Statistical error in simulations of Poisson processes: Example of diffusion in solids

Johan O. Nilsson,^{1,*} Mikael Leetmaa,² Olga Yu. Vekilova,¹ Sergei I. Simak,³ and Natalia V. Skorodumova^{1,2}

¹*Department of Materials Science and Engineering, KTH - Royal Institute of Technology, Brinellvägen 23, 100 44 Stockholm, Sweden*

²*Department of Physics and Astronomy, Uppsala University, Box 516, 751 20 Uppsala, Sweden*

³*Department of Physics, Chemistry and Biology (IFM), Linköping University, 581 83, Linköping, Sweden*

(Received 15 April 2016; revised manuscript received 1 July 2016; published 31 August 2016)

Simulations of diffusion in solids often produce poor statistics of diffusion events. We present an analytical expression for the statistical error in ion conductivity obtained in such simulations. The error expression is not restricted to any computational method in particular, but valid in the context of simulation of Poisson processes in general. This analytical error expression is verified numerically for the case of Gd-doped ceria by running a large number of kinetic Monte Carlo calculations.

DOI: [10.1103/PhysRevB.94.085206](https://doi.org/10.1103/PhysRevB.94.085206)

I. INTRODUCTION

Diffusion processes are very common in nature and central in a multitude of research fields; from biology and the pharmaceutical sciences, to defect segregation in steels and alloys, to clean energy technologies and catalysis. Detailed insight into diffusion properties is therefore crucial for many technological applications as well as for a fundamental understanding of materials.

Atomic scale computer simulation methods have successfully been used in the study of ion transport properties in solids. The molecular dynamics (MD) technique in conjunction with interatomic forces described by simple empirical pair potentials (classical MD) is a popular simulation method for studying transport properties [1]. Many chemically complex materials, though, are poorly described by classical potentials, which warrant the use of a quantum mechanical description of the interatomic interactions (*ab initio* MD) [2]. Despite its merits, the use of *ab initio* MD is severely limited by a high computational cost that effectively prevents access to time scales necessary for simulating diffusion. This problem can be addressed by a variety of methods [3–5]. Specifically, the problem can be eased by accelerating the diffusion process by the application of an external field that takes the system out of equilibrium (nonequilibrium MD) [6,7]. The kinetic Monte Carlo (KMC) technique is another popular simulation method that allows for the simulation of very large length and time scales [8–10]. Although not a quantum mechanical method *per se*, it can be used with input rates calculated employing quantum mechanical techniques.

The issue of convergence and statistical error estimation is important to many simulation methods. Typically, an observable is computed as an average over recorded data, and therefore questions of the type, *was the simulation long enough?* or, *was sufficient data gathered?* naturally arise. The error in calculating observables, such as diffusion constant or conductivity, as time averages over a finite simulation time have been investigated by Zwanzig and Ailawadi in the context of a general time correlation function of some dynamical variable for a purely hypothetical system [11] and by Pranami and Lamm in the context of MD simulations of

a simple Lennard-Jones fluid. In the latter case the diffusion constant was calculated to assess whether the simulation was long enough for the ergodic hypothesis to be valid [12]. Müller-Krumbhaar and Binder have estimated the error in calculating ensemble averages as averages over a finite number of ensemble members in a Monte Carlo study of magnetization in an Ising spin lattice, and they have presented the error in magnetization as a function of system size [13].

Here we focus on ion diffusion in oxides and address the question: What is the error in calculated ion flux (conductivity) resulting from limited statistics of diffusive events? We study the statistical error using Gd-doped ceria as an example system directly relevant for applications [14], and we present a simple analytical expression to estimate this error. For the case of Gd-doped ceria we compare the error estimates obtained using our expression with those obtained numerically using two different computational methods: KMC and nonequilibrium *ab initio* MD (referred to as NEMD here). The analysis presented here is neither restricted to any specific method like MD or KMC, nor to the diffusion process in crystals itself, but rather it is valid in the general context of rare, Poissonian observations.

The paper is organized as follows. Section II contains details of the simulated system, the computational technicalities of the DFT calculations, description of the KMC method and the NEMD method, and a brief summary of statistical uncertainty and the Poisson process. Results are presented in Sec. III in which the connection between the KMC and NEMD methods is shown and an analytical expression for the error resulting from limited simulation length is presented. Error estimates calculated with the analytical expression are compared to the error estimates calculated numerically by running a large number of KMC simulations. Our findings are summarized in Sec. IV.

