V_l, of a supercooled liquid. Assuming that the nuclei formed do not significantly deplete the volume V_l accessible for further nucleation,

$$S(t) \equiv N(t)/J_sV_l = \int_o^t R(t')dt' \tag{4a}$$

Integration of $R(t)$ yields

$$S(t) = t\left(1 - \frac{1}{2}\mathrm{erfc}\left[\frac{\ln(t/t_o) - a}{\sqrt{2b^2}}\right]\right) - t_o\left(1 - \frac{1}{2}\mathrm{erfc}\left[\frac{\ln(t/t_o) + a}{\sqrt{2b^2}}\right]\right) \tag{4b}$$

Wu's parameters a and b are defined in terms of a quantity we choose to call the "reduced moment" M_R, such that

$$a = -\frac{1}{2}\ln(M_R) \tag{5}$$

and

$$b^2 = \ln(M_R) \tag{6}$$

Here $M_R = 2M/t_o^2$, where the moment M is a quantity Wu regarded as a free parameter to be derived in the analysis of experimental data. It is evident that the lowest value the reduced moment can have is unity.

If, instead of a system consisting of Wu's large fixed volume, the system is a set of N_o supercooled clusters, each of whose volumes is V_c, then the nucleation expression must be modified. When one cluster is removed from the set after a critical nucleus has formed in it, the volume remaining in the set becomes N_lV_c, where N_l is the number of liquid clusters left in the set. Therefore, instead of eqs 3a and 3b, we have

$$R(t) = \frac{J(t)}{J_s} = \frac{1}{J_s}\frac{dN/dt}{N_lV_c} = \frac{-(dN_l/N_l)/dt}{K} \tag{7}$$

By virtue of the definition $S(t) \equiv \int_o^t R(t')\,dt'$, rearrangement and integration of eq 7 produces eq 2 above. Note that eq 2 reduces to eq 1 as M_R approaches unity, or when t becomes very large. Note also that N_l has been treated as a continuous function of t, an approximation that becomes more accurate as the number of clusters in a set becomes large. How $S(t)$, and thereby $-\ln[N_l(t)/N_o]$, varies with M_R for a given t is shown in Figure 1.

In analyzing simulations for a set of N_o clusters, the sequence of numbers N_l is known exactly, whereas the stochastically determined times are very much a matter of chance and can vary widely from set to set. Therefore it is reasonable in least squares analyses to consider the nucleation time as the uncertain "y" variable and the quantity $g_l(t)$ of eq 2b to be the accurately known "x" variable. To carry out least squares analyses, then, it is necessary to invert eq 7 to the form $t(g_l)$. As shown in paper I, an empirical expression for the reduced time t/t_o as a function of g_l, namely

$$t/t_o \approx 1 + g_l/Kt_o - (1 - 0.5/M_R^{2.5}) \times \exp[-1.82(g_l/Kt_o)^{1/2}/(M_R - 1)^{0.41}] \tag{8}$$

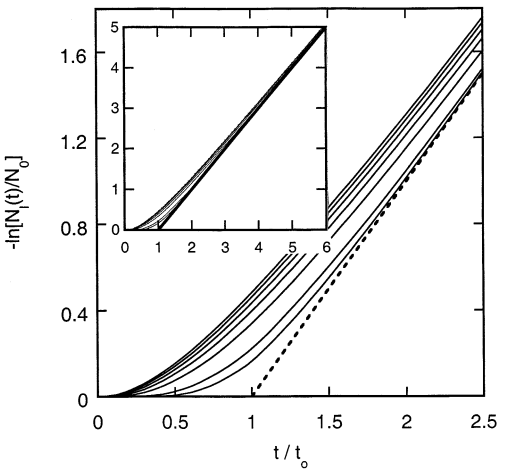

Figure 1. Dependence of the decaying population N_l of liquid clusters on the reduced moment, M_R. In all curves, K is taken to be 1. The dashed line in the main plot corresponds to the limiting moment, unity, where the decay of N_l is exponential. The next two values of the moment are in the range of those suggested in prior work, namely 1.2 and 1.4. Intervals between successive M_R values in the remainder of the curves, starting with 2.2, are all 0.8. In the insert are plotted the same curves out to much larger values of the time. This is to show that the curves ultimately do approach the limiting value, that for $M_R = 1$, here represented by a heavy solid line.

is sufficiently accurate for use in least squares calculations over the range of M_R expected to be physically significant.

