



ELSEVIER

Contents lists available at ScienceDirect

Journal of Computational Physics

www.elsevier.com/locate/jcp



Quantification of sampling uncertainty for molecular dynamics simulation: Time-dependent diffusion coefficient in simple fluids

Changho Kim^a, Oleg Borodin^b, George Em Karniadakis^{a,*}^a Division of Applied Mathematics, Brown University, 182 George St., Providence, RI 02912, USA^b Electrochemistry Branch, Army Research Laboratory, 2800 Powder Mill Rd., Adelphi, MD 20783, USA

ARTICLE INFO

Article history:

Received 12 March 2015

Received in revised form 27 July 2015

Accepted 10 September 2015

Available online 16 September 2015

Keywords:

Statistical error

Velocity autocorrelation function

Mean-squared displacement

Uncertainty quantification

Gaussian process approximation

Self-diffusion

ABSTRACT

We analyze two standard methods to compute the diffusion coefficient of a tracer particle in a medium from molecular dynamics (MD) simulation, the velocity autocorrelation function (VACF) method, and the mean-squared displacement (MSD) method. We show that they are equivalent in the sense that they provide the same mean values with the same level of statistical errors. We obtain analytic expressions for the level of the statistical errors present in the time-dependent diffusion coefficient as well as the VACF and the MSD. Under the assumption that the velocity of the tracer particle is a Gaussian process, all results are expressed in terms of the VACF. Hence, the standard errors of all relevant quantities are computable once the VACF is obtained from MD simulation. By using analytic models described by the Langevin equations driven by Gaussian white noise and Poissonian white shot noise, we verify our theoretical error estimates and discuss the non-Gaussianity effect in the error estimates when the Gaussian process approximation does not hold exactly. For validation, we perform MD simulations for the self-diffusion of a Lennard-Jones fluid and the diffusion of a large and massive colloid particle suspended in the fluid. Our theoretical framework is also applicable to mesoscopic simulations, e.g., Langevin dynamics and dissipative particle dynamics.

© 2015 Elsevier Inc. All rights reserved.

1. Introduction

The estimation of transport coefficients is one of the most notable applications of the molecular dynamics (MD) simulation technique [1–5]. They can be calculated through nonequilibrium MD simulation methods [5], which are direct and intuitive but require the generation of nonequilibrium steady states. Alternatively, the transport coefficients can be estimated from equilibrium MD simulations by using relations derived from statistical mechanics [1–4].

For diffusion, the corresponding transport coefficient is the diffusion coefficient D . Phenomenologically, D is defined through Fick's law, which leads to the diffusion equation:

$$\frac{\partial}{\partial t} \rho(\mathbf{x}, t) = D \Delta \rho(\mathbf{x}, t), \quad (1)$$

* Corresponding author.

E-mail address: george_karniadakis@brown.edu (G.E. Karniadakis).

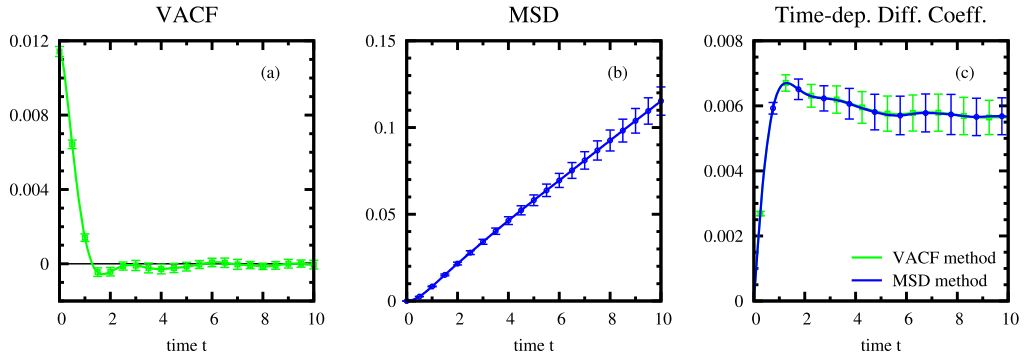


Fig. 1. Typical MD simulation results for the evaluation of the diffusion coefficient of a tracer particle in a fluid by the VACF and MSD methods. The diffusion coefficient is calculated either from the time integral of the VACF or from the long-time slope of the MSD, see Eq. (5). For detailed description of the system, see Section 4.1. The standard errors are estimated from 16 independent MD runs and the error bars denote two standard deviations.