II. METHODOLOGY

Diffusion in our study system (Gd-doped ceria) was simulated using two different approaches. These were based on either a KMC method or a NEMD method. Although the two approaches differ conceptually in how diffusion is simulated they both rely on the calculation of the electronic structure of the system, which in both cases were performed with density functional theory (DFT) [15,16].

*Corresponding author: johann2@kth.se

FIG. 1. The fluorite structure. The $2 \times 2 \times 2$ (96 sites) cerium oxide supercell structure is shown in thin lines and the cerium oxide unit cell in thick lines. Large green circles are Ce atoms and red small circles are O atoms. From the supercell one O atom is removed to create a vacancy, and two Ce atoms are replaced by Gd atoms at positions $(0, 0, 0)$ and $(a/2, a/2, 0)$, marked by small arrows in the figure. a is the lattice constant.

A. The structure

The Gd-doped ceria (CeO_2) supercell consisted of $2 \times 2 \times 2$ fluorite unit cells with one oxygen atom removed and with two Ce atoms replaced by Gd atoms positioned at the nearest-neighbor positions $(0, 0, 0)$ and $(a/2, a/2, 0)$, where a is the lattice constant. The total number of atoms in the cell was 95, comprising 30 Ce atoms, 2 Gd atoms, and 63 O atoms with the chemical formula $\text{Gd}_{0.0625}\text{Ce}_{0.9375}\text{O}_{2-\delta}$, where $\delta = 0.03125$ (see Fig. 1). The dopant concentration was thus 6.25 at. %. The motivation for doping ceria with trivalent rare earth elements such as Gd is that this introduces oxygen vacancies in the system that facilitates the diffusion of the oxygen ions. The calculated equilibrium lattice constant of $a = 5.47 \text{ \AA}$ was used as in Ref. [7].

B. DFT parameters

For the DFT calculations we utilized the Vienna *ab initio* simulation package (VASP) [17–19] with the projector augmented wave (PAW) method [20]. Electronic exchange and correlation effects were modeled within the generalized gradient approximation (GGA) as parameterized by Perdew, Burke, and Ernzerhof (PBE) [21]. For the KMC input calculations of diffusion barriers at 0 K, a plane-wave cutoff of 500 eV and $2 \times 2 \times 2$ gamma centered k -point mesh were used, and for the NEMD calculations a plane-wave cutoff of 400 eV and $1 \times 1 \times 1$ k -point mesh (i.e., Γ -point) were used. The diffusion barriers calculated at 0 K using these two sets of parameters differ only by ~ 0.01 eV.

C. NEMD

An *ab initio* NEMD method based on the color diffusion algorithm [6,22] implemented by us in VASP for a previous

study [7] has been used to simulate diffusion in Gd-doped ceria. One can find a detailed description of the method and its important parameters in Ref. [7]; here we just outline the main features relevant for this work.

In this method an external field accelerates the otherwise inhibitive slow diffusion processes, without altering the diffusion mechanism itself. This is achieved by assigning fictitious so-called color charges c_i to the diffusing atoms that are acted on by a fictitious external color field $\mathbf{F}_{\text{ext}}$. The field accelerates the color charged atoms and speeds up their diffusion, while leaving the other atoms mostly unperturbed. The temperature of the system was kept constant by a velocity-rescaling thermostat [7].

The response of the system to the field is a color flux in the direction of the field defined as

$$J_c(t) = \frac{1}{V} \sum_{i=1}^I c_i v_i(t), \quad (1)$$

where I is the number of atoms in the cell, $v_i(t)$ is the projection of the velocity of atom i onto the direction of the field, and V is the volume of the cell. After an initial transient period the system reaches a steady state with constant temperature and constant color flux. Linear response theory in the steady state provides the connection between the color flux and the ion conductivity Q

$$Q = (Ze)^2 \lim_{t \rightarrow \infty} \lim_{F_{\text{ext}} \rightarrow 0} \frac{\langle J_c \rangle}{F_{\text{ext}}}, \quad (2)$$

where Ze is the charge of the diffusing atoms in elementary charges and the angular brackets $\langle \dots \rangle$ denote averaging over time. The onset of the linear regime in Gd-doped ceria was found at $F_{\text{ext}} = 0.5 \text{ eV/\AA}$ for a field in the $[111]$ direction. Below this field strength the response is predominantly linear, and for fields larger than $0.6 \text{ eV/\AA}$ the response of the system is strongly nonlinear [7]. For the analysis here we have chosen results obtained at $T = 1063 \text{ K}$.