Generation of Model Stochastic Times. A requisite for testing the method of moments is a procedure for selecting large numbers of sets of stochastic nucleation times (SNTs) of various numbers, N_o, of independent clusters per set. This provides as much "data" as desired, all of it based on known values for rate, time lag, and reduced moment. To emphasize what was mentioned above, fitting such data by least squares reveals the actual errors in the derived parameter values in individual sets, and provides the standard deviations, σ^{ls}, derived from the residuals encountered in individual data sets. More importantly, from the actual errors in parameters, it yields the expectation values of the standard deviations σ^{ϵ} derived from the variance in the least squares parameters over an ensemble of sets. Presumably this estimate of uncertainty can be applied to cluster nucleation data from MD simulations, where "true" parameter values are unknown. Initial results for SNT analyses presented in paper I confirmed the utility of the approach. However, they employed an inefficient and numerically problematic way to create the sets of SNTs, and were based on weights that were decidedly nonoptimum for the transient nucleation model. An improved procedure for determining sets of SNTs is outlined next.

The general scheme for generating suitable sets of stochastic times is to apportion time, from 0 to ∞, into an arbitrarily large number, N_B, of equally probable intervals or time bins. We have usually taken N_B to be 10 000. Here, probability refers to the likelihood of a nucleation event occurring within the time spanned by each bin. An event time, t_k, lying somewhere within bin k, is ascribed to each bin. Typically each t_k is positioned at the same fraction, f_B, of the bin's span from the start of the bin. Individual sets of stochastic times, N_o in number, are produced as follows. First N_o bin numbers, n, are chosen randomly from the entire set of N_B bins. To mimic sizes of sets characteristic of MD simulations, the number of time bins should greatly exceed N_o, although calculations to establish limiting values of K^{ls} and t_o^{ls} have been carried out for sets with N_o as large as 20 000. The N_o stochastic nucleation times are simply the values of the times, t_n, associated with the chosen bins. Set selection is repeated until a statistically meaningful ensemble of sets of simulated nucleation times $\{t_n\}$ is formed. Each set is ordered in time and analyzed in the same way as a set from an MD simulation. As will be seen, the new method for determining values of bin event times, t_k, allows one either to determine t's for the entire set of bins or, if desired, just for individual randomly selected bins.

In the earlier paper,[1] time apportionment was based on integration of the Wu probability distribution, $P(t)$, suitably modified for clusters. As derived in paper I,

$$P(t) = KR(t) \exp[-KS(t)] \tag{9}$$

The desired t_k values are the integration limits for which the expressions

$$\int_0^{t_k} P(t')\, dt' = (k - 1 + f_B)/N_B,\ k = 1, ..., N_B \tag{10}$$

are satisfied.

Solving eq 10 for the integration limits, t_k, by numerical integration of $P(t)$ suffers some practical deficiencies. The method is prone to systematic, often cumulative, errors, including round-off errors in accumulating the running total of the integration limit. A method that avoids the troublesome numerical integration process is as follows. It has proven to be orders of magnitude more accurate, as well as simpler to apply.

Consider a term, h_k, analogous to g_l in eq 2b,

$$h_k \equiv -\ln[N_k(t_k)/N_B] \tag{11a}$$

in which k indexes the nucleation events, and the total number of nucleation events is now N_B rather than N_o. In this expression $N_k(t_k)$ represents, for N_B evolving clusters, the average number of unnucleated clusters remaining at some arbitrary fraction, f_B, of the way across bin k, namely,

$$N_k(t) = [N_B - (k - 1 + f_B)] \tag{11b}$$

Taking $f_B = 0$ corresponds to locating the kth nucleation event at the beginning of bin k. For the particular case of N_B equally probable nucleation events distributed uniformly over the N_B time bins eq 2, evaluated at these "average" event times, becomes

$$KS(t_k) = h_k = -\ln([N_B - (k - 1 + f_B)]/N_B) \tag{12}$$

Finding the desired time apportionment reduces to finding values of t_k that satisfy eq 12. Because $S(t)$ is a smooth, monotonically increasing function with a limiting expression

$$\lim_{t \to \infty} S(t) = (t - t_o) \tag{13}$$

solutions of eq 12 are easily found. If eq 8 were exact, it would be possible to use it for calculating the desired t_k values directly, by substituting the expression for h_k in place of g_l, but a more precise evaluation of bin event times is desired.

Simple search techniques suffice for locating individual t_k values, whether for randomly selected k's or for the complete set in succession. Once a pair of t's is found such that their corresponding $S(t)$ values bracket the target h_k value, the time interval can be successively halved, each time choosing the half whose corresponding $S(t)$ values still bracket the target h_k value. In the absence of any prior knowledge of narrower limits for t_k, one can start with the full range, $0 < t_k < t_{N_B}$ or with $t_{k-1} < t_k \leq t_{N_B}$. For large N_B substituting the limiting expression for

$S(t)$ into eq 12 leads to a simple expression for t_{N_B}. When calculating a full set of t_k's, a number of ways of estimating the width of the next bin or locating the next t_k are available, including application of eq 8 or eq 17 below. A faster, more efficient search scheme than the indicated geometric progression combines finding a partially reduced time span, then fitting selected $[t,S(t)]$ points by a low-order polynomial and interpolating to the final t_k value, or even iterating the polynomial plus interpolation steps, with reduced time spans for each iteration. As a practical matter, all of the numerical integrations of the previous method (the "P-method"), and all of the calculations of the present method (the "S-method") were carried out in reduced time, rate, and volume (i.e., assigning unity to the values for t_o and K).