where $\rho(\mathbf{x}, t)$ is the (normalized) concentration of a diffusing species. Then, the second moment of $\rho(\mathbf{x}, t)$ increases linearly with time [3]:

$$\frac{d}{dt} \langle \mathbf{x}^2(t) \rangle_{\rho} = \frac{d}{dt} \int \mathbf{x}^2 \rho(\mathbf{x}, t) d\mathbf{x} = 2dD, \quad (2)$$

where d is the dimensionality of the space. Microscopically, the diffusion coefficient D of a tracer particle in a medium is defined by the long-time slope of its mean-squared displacement (MSD) $\langle [\mathbf{x}(t) - \mathbf{x}(0)]^2 \rangle$:

$$D = \frac{1}{2d} \lim_{t \rightarrow \infty} \frac{\langle [\mathbf{x}(t) - \mathbf{x}(0)]^2 \rangle}{t} = \frac{1}{2d} \lim_{t \rightarrow \infty} \frac{d}{dt} \langle [\mathbf{x}(t) - \mathbf{x}(0)]^2 \rangle, \quad (3)$$

where $\mathbf{x}(t)$ denotes the position vector of the particle and the brackets denote the equilibrium average. In this case, diffusion occurs in the absence of concentration gradient (i.e., in equilibrium) and it is due to a spontaneous mixing of molecules. It is also noted that although Eq. (3) looks similar to Eq. (2), on the microscopic time scale, it takes time for the MSD to exhibit linear growth with time. Since the MSD can be expressed as a double time integral of the velocity autocorrelation function (VACF) $\langle \mathbf{v}(0) \cdot \mathbf{v}(t) \rangle$ [3], D is also expressed as

$$D = \frac{1}{d} \int_0^{\infty} \langle \mathbf{v}(0) \cdot \mathbf{v}(t) \rangle dt, \quad (4)$$

where $\mathbf{v}(t)$ denotes the velocity vector of the particle. Hence, in equilibrium MD simulation, the diffusion coefficient is calculated either by calculating the MSD and using Eq. (3) or by calculating the VACF and using Eq. (4). More specifically, the time-dependent diffusion coefficient $D(t)$, which is defined as

$$D(t) = \frac{1}{2d} \frac{d}{dt} \langle [\mathbf{x}(t) - \mathbf{x}(0)]^2 \rangle = \frac{1}{d} \int_0^t \langle \mathbf{v}(0) \cdot \mathbf{v}(t') \rangle dt', \quad (5)$$

is calculated either from the MSD or from the VACF and then D is estimated from the long-time plateau value of $D(t)$, see Fig. 1.

Similarly, other transport coefficients can be estimated by equilibrium MD simulation through the following two types of relations. In the Green–Kubo relations [6,7], the transport coefficients are expressed as the time integrals of corresponding time autocorrelation functions (as in Eq. (4)). On the other hand, in the Einstein–Kubo–Helfand relations [8–10], they are expressed as the long-time slopes of the mean-squared differences of corresponding Helfand moments (as in Eq. (3)).

1.1. Computational considerations

When transport coefficients are computed from an equilibrium MD simulation, various types of errors may be involved. One possible source of errors is inaccuracy of the MD simulation model [11], which is due to approximate interaction potential function [12], finite system size [13,14], large time step size, and insufficiently equilibrated initial configuration, to name but a few. Another source of errors comes from evaluation methods for transport coefficients: truncation of the time integral in the Green–Kubo relations or evaluation of the slope at a finite time in the Einstein–Kubo–Helfand relations. These two types of errors are reproducible and called *systematic errors*. On the other hand, the estimated values of transport coefficients are subject to random fluctuations as the estimated time correlation functions or mean-squared differences are.

Since the latter quantities are defined in terms of equilibrium averages of the system and the equilibrium MD simulation technique is based on sampling method, finite-sampling inevitably introduces *statistical errors*.