The times between diffusive jumps were extracted from the atomic trajectories of the NEMD simulations by constructing cubical boxes centered on the ideal oxygen positions that completely fill the space. The time at which an oxygen atom crossed from one box to another was registered as t_j , and we defined the *waiting time*, i.e., the time between jumps, as the time between consecutive box crossings $t_j - t_{j-1}$ (some authors use the term ‘time of first escape’). We acknowledge that the extracted times depend on the shape and size of the boxes constructed around the oxygen sites. Here we choose the simplest box construction, which should be sufficient for our analysis.

D. KMC

The KMC method simulates the dynamic evolution of a system by stochastically exploring a preconstructed state space, with a probability of transition between the states dictated by predefined input rates. In KMC, the vibrational motion of the atoms around their equilibrium positions is discarded, and only the diffusive events themselves are simulated. Every simulation step in KMC therefore corresponds to a diffusive jump. The input rates k_{ij} for all the possible transitions $i \rightarrow j$ of the system (Sec. II A) were constructed from static diffusion

barriers E_{ij} as $k_{ij} = \nu \cdot \exp(-\frac{E_{ij}}{k_B T})$, where k_B is Boltzmann's constant, T is temperature, and ν is the prefactor. The KMC simulations were performed for the supercell used in the NEMD calculations (Sec. II A). There are 384 barriers for oxygen diffusion in our supercell in total, out of which 60 are symmetrically unique. The full set of barriers was calculated at 0 K using VASP (Sec. II B) by freezing the coordinate of the diffusing atom in five positions along the diffusion pathway made up by a straight line between the ideal oxygen lattice positions. The diffusing atom was allowed to relax in a plane perpendicular to the diffusion pathway as the system was structurally optimized. Finally, a polynomial was fitted to the energies of the optimized structures to calculate the diffusion barrier height as the difference between the equilibrium position energy and the top of the energy polynomial. The temperature for the simulation was chosen as in the NEMD calculations, $T = 1063$ K. Typical prefactors for diffusion are in the range [10] 10^{12} – 10^{13} Hz. To see the influence of the prefactor on the mean time step (see Sec. II E), we have used several values in this work: 3×10^{12} , 4×10^{12} , 5×10^{12} , 6×10^{12} Hz. The KMC simulations were performed with the *KMCLib* framework for lattice KMC simulations [23–25].

E. Sample mean and variance

Below we give a brief summary of statistical uncertainty, the textbook knowledge, which will facilitate the presentation of our results. Here, the quantity of interest is the waiting time, or the time between diffusive jumps, τ .

Often it is not possible to assess the whole population of a random variable of interest (here, all possible waiting times) and calculate the population mean (μ) and population variance (σ^2) in diffusion simulations. Usually we obtain only a limited sample of waiting times and use the sample mean ($\bar{\tau}$) and sample variance (s^2) to estimate the mean and variance of the whole population.

Let $\{\tau_1, \tau_2, \dots, \tau_n\}$ be a random sample of size n from a population with mean μ and variance σ^2 . The sample mean [26] is defined to be $\bar{\tau} = \frac{1}{n} \sum_{i=1}^n \tau_i$, and the sample variance [26] is $s^2 = 1/(n-1) \sum_{i=1}^n (\tau_i - \bar{\tau})^2$. Naturally, for a large sample, the sample mean will approach the population mean $\mu = \lim_{n \rightarrow \infty} \bar{\tau}$.

Similarly, for the population of the sample mean $\bar{\tau}$, if $\{\bar{\tau}_1, \bar{\tau}_2, \dots, \bar{\tau}_N\}$ are N measurements of $\bar{\tau}$, we have that an estimate of the mean of the sample means is $\mu_{\bar{\tau}} = 1/N \sum_{i=1}^N \bar{\tau}_i$, and that an estimate of the variance of the sample means is $\sigma_{\bar{\tau}}^2 = 1/(N-1) \sum_{i=1}^N (\bar{\tau}_i - \mu_{\bar{\tau}})^2$.