Error Matrix. The manner in which data are weighted and standard deviations are calculated can have a significant effect on derived parameters and their apparent uncertainties. All calculations of σ^{ls} reported here were based on the "bona fide error matrix" $\mathbf{M}_x^W$, for calculating parameter standard deviations and correlation coefficients from residuals where, in the case of the derivation of two parameters,

$$\mathbf{M}_x^{\mathbf{W}} \equiv \begin{pmatrix} {\sigma_K}^2 & \sigma_K \sigma_{t_o} \rho_{K,t_o} \\ \sigma_K \sigma_{t_o} \rho_{K,t_o} & {\sigma_{t_o}}^2 \end{pmatrix}$$

$$= \mathbf{B}^{-1}\mathbf{A}'\mathbf{W}\mathbf{M}_f\mathbf{W}\mathbf{A}\mathbf{B}^{-1} \tag{14}$$

with **B**, **A**, **W**, and $\mathbf{M}_f$ representing, respectively, the information matrix, the design matrix, the weight matrix, and the matrix of errors in observations.[6] The matrix $\mathbf{M}_x^W$ is valid even for nonoptimum weights as opposed to the false, or "zero-order" error matrix $\mathbf{M}_x^o$,

$$\mathbf{M}_x^{\,o} = \mathbf{B}^{-1}\mathbf{V}'\mathbf{W}\mathbf{V}/(n-m) \tag{15}$$

where **V** represents the matrix of residuals, and n and m, the number of observations and the number of derived parameters. Equation 15 is widely used but valid only if the weights are optimum,[6] that is (in the case of uncorrelated errors in observations), weights must be proportional to the inverse of the variances in observations in $\mathbf{M}_f$. Nucleation times in separate clusters of molecular dynamics simulations are uncorrelated, assuming the starting configurations are prepared properly, so that a diagonal weight matrix is appropriate. No uncertainty is attached to the values for $g_l(t) = -\ln(N_l(t)/N_o)$, where N_o is the number of clusters in the set, and N_l the number of unnucleated clusters just prior to the lth nucleation. However, the nucleation times, t_l, occurring purely by chance, are subject to substantial uncertainties. An estimate of the variance in individual stochastic nucleation times is presented next.

Weight Functions. A procedure to devise a weight function taking the uncertainty in nucleation times into account is as follows. The optimum weight w_l is inversely proportional to the variance expected for the time, t_l. Variances in stochastic times are related to the average intervals between such times. This correspondence is suggested by the following argument. In least squares analyses of sets of N_o events, the calculated times t_l to be compared with the observed times (in a molecular dynamics run or in a model set of stochastic times), can be envisaged as belonging to time bins in a bin array with N_B being just N_o. Since, on average, whether in MD or model runs, the bins tend to be sampled evenly over the bin array, the typical residual for time t_l tends to have the magnitude of the breadth of bin l. This breadth is the difference between times t_{l+1} and t_l, each time calculated from eq 8. Therefore, a plausible weight for event l would be proportional to the inverse square of this interval, or

$$w_l = C/(t_{l+1} - t_l)^2 \tag{16}$$

where C is an arbitrary constant, and the times are those from eq 8. Although these intervals depend on the quantities to be derived, namely the nucleation rate and the time lag, as well as the moment, it is simple to cycle to self-consistency during the least squares routine. Inasmuch as the time width of each bin for the stochastic generation of nucleation times is defined by the integral

$$\int_{t_l}^{t_{l+1}} P(t')\,dt' = 1/N_B$$

$$\approx \bar{P}\Delta t \tag{17}$$

the time spread Δt over a bin is roughly proportional to the inverse of the nucleation probability function, $P(t)$, a function given explicitly in eq 9. Therefore, the weights are small when P is small, that is, when the nucleation time is either much smaller than or very much larger than the nucleation time lag.

Determination of Reduced Moment. One possible method is suggested by the fact that the reduced moment is related to the number of events occurring before the time lag t_o is reached. For example, if M_R were unity, there would be no such events whatsoever. As M_R increases, so does the fraction, F_r, of nucleation events occurring before t_o. Statistically, this fraction is

$$F_r = 1 - \exp\left[-Kt_o \mathrm{erf}\left(\frac{\sqrt{\ln(M_R)/2}}{2}\right)\right] \tag{18}$$

This method would work very well if the time lag were known independently of the least-squares analysis of the nucleation data. Unfortunately, the time lag can be determined from MD data only by fitting the observed set of nucleation times with the representation of eq 8 or equivalent. It turns out that in such fittings by least squares, K and t_o are highly correlated, with a correlation coefficient of the order of 0.9 or higher. As the parameter M_R is increased in least squares analyses, the value of t_o derived also increases as explained in paper I.[1] Therefore, the product Kt_o increases even faster with the result that the theoretical expression (eq 18) for F_r closely parallels the observed number of events occurring before the derived time lag, almost irrespective of the value assumed for M_R. This remarkable parallel is illustrated in Table 1 for the case of 150 nucleation events in molten $(RbCl)_{108}$ clusters.[4] Such a parallel was also seen in a series of runs on liquid $(SeF_6)_{550}$ clusters. Therefore, this approach to determining the reduced moment is ineffective.

Despite the above result, there nevertheless is information about M_R in the data. If the least squares analyses are carried out for a series of values of M_R, one of the representations of the data is better than the others. The region of transient nucleation is more sensitive to the value of M_R than the region at higher times. After determining the value of t_o corresponding to each assumed moment from the full set of data, it is worthwhile to calculate the sums of squares of the residuals in the region of transient nucleation. This procedure yields a more definitive result than the first approach described above, as suggested in Table 2 and shown more objectively in a subsequent section. Still, the determination of reduced moment is very noisy unless N_o is in the neighborhood of thousands of

TABLE 1: Fraction, F_r, of Nucleation Events Occurring before Time t_o Is Reached, as a Function of Assumed Reduced Moment M_R[a]

M_R	t_o (ps)	Kt_o	F_r [eq 18]	F_r (MD result)
1.08	64.9	0.957	0.10	0.11
1.2	74.5	1.235	0.18	0.21
1.4	93.2	1.845	0.34	0.34
1.6	118.2	2.801	0.52	0.54
1.9	175.2	5.694	0.83	0.83

[a] Comparison of theoretical expectation values of eq 18 with those derived from optimally weighted least squares analyses of a set of 150 MD nucleation runs for $(RbCl)_{108}$ clusters.[4] In each case, weights and derived results for t_o and K are consistent with the assumed reduced moment. Had the time lag been known independently of the assumed reduced moment, the fraction of events before t_o could have served to determine the reduced moment for the MD run. Since the MD least squares fraction F_r rises in parallel with that based on eq 18, it can be seen that F_r provides no basis for estimating the moment.

TABLE 2: Estimation of Transient Nucleation Parameter M_R from Results of Optimally Weighted Least Squares Analyses of a Set of 150 MD Nucleation Runs for $(RbCl)_{108}$ Clusters[a,b]

M_R	ρ_{tK}	$\sigma(t,$ ps)	$\sigma^2_{0.3}(t)$
1.08	0.837	3.27	7.10
1.2	0.882	2.44	3.33
1.4	0.950	2.13	2.65
1.6	0.979	2.36	5.24
1.9	0.992	2.79	8.64

[a] Ref 4. [b] Listed for each assumed moment are the correlation coefficients ρ_{tK} associated with the derived time lag and nucleation rate, the weighted standard deviation $\sigma(t)$ between observed and calculated nucleation times, and the weighted variance $\sigma^2_{0.3}(t)$ in the transient range covering the first 30% of the nucleation events. The latter quantity is generally more discriminating than the overall standard deviation.

independent events. If least squares analyses are carried out *only* for times in the transient region, t_o and K adjust themselves to fit the data almost equally well for any plausible value of the reduced moment. Least squares refinements including times beyond the transient region are needed to place constraints on t_0 and K.

Results

Construction of Model Stochastic Times. Bin times calculated via the S-method exhibit a noise level consistent with the degree of precision selected for the computation, and are free of systematic errors, provided that the routine adopted for the erfc function is accurate. The routine selected for the present computations was taken from ref 7. On the other hand, errors we encountered when the integration method was used were many orders of magnitude greater than for the S-method.

Weights. Listed in Table 3 is an illustrative example analyzing the times of nucleation events in the freezing of 100 SeF_6 clusters each composed of 550 molecules.[5] Standard deviations in the derived parameters K and t_o, as well as the standard deviations for the residuals in nucleation times, are tabulated. Standard deviations were calculated from residuals via the "bona fide error matrix" $\mathbf{M}_x^W$, eq 14. Standard deviations in Table 3 can be seen to be appreciably lower when the suggested weights are adopted than when the weights applied were either unit weights or the arctangent weights that had been reported to be satisfactory for the analyses of stochastic data ignoring transient nucleation.[3] The very large effect of choice of weights is due mainly to the overemphasis of data at large times unless optimum weights are adopted. Large nucleation

TABLE 3: Illustration of Effects of Different Weight Functions w on Derived Parameters and Their Standard Deviations[a]

w	$10^3 K^b$	$10^3 \sigma_K^{ls}$	$10^3 \sigma_K^{\epsilon}$	t_o[c]	$\sigma_{t_o}^{ls}$	$\sigma_{t_o}^{\epsilon}$	ρ_{K,t_o}	$\sigma(t)^{c,d}$
opt[e]	5.01	0.046	0.81	116	1.65	15	0.867	6.1
atan[f]	5.68	0.212	0.85	136	5.94	21	0.972	15.3
unity	5.94	0.287	1.02	149	9.24	31	0.979	20.3