For a large enough sample size N , the central limit theorem [26] states that the distribution of the sample mean $\bar{\tau}$ will approach a normal distribution, no matter what the underlying distribution of τ is, as long as the distribution of τ has a finite variance σ^2 , and as long as the τ_i are independent. We thus have $\bar{\tau} \sim \text{Normal}(\mu, \sigma^2/N)$, i.e., for the mean and variance of the sample mean [26]

$$\begin{aligned} \lim_{N \rightarrow \infty} \mu_{\bar{\tau}} &= \mu, \\ \lim_{N \rightarrow \infty} \sigma_{\bar{\tau}}^2 &= \frac{\sigma^2}{N}. \end{aligned} \quad (3)$$

Although the mean of the distribution of $\bar{\tau}$ is identical to the mean of the underlying distribution, the variance of the distribution of $\bar{\tau}$ is much smaller for large sample sizes since there is a division by N . In the rest of this paper we will refer to $\bar{\tau}$ as ‘mean waiting time.’

F. Poisson process

The self-diffusion process in solid ion conductors like the one studied in this work is described by a Poisson process. This implies that the diffusive events are independent and occur at a rate $\lambda > 0$ and that the waiting time between the jumps should be exponentially distributed, $\tau \sim \text{Exp}(\lambda)$. The probability density function for the exponential distribution is [27]

$$f(\tau; \lambda) = \begin{cases} \lambda \cdot e^{-\lambda \cdot \tau} & \text{if } \tau \geq 0, \\ 0, & \text{otherwise.} \end{cases} \quad (4)$$

The mean of an exponentially distributed random variable τ is given by $E[\tau] = \frac{1}{\lambda}$, and the variance is given by $\text{Var}[\tau] = \frac{1}{\lambda^2}$, where λ is the rate parameter of the distribution [27].

For a sample of N exponentially distributed independent random variables with a rate parameter λ , we define the error in the mean $\epsilon_{\bar{\tau}}$ resulting from the (finite) sample size as

$$\epsilon_{\bar{\tau}} = \frac{\sigma_{\bar{\tau}}}{\mu_{\bar{\tau}}} = \frac{\sigma/\sqrt{N}}{\mu} = \frac{1}{\sqrt{N}}, \quad (5)$$

where we assume that the central limit theorem holds and apply Eq. (3) with $\sigma = \mu = 1/\lambda$. For the extent of this paper, errors will be derived from this definition.

III. RESULTS

In this section we establish that the KMC method and the NEMD method capture the same diffusion process in our doped ceria system, and we calculate the error (numerically) in mean waiting time resulting from limited statistics using KMC (Sec. III A). Thereafter, we show that the mean waiting time is (inversely) proportional to the ion flux (Sec. III B), which means that the error in mean waiting time is the same as the error in ion flux. Finally, we derive an analytical expression for the error in mean waiting time (i.e., error in flux) resulting from limited statistics, and the analytical error in flux is compared to the error calculated numerically (Sec. III C).

A. Two methods, the same process

Given that the vibrational motion of the diffusing atoms about the ideal lattice positions integrates out, and that a diffusing atom spends enough time to thermalize, and thus “loose memory” of the previous diffusive event, before the next occurs; a MD simulation of diffusion will capture the Markovian [26] nature of the Poisson process. In KMC, the simulation of a Poisson process is by construction. The characteristic exponential distribution of waiting times between diffusive events can be probed by plotting the waiting times.

The distribution of the waiting time (time between diffusive jumps) obtained from the NEMD simulation for the color field 0.5 eV/Å with 208 diffusive jumps is shown in Fig. 2(a). The distribution obtained in the KMC simulations with 10 000 000

FIG. 2. Waiting time distributions originating from: (a) NEMD simulation with field strength $F_{\text{ext}} = 0.5 \text{ eV/\AA}$ (distributions obtained for other color field strengths in the linear regime are similar), and (b) a KMC simulation of $n = 10\,000\,000$ steps with prefactor $\nu = 5 \times 10^{12} \text{ Hz}$. The waiting times are binned and presented in a histogram (green bars). Red solid line shows a fitted exponential function. Black dashed vertical line shows the mean waiting time $\bar{\tau}$.

steps is shown in Fig. 2(b). In both figures an exponential distribution corresponding to Eq. (4) is fitted to the data [red lines in Figs. 2(a) and 2(b)]. Even though the data is binned differently for each case, we can see that, qualitatively, the two methods reproduce to the same (exponential) distribution and thus describe the same process: the Poisson process. This provides a rationale for using the KMC method as a statistical reference for the NEMD method.

Given that the computationally heavy DFT calculations to calculate the diffusion barriers have been performed beforehand, then the subsequent KMC simulations are very fast. Therefore, the KMC method allows for the gathering of vast statistics. We have used a KMC simulation with $n = 10\,000\,000$ steps to obtain the reference mean waiting time between jumps, which is found to be $\bar{\tau} = 3.90 \text{ ps}$ with prefactor $\nu = 5 \times 10^{12} \text{ Hz}$. The convergence of the mean

FIG. 3. Red solid line is mean waiting time versus simulation length for a KMC simulation with prefactor $\nu = 5 \times 10^{12} \text{ Hz}$. Dashed black line is the mean waiting over the whole interval.

waiting time for this calculation can be seen in Fig. 3. One should note that the choice of exponential prefactor ν affects the resulting mean waiting time of the KMC simulation. A prefactor $\nu = 3 \times 10^{12} - 6 \times 10^{12} \text{ Hz}$ gives a mean waiting time $\bar{\tau}(\nu) = 3.2 - 6.6 \text{ ps}$ (see Fig. 4).

At the same time, the mean waiting time from the NEMD simulation depends on the choice of external field F_{ext} . The mean waiting time from KMC for different prefactors, and the mean waiting time from NEMD for different field strengths are shown in Fig. 4. The extrapolated mean waiting time at zero field from the NEMD simulation is $\bar{\tau}_{\text{NEMD}}(F_{\text{ext}} = 0) \approx 6 \text{ ps}$, which is in the range of the KMC mean waiting times for reasonable choices of prefactors, suggesting that the application of the field effectively can be seen as a time scaling. The extrapolated NEMD mean waiting time corresponds to a prefactor of about $3.4 \times 10^{12} \text{ Hz}$. One can also notice that the scaling is linear for the color field strengths in the linear regime (0.3, 0.4 and 0.5 $\text{eV/\AA}$) but not for higher fields (0.6 $\text{eV/\AA}$) when the time between jumps starts to approach zero. The dashed vertical line in Fig. 2(a) corresponds to the blue dot at $F_{\text{ext}} = 0.5 \text{ eV/\AA}$ in Fig. 4.

Now we are in the position to assess the error in the mean waiting time related to the limited statistics. To do this we performed KMC simulations with a small number

FIG. 4. Mean waiting time $\bar{\tau}$ vs field strength from NEMD (blue dots). The dashed line is a linear fit to the points in the linear response regime, i.e., $F_{\text{ext}} \leq 0.5 \text{ eV/\AA}$. Triangles are mean waiting time from KMC (zero field) with prefactors: 3×10^{12} , 4×10^{12} , 5×10^{12} , $6 \times 10^{12} \text{ Hz}$, from top to bottom. Error bars are calculated according to Eq. (11) and lie within the plot symbols if not visible.

FIG. 5. Histogram of mean waiting time $\bar{\tau}$ for $N = 50\,000$ KMC simulations with prefactor $\nu = 5 \times 10^{12}$ Hz (green bars). The mean $\mu_{\bar{\tau}}$ and the standard deviation $\sigma_{\bar{\tau}}$ of the mean waiting time is shown in each figure. The number in parenthesis is $\sigma_{\bar{\tau}}/\mu_{\bar{\tau}}$ expressed in percent. Vertical dashed lines mark $\mu_{\bar{\tau}} \pm \sigma_{\bar{\tau}}$. The red solid line is the normal distribution $Normal(\mu_{\bar{\tau}}, \sigma_{\bar{\tau}}^2)$. The subfigures are for different simulation lengths (diffusive jumps) n .

of steps (diffusive jumps) $n = \{50, 100, 150, 200, 500, 1000\}$. The choices of n here were motivated by the number of jumps accessible in NEMD simulations. For each simulation length n we performed $N = 50\,000$ independent simulations. The mean waiting time $\bar{\tau}_1, \bar{\tau}_2, \dots, \bar{\tau}_N$ for each simulation was calculated as $\bar{\tau} = \frac{1}{n} \sum_i^n \tau_i$. The distribution of $\bar{\tau}$'s for different n is shown in Fig. 5. The mean $\mu_{\bar{\tau}}$ and the standard deviation $\sigma_{\bar{\tau}}$ of the mean waiting time for these simulations are shown in Fig. 5. One can see that the error varies from $\pm 14\%$ for 50 jumps to just $\pm 3\%$ for 1000 jumps. In our calculations of ion conductivity in Ref. [7] we counted 208 diffusive jumps at $T = 1063$ K and $F_{\text{ext}} = 0.5$ eV/Å for a simulation of $\sim 40\,000$ cpu hours. From Fig. 5 we see that this statistics corresponds to an error of about $\pm 7\%$. Moreover, to get the error down to $\pm 3\%$ we would require about 1000 jumps, which in turn would require a simulation of some 200 000 cpu hours (about 22 cpu years). In Sec. III B and III C we will show that the error in time directly translates to the error in flux and conductivity.