[a] A set of 100 nucleation events in $(SeF_6)_{550}$ clusters[5] was analyzed by least squares using the optimum weight, eq 16, the arctangent weight proposed in ref 3, or unit weight, suggested in ref 1. Comparisons are based on the reduced moment 1.4 but similar results are found for the other reduced moments considered in Tables 1 and 2. Standard deviations σ^{ls} are derived from the least squares error matrix. The much larger standard deviations σ^{ϵ} are estimated via eqs 19 and 20 from the expectation values of the variance found for the kinetic parameters in analyses of large sets of stochastically generated nucleation times. [b] K in ps^{-1}. [c] Times in ps. [d] Standard deviation in time. [e] Equation 16. [f] See ref 1.

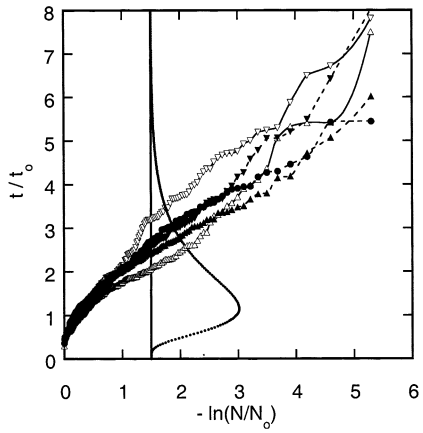

Figure 2. Stochastic sets of nucleation times, each for 200 events. Plots do not illustrate the representative scatter but illustrate sets corresponding to the maximum (1.63, △), and minimum (0.66, ▽) K values (open markers on solid curves), and maximum (1.50, ▲), minimum (0.64, ▼), and midrange t_o (●) values (solid markers on dashed curves), encountered in 10 000 random model sets, all generated with the same input rate and time lag. Also plotted on a vertical axis is the probability distribution, $P(t)$ of eq 9. The density of event times tends to be proportional to $P(t)$, and the optimum weights in least squares analyses are proportional to the square of $P(t)$.

times are associated with very large uncertainties, a consequence of the low probability of nucleation (eq 9) in any particular interval of time when t is large.

Uncertainties. Sets of N_o nucleation times generated from N_o random hits on time bins closely resemble the sets of N_o times acquired in MD simulations. Some idea about the distribution of times is conveyed by the 5 sets, each of 200 times, illustrated in Figure 2. The rather large difference between the sets plotted is not indicative of the characteristic stochastic scatter but rather, illustrates the extremes found in 10,000 sets. Shown are the sets that yielded the maximum and minimum values of K in least squares analyses, and the maximum, minimum, and a mid range value of t_o. Most sets in the ensemble resemble that for the mid range t_o more closely than they do any of the illustrated extremes.

Also displayed on a vertical axis in the figure is the probability distribution (eq 9) in time, t, according to the Wu moment theory.[2] The density of points along the time axis is based on this distribution. Moreover, since the optimum weights of points are proportional to the square of this distribution, it can be seen how little the points at large times influence the least squares analyses.

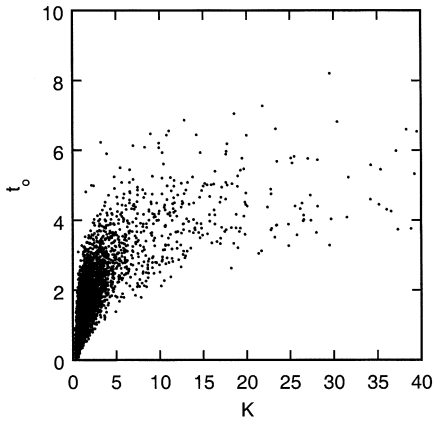

Figure 3. Distribution of least squares values of K and associated t_o values acquired in analyses of 1000 sets, each of $N_o = 10$ stochastic nucleation events. For each set the event times were based on input K, and t_o values of unity. Two percent of the points lie at K values outside the figure. Clearly, sets of only 10 events are entirely insufficient to establish rates and time lags.

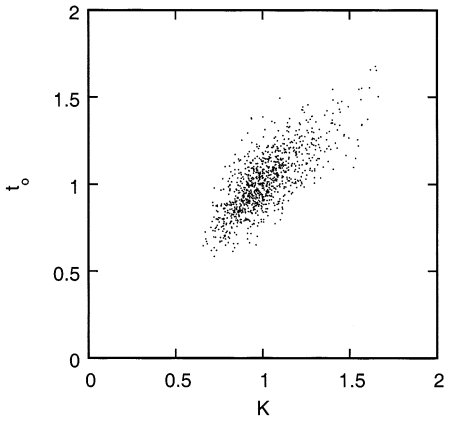

Figure 4. Distribution of least squares values of K and associated t_o values acquired in analyses of 1000 sets, each of $N_o = 100$ stochastic nucleation events. For each set the event times were based on input K, and t_o values of unity.

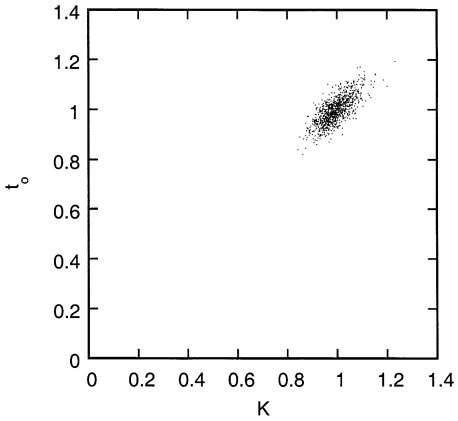

Figure 5. Distribution of least squares values of K and associated t_o values acquired in analyses of 1000 sets, each of $N_o = 1000$ stochastic nucleation events. For each set the event times were based on input K, and t_o values of unity.

A more direct portrayal of the distribution in results for the derived kinetic parameters, as well as the dependence on set size is shown in Figures3−5. Plotted are the (K, t_o) pairs found in 1000 sets of runs for set sizes of $N_o = 10$, 100, and 1,000 clusters. Most obvious is the decrease in scatter with increasing set size. Next, is the strong correlation between K and t_o values,

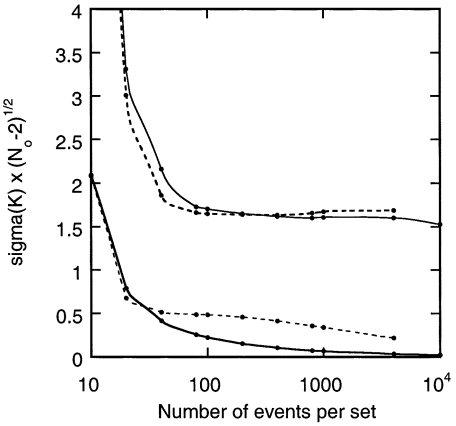

Figure 6. Dependence on sizes of data sets, of uncertainties in nucleation rates associated with analyses of sets of N_o stochastic nucleation events. Solid lines correspond to optimum weighting, and dashed, to unit weights. The upper curves are for the true mean standard deviations, σ_K^{ϵ}, and the lower, the mean standard deviations determined from the least squares error matrices. Each generation of stochastic nucleation times was based on a K value of unity. All values plotted are σ values multiplied by the factor $(N_o - 2)^{1/2}$ to find whether the "root N law" applies.

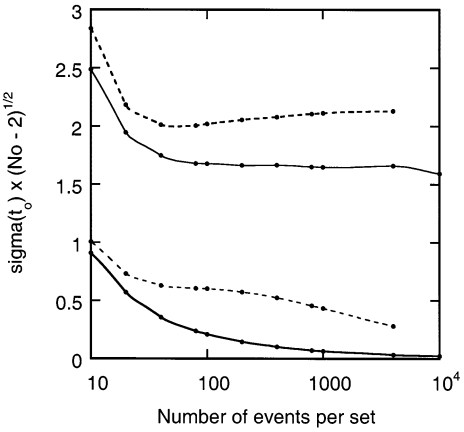

Figure 7. Dependence on sizes of data sets, of uncertainties in time lags associated with analyses of sets of N_o stochastic nucleation events. Solid lines correspond to optimum weighting, and dashed, to unit weights. The upper curves are for the true mean standard deviations, $\sigma_{t_o}^{\epsilon}$, and the lower, the mean standard deviations determined from the least squares error matrices. Each generation of stochastic nucleation times was based on a t_o value of unity. All values plotted are σ values multiplied by the factor $(N_o - 2)^{1/2}$ to find whether the "root N law" applies.

although the correlation breaks down in the ensemble of sets of only 10 nucleation events where the exceedingly long tail of the distribution in K from least squares analyses is only hinted at in the figure. Values of K as large as 150 times that fed into the generation of stochastic times have been seen. It is clear that sets of as few as 10 events cannot be expected to yield reliable kinetic parameters.

Figures 3−5 qualitatively illustrate the magnitude of the dispersion of least squares parameters associated with the stochastic nature of nucleation. A more quantitative portrayal of uncertainties is given in Figures 6 for K, and Figure 7 for t_o. Uncertainties plotted are multiplied by the factor $(N_o - 2)^{1/2}$ to find whether they follow what statisticians refer to as the "root N law" observed in a wide variety of problems of physical interest.[8] It is apparent that the expectation values σ^{ϵ} do tend to follow this law within statistical error except for very small N_o, but the σ^{ls} values do not. The σ^{ls} values obtained from the least-

squares error matrix depend, as expected, upon the weighting of data, with optimum weights yielding appreciably smaller standard deviations (over the practical range of N_o) than do the unit weights used in paper I. More important is the fact that the expectation values σ^{ϵ} are considerably larger than the σ^{ls}. These expectation values, whose magnitudes are manifested in the dispersion of points in Figures 3−5, are the more valid measure of the real uncertainty. Moreover, these magnitudes are not strongly dependent on whether the optimum or unit weights are adopted. Although this might suggest the weighting scheme is of little practical importance, the kinetic parameters derived do depend appreciably on the weights (Table 3) and the optimum weights are preferred.