B. Defining ion flux through mean waiting time

In a typical molecular dynamics simulation of diffusion in a solid, an atom will spend most of its time vibrating about an ideal lattice position. Eventually, the atom will gain enough thermal energy to overcome the diffusion barrier and move to a neighboring vacant site. This can be exploited to reformulate an expression for the color flux.

Unaltered, the expression for the time-averaged color flux in NEMD, via Eq. (1), is

$$\langle J \rangle = \frac{1}{V} \frac{1}{K} \sum_{i,k}^{I,K} c_i v_{i,k}, \quad (6)$$

where c_i is the color charge of atom i , k is the k th step in the simulation, and $v_{i,k}$ is the velocity of atom i at step k . The number of atoms is I , and the number of MD time steps is K . V is the volume of the cell. We can without loss of accuracy look *only* at the movement of the jumping atom (index i disappears). Since we are working with a numerical method, velocities are approximated as $v_k = \left(\frac{dr}{dt}\right)_k = \frac{\Delta r_k}{\Delta t_k}$, where r is displacement and t time. Furthermore, since time is discretized in steps of equal length, $\Delta t_k = \Delta t$, we can write

$$\begin{aligned} \langle J \rangle &= \frac{c}{VK} \left(\left(\frac{\Delta r}{\Delta t} \right)_1 + \left(\frac{\Delta r}{\Delta t} \right)_2 + \dots + \left(\frac{\Delta r}{\Delta t} \right)_K \right) \\ &= \frac{c}{V} \frac{1}{K \Delta t} (\Delta r_1 + \Delta r_2 + \dots + \Delta r_K). \end{aligned} \quad (7)$$

The total time $K \Delta t$ can be expressed in the times between the jumps τ_j as $K \Delta t = \sum_j^n \tau_j$, where n is the total number of jumps. The total displacement, $(\Delta r_1 + \Delta r_2 + \dots + \Delta r_K) = \sum_i^K \Delta r_i$, can also similarly be expressed in the jump lengths δr_j as $\sum_i^K \Delta r_i = \sum_j^n \delta r_j = n \cdot \langle d \rangle$, where $\langle d \rangle$ is the average jump length. Finally, we can write the mean flux as

$$\langle J \rangle = \frac{c}{V} \frac{n \langle d \rangle}{\sum_j^n \tau_j} = \frac{c \langle d \rangle}{V} \frac{1}{\langle \tau \rangle}. \quad (8)$$

For our system [7] we have $V = 1309.3 \text{ Å}^3$, $c = +1$, $\langle d \rangle = a/2 = 2.735 \text{ Å}$, which is the ideal oxygen-oxygen nearest neighbor distance in our simulation cell. The mean waiting time ($\bar{\tau}$) is $\langle \tau \rangle = 0.62$ ps for field strength $F_{\text{ext}} = 0.5$ eV/Å. Putting this into Eq. (8) and projecting onto the [111] direction of the field gives $\langle J \rangle_{\bar{\tau}} = 0.19 \times 10^{-2} \text{ Å}^{-2} \text{ ps}^{-1}$. In comparison, calculating the flux via the velocities (v) using Eq. (1) gives $\langle J \rangle_v = 0.15 \times 10^{-2} \text{ Å}^{-2} \text{ ps}^{-1}$. The discrepancy between the flux values from the two different methods is the consequence of the approximation we used in calculating the mean waiting time (see Sec. II C). Calculating the ion conductivity from the mean waiting time using Eq. (2) $Q_{\bar{\tau}} = 25 \times 10^{-2} \text{ S cm}^{-1}$, and calculating the ion conductivity from the velocities using Eq. (2) gives $Q_v = 19 \times 10^{-2} \text{ S cm}^{-1}$. Again, as with the flux, a small discrepancy between Q_v and $Q_{\bar{\tau}}$ is observed.