Another result confirmed by application of the stochastic model is that both the σ^{ϵ} and the σ^{ls} values scale with the kinetic parameters. That is, if K (or t_o) is doubled, so also are the corresponding values of σ^{ϵ} and σ^{ls}. Therefore, in the inference of uncertainties in MD results, where a measure of σ^{ϵ} is not available, a reasonable estimate of this uncertainty can be obtained from either of two relations, or

$$\sigma^{\epsilon}_{K.MD} = \left(\frac{K^{ls}_{MD}}{K^{ls}_{stoch}}\right)\sigma^{\epsilon}_{K,stoch} \tag{19}$$

$$\sigma^{\epsilon}_{K.MD} = \left(\frac{\sigma^{ls}_{K,MD}}{\sigma^{ls}_{K,\,stoch}}\right)\sigma^{\epsilon}_{K,stoch} \tag{20}$$

and similarly for t_o, where the subscript "stoch" refers to a value derived from a sets of model stochastic nucleation times. Since the dispersion from the stochastic runs is known both for the parameters K and t_o, and also for the σ^{ls} values for each parameter, a weighted mean of results from eqs 19 and 20 can be applied to the estimate of the MD expectation values, σ^{ϵ}. This averaging was done to obtain the entries in Table 3.

Reduced Moments. In paper I it was stated that least squares refinements in which all three kinetic parameters, K, t_o, and M_R, were refined together were too ill-conditioned to be worthwhile. Now that greater numbers on nucleation events can be acquired, that conclusion may no longer be entirely true, although such analyses are likely to be unstable even for currently accessible sizes for MD data sets. In the present paper such refinements including M_R were not attempted but, rather, information about the reduced moment was extracted as outlined in the foregoing. Results illustrating the sensitivity to M_R of the overall standard deviation of nucleation times and of the sums of squares of residuals in times in the transient region are illustrated in Table 2 and, more definitively, in Figures 8 and 9.

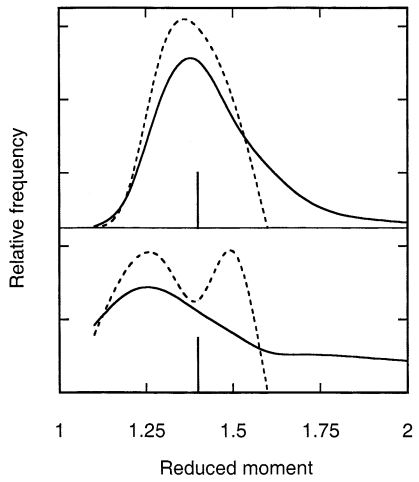

Figure 8. Histograms of the values of M_R derived from least-squares fits of stochastically generated nucleation times for thousands of sets of times when the value of M_R input into the stochastic generation was 1.4, the value indicated by the heavy vertical lines. In the lower two curves the sets of nucleation times included only 200 nucleation events. The upper two curves correspond to sets of 800 nucleation events. A truly precise determination would require sets of thousands of events. The criteria for selecting the M_R value from the least-squares analysis of any given set were (solid curves) the minimum standard deviation in time over the entire set, and (dashed curves) the minimum variance in time in the transient regime, namely the first 20% of the events.

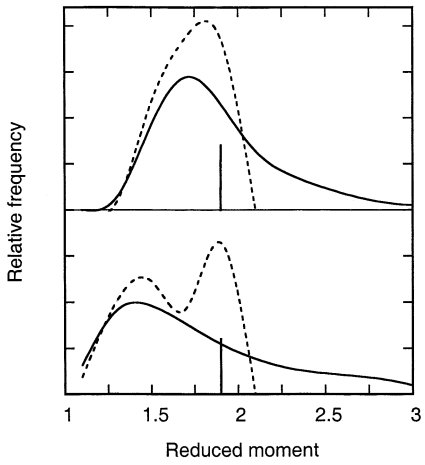

Figure 9. Same quantities as in Figure 8 except that the value of M_R input into the stochastic generation of nucleation times was 1.9.