Note that the purpose of this exercise of redefining the ion flux through the mean waiting time is the following: There is no straightforward way to access the statistical error in Q_v , but for the error in $Q_{\bar{\tau}}$ there is. An expression for the error in $Q_{\bar{\tau}}$ will be derived below.

TABLE I. Statistical error for simulations of different length n . $\epsilon_{(J)}^{\text{num}}$ is the numerical relative error and $\epsilon_{(J)}^{\text{analyt}}$ is the analytical relative error defined in this section.

n	$\epsilon_{(J)}^{\text{num}}$ [%]	$\epsilon_{(J)}^{\text{analyt}}$ [%]
50	-12.5, 16.7	-12.4, 16.5
100	-9.2, 11.3	-9.0, 11.1
150	-7.7, 9.0	-7.5, 8.9
200	-6.7, 7.7	-6.6, 7.6
500	-4.3, 4.7	-4.3, 4.7
1000	-3.1, 3.3	-3.1, 3.3

C. Analytical and numerical ways to determine statistical error

The error in flux resulting from the finite sample size (n) can be estimated both numerically and analytically from the mean waiting time ($\mu_{\bar{\tau}}$), the standard deviation of the mean ($\sigma_{\bar{\tau}}$), and the error in the mean ($\epsilon_{\bar{\tau}}$). The relative error in the flux is defined as

$$\epsilon_{(J)} = \frac{\langle J \rangle^{\text{calc.}} - \langle J \rangle^{\text{true}}}{\langle J \rangle^{\text{true}}}, \quad (9)$$

where $\langle J \rangle^{\text{calc.}}$ is the calculated or measured mean flux (corresponding to ‘sample’ as introduced in Sec. II E), and $\langle J \rangle^{\text{true}}$ is the true mean flux (corresponding to ‘population’ as introduced in Sec. II E). Using the definition of the flux from Eq. (8) with $\langle \tau \rangle^{\text{true}} = \mu_{\bar{\tau}}$, and $\langle \tau \rangle^{\text{calc.}} = \mu_{\bar{\tau}} \pm \epsilon_{\bar{\tau}} \mu_{\bar{\tau}}$, the relative error in the flux, Eq. (9), becomes

$$\epsilon_{(J)} = \frac{\mp \epsilon_{\bar{\tau}}}{1 \pm \epsilon_{\bar{\tau}}}. \quad (10)$$

Up to here the analysis is completely general, i.e., the same for the numerical and analytical treatment of the error.

Numerically, we calculate the relative error in the flux using $\epsilon_{\bar{\tau}} = \sigma_{\bar{\tau}}/\mu_{\bar{\tau}}$ from Eq. (5), where $\mu_{\bar{\tau}}$ and $\sigma_{\bar{\tau}}$ are calculated as described in Sec. II E. The resulting numerical relative errors $\epsilon_{(J)}^{\text{num}}$ are tabulated in Table I.

The error in the flux can also be estimated analytically. Using $\epsilon_{\bar{\tau}} = 1/\sqrt{n}$ from Eq. (5) in Eq. (10) we get

$$\epsilon_{(J)}^{\text{analyt}} = \frac{\mp 1}{\sqrt{n} \pm 1}. \quad (11)$$

The analytical relative errors calculated using this expression are presented in Table I for different simulation lengths,

alongside the corresponding numerical relative errors. It is clear that the two approaches give the same error estimation, but where the assessment of the numerical error required 50 000 samples, the analytical error estimation is instantly available. Note that, following the treatment here, *any* quantity that can be expressed in terms of the inverse mean waiting time will have an error $\epsilon_{(J)}^{\text{analyt}}$ as given by Eq. (11).

IV. CONCLUSIONS

The ability to gauge the impact of limited statistics on the error in calculated observables is important in any computational study of diffusion in solids. This is especially true in the context of *ab initio* computational methods to study diffusion, where computational cost is a limiting factor. Knowing the statistical error can also help balancing computational accuracy against statistical quality when choosing the most appropriate computational method for a specific system.

Our study demonstrates that both the NEMD method and the KMC method describe a Poisson process for our particular system. Based on this assertion, we use the KMC method to numerically verify the analytical expression for the relative error of the flux or conductivity suggested by us: $\epsilon_{(J)}^{\text{analyt}} = \mp 1/(\sqrt{n} \pm 1)$, where n is the number of diffusive events. It is worth pointing out that the numerical assessment of the error required 50 000 samples, whereas the analytical error estimate is instantly available. The analysis presented here is not limited to any specific method like MD or KMC, but is valid in the context of simulation of Poisson processes in general.