Discussion

Reduced Moments. Very little information about experimental values of the reduced moment is available in the literature. Wu gave no guidelines for estimating the moment but suggested that it be determined from experiments. Kashchiev's theoretical result,[9] which corresponds to a reduced moment of very nearly 1.4, has been cited as being in good agreement with experimental data,[10] but a later publication[2] claimed that the Kashchiev result is based on rather crude approximations and, therefore cannot be assumed to be correct. From the results in Table 2, it appears that the best fit for a system of 150 independent $(RbCl)_{108}$ clusters[4] is consistent with the Kashchiev prediction, but 150 nucleation events are far too few to be definitive. Preliminary results for 800 nucleation events in much larger clusters of SeF_6 have suggested a similar value for the moment, while a numerical simulation of the crystallization of lithium disilicate glass[11] can be represented quite well by a value of M_R in the range of 1.1 to 1.2.

Uncertainties. One of the most important results of the present work is the finding that the inference of uncertainty σ^{ls} from least squares residuals for a given set of data, as expressed in the error matrix, far underestimates the true statistical uncertainty σ^{ϵ}, as illustrated in Figures 3−7. It is often assumed that errors yielded by the error matrix are "random errors" while those over and above such random errors are "systematic errors" either stemming from some deviation of conditions from those assumed for the data or some error in the theoretical expression used to define the residuals. In the present treatment, however, the data points are *all* based entirely on random selections of time bins, and hence errors above those from the error matrix do not conform to the concept of "systematic errors". The expectation values of uncertainties, σ^{ϵ}, do follow the "root N law" expected for random data, while the least squares values utterly fail to, principally because the variance of the residuals

falls off markedly as the number of events, N_o, in a set increases. This behavior arises because a small N_o corresponds to a such a sparse selection of time bins that a potentially very unfaithful representation of eq 2 is generated. For a very large set where the sampling of time bins becomes more uniform, the distribution of times approaches the Wu transient nucleation distribution built into the construction of the time bins. While this rationale partly accounts for the diminishing mean-square residuals found as N_o increases, it does not fully explain why the error matrix is such an unsatisfactory gauge of the true uncertainty. The most important conclusion concerning the determination of rates and time lags, then, is that the variance $(\sigma^{\epsilon})^2$ is the appropriate indicator of uncertainty, not the elements of the least squares error matrix. Inference of this important information, then, requires the present stochastic analysis, for it is beyond the capacity of experiments to determine.

The same conclusion applies to the determination of the reduced moment. This moment is too strongly correlated with the other kinetic parameters to make its simultaneous derivation with K and t_o advisable when applying current techniques. On the other hand, if very large sets of nucleation times are available, this moment *can* be determined.

Figures 8 and 9 compare results of determining M_R by the two criteria mentioned in a previous section. These figures were generated by carrying out many thousands of sets of stochastic runs. If M_R is inferred from the best fit of an entire "experimental" curve of $-\ln[N_n(t)/N_o]$ vs t, with the theoretical curve of eq 8, it can be seen that the uncertainty can be very great because of the very broad foot of the histogram illustrated. The broadness of this foot is related to the dependence of $-\ln[N_n(t)/N_o]$ on M_R shown in Figure 1. As M_R increases by fixed increments, the curves crowd closer and closer together. In contrast, when M_R is determined from the minimum misfit in the transient regime, the broad foot of the histogram disappears. Therefore, as stated previously, the latter method is preferred. Why it leads to a bimodal distribution of moment when the number of events is only 200 is not clear. What is clear is that a precise determination of the moment characterizing the transient regime requires the acquisition of a very large number of nucleation events.

Acknowledgment. We thank Drs. Jinfan Huang and Yaroslav Chushak and Mr. Guarav Shah for permission to cite some of the preliminary results from their MD runs on RbCl and SeF_6 clusters.

References and Notes

(1) Bartell, L. S. *J. Phys. Chem. A* **2002**, *106*, 10893.

(2) Wu, D. In *Solid State Physics*; Ehrenreich, H., Spaepen, F., Eds.; Academic Press: New York, 1997; Vol. 50, p 38.

(3) Chushak, Y. G.; Santikary, P.; Bartell, L. S. *J. Phys. Chem. A* **1999**, *103*, 5636.

(4) Huang, J.; Bartell, L. S. Unpublished data, 2002.

(5) Chushak, Y. G.; Shah, G.; Bartell, L. S. Unpublished data, 2002.

(6) Bartell, L. S.; Anashkin, M. G. *J. Mol. Struct.* **1973**, *17*, 193.

(7) *Numerical Recipes. The Art of Scientific Computing*; Press, William H., Flannery, B. P., Teukolsky, S. A., Vetterling, W. T., Eds.; Cambridge University Press: Cambridge, 1986; p 164.

(8) Roe, B. P. *Probability and Statistics in Experimental Physics*; Springer-Verlag: New York, 1992.

(9) Kashchiev, D. *Surf. Sci.* **1969**, *14*, 209.

(10) Kelton, K. F.; Greer. A. L.; Thompson, C. V. *J. Chem. Phys.* **1983**, *79*, 6261.

(11) Greer, A. L.; Kelton, K. F. *J. Am. Ceram. Soc.* **1991**, *74*, 1015.