ACKNOWLEDGMENTS

N.V.S., O.Y.V., and J.O.N. acknowledge the financial support of Swedish Energy Agency (STEM) (Project No. 35515-1), Carl Tryggers Foundation (Project No. CTS 14:433), and the Swedish Research Council (VR) (Project No. 2014-5993). The support from the Swedish Research Council (VR) Project No. 2011-42-59, LiLi-NFM, and the Swedish Government Strategic Research Area in Materials Science on Functional Materials at Linköping University (Faculty Grant SFO-Mat-LiU No. 2009 00971) are acknowledged by S.I.S. We also thank the Swedish National Infrastructure for Computing (SNIC) for providing computational resources.

-
- [1] D. Frenkel and B. Smit, in *Understanding Molecular Simulation*, edited by D. Frenkel and B. Smit (Academic Press, San Diego, 2002) 2nd ed., p. 63.
 - [2] D. Marx and J. Hutter, *Modern Methods and Algorithms of Quantum Chemistry* (John von Neumann Institute for Computing, Jülich, 2000), pp. 301–449.
 - [3] A. F. Voter, *Phys. Rev. B* **57**, R13985 (1998).
 - [4] A. F. Voter, *J. Chem. Phys.* **106**, 4665 (1997).
 - [5] M. R. Sørensen and A. F. Voter, *J. Chem. Phys.* **112**, 9599 (2000).
 - [6] D. J. Evans and G. P. Morriss, *Statistical Mechanics of Nonequilibrium Liquids* (Cambridge University Press, Cambridge, 2008).
 - [7] J. O. Nilsson, O. Y. Vekilova, O. Hellman, J. Klarbring, S. I. Simak, and N. V. Skorodumova, *Phys. Rev. B* **93**, 024102 (2016).
 - [8] A. Bortz, M. Kalos, and J. Lebowitz, *J. Comput. Phys.* **17**, 10 (1975).
 - [9] D. T. Gillespie, *J. Comput. Phys.* **22**, 403 (1976).
 - [10] A. F. Voter, *Radiation Effects in Solids* (Springer Netherlands, Dordrecht, 2007), Chap. 1, p. 1.

- [11] R. Zwanzig and N. K. Ailawadi, *Phys. Rev.* **182**, 280 (1969).
- [12] G. Pranami and M. H. Lamm, *J. Chem. Theory Comput.* **11**, 4586 (2015).
- [13] H. Müller-Krumbhaar and K. Binder, *J. Stat. Phys.* **8**, 1 (1973).
- [14] B. C. H. Steele and A. Heinzel, *Nature (London)* **414**, 345 (2001).
- [15] P. Hohenberg and W. Kohn, *Phys. Rev.* **136**, B864 (1964).
- [16] W. Kohn and L. J. Sham, *Phys. Rev.* **140**, A1133 (1965).
- [17] G. Kresse and J. Hafner, *Phys. Rev. B* **48**, 13115 (1993).
- [18] G. Kresse and J. Furthmüller, *Phys. Rev. B* **54**, 11169 (1996).
- [19] G. Kresse and J. Furthmüller, *Comput. Mater. Sci.* **6**, 15 (1996).
- [20] P. E. Blöchl, *Phys. Rev. B* **50**, 17953 (1994).
- [21] J. P. Perdew, K. Burke, and M. Ernzerhof, *Phys. Rev. Lett.* **77**, 3865 (1996).
- [22] P. C. Aeberhard, S. R. Williams, D. J. Evans, K. Refson, and W. I. F. David, *Phys. Rev. Lett.* **108**, 095901 (2012).
- [23] M. Leetmaa and N. V. Skorodumova, *Comput. Phys. Commun.* **185**, 2340 (2014).
- [24] M. Leetmaa and N. V. Skorodumova, *Comput. Phys. Commun.* **196**, 611 (2015).
- [25] M. Leetmaa and N. V. Skorodumova, *Comput. Phys. Commun.* **191**, 119 (2015).
- [26] D. Zwillinger, *CRC Standard Mathematical Tables and Formulae*, 32nd ed. (CRC Press, Boca Raton, 2011), p. 507.
- [27] W. Feller, *An Introduction to Probability Theory and Its Applications*, 2nd ed., Vol. II (Wiley, New York, 1971), p. 8.