The sampling uncertainty can be quantified in terms of the standard error (i.e., the standard deviation of the statistical error). It has been known that when the length of a sample trajectory, \mathcal{T} , or the number of independent sample trajectories, \mathcal{N} , increases, the standard error of an averaged quantity decreases like $\mathcal{T}^{-1/2}$ or $\mathcal{N}^{-1/2}$ [1,3,15,16]. It is noted that, in order to computationally investigate systematic errors, the level of statistical errors should be suppressed by increasing the values of \mathcal{T} and/or \mathcal{N} . Also, quantifying statistical errors is crucial for the accurate estimation of transport coefficients. More generally, for a total uncertainty quantification analysis, we need to account both for parametric uncertainties, e.g., the form of potentials, boundary conditions, etc., as well as sampling errors. Recently, an uncertainty quantification framework based on ANOVA (analysis of variance) and Sobol's variance decomposition has been proposed for the separation of parametric uncertainty from statistical uncertainty due to intrinsic fluctuations of the system [17].

In the present paper, we mainly investigate statistical errors by theoretical and computational means. We demonstrate our approach by analyzing the VACF and MSD methods for the diffusion coefficient D of a tracer particle in a fluid. We introduce the motivating questions for the present paper by answering a natural question at this point: which method is better? To this end, in Fig. 1, we compare typical MD simulation results obtained by the two methods. The standard errors of each method were estimated from a number of independent MD runs. These results suggest that the VACF and MSD methods seem to produce the same mean values with the same level of the statistical errors. In fact, the coincidence is not accidental but is typically observed in various molecular systems.

We shall confirm the equivalence of the two methods by addressing the following two questions.

- How are the statistical errors in the original MD data (i.e., the VACF or the MSD) propagated into the time-integrated or time-differentiated data (i.e., $D(t)$)?

It is important to note that, in principle, even if we know the standard errors (i.e., error bars) of the VACF or the MSD, we cannot calculate the standard errors (i.e., error bars) of $D(t)$. This is because the statistical errors, $\varepsilon(t_1)$ and $\varepsilon(t_2)$, at different times in the original data are correlated with each other and thus we need to know the *error correlation function* $\langle \varepsilon(t_1)\varepsilon(t_2) \rangle$. We shall derive relations between the standard errors of $D(t)$ and the error correlation functions of the VACF and the MSD. Hence, from the latter functions, we can calculate the standard errors of $D(t)$ as well as the VACF and the MSD.

- Can we calculate the error correlation functions of the VACF and the MSD from the VACF under some reasonable assumptions?

If this is the case, once the VACF is calculated in order to estimate $D(t)$, the standard errors of both quantities can be readily estimated without an additional MD ensemble run. For various averaging procedures such as time-averaging and ensemble-averaging, we shall derive expressions for the error correlation functions of the VACF and the MSD. Combining these expressions with the aforementioned relations, we shall obtain expressions for the standard errors of $D(t)$.

1.2. Theoretical considerations

Theoretical analysis on the statistical errors in MD simulation data dates back to early days of MD simulations. Zwanzig and Ailawadi [16] have investigated the statistical errors occurring in the evaluation of time correlation function $\langle A(0)A(t) \rangle$ for a general property A . For time-averaging procedure over a long time interval of length \mathcal{T} , they have obtained an expression for the error correlation function of $\langle A(0)A(t) \rangle$ in terms of $\langle A(0)A(t) \rangle$. As expected, the error correlation function has been shown to be inversely proportional to \mathcal{T} . In the derivation, they have introduced the Gaussian process approximation (GPA) of the underlying process $A(t)$. Since all high-order statistics can be expressed in terms of the second-order statistics under the GPA, they have expressed the four-time correlation function $\langle A(0)A(t_1)A(t_2)A(t_3) \rangle$ in terms of $\langle A(0)A(t) \rangle$. Frenkel [18] has applied their results to two types of correlation functions for a system consisting of N identical particles. For a correlation function of a single-particle property (e.g., the VACF), he has shown that averaging over particles also introduces the $N^{-1/2}$ -scaling behavior in the standard errors. However, for an N -particle property (e.g., the total dipole moment), he has concluded that such a scaling behavior is not expected. Recently, Jones and Mandadapu [19] have employed the error analysis of Zwanzig and Ailawadi [16] and Frenkel [18] and investigated the statistical errors present in the time integral of a time correlation function. They have obtained an upper bound for the variance of the statistical error and employed it for the development of an adaptive Green–Kubo methodology to calculate transport coefficients. To the best of our knowledge, despite extensive use of the MSD method, systematic error analysis of the method has not been available.

Through our theoretical approach, we present a complete systematic analysis of the statistical errors occurring in the VACF and MSD methods. By assuming the GPA and following Zwanzig and Ailawadi [16] and Frenkel [18], we obtain *explicit* expressions for the standard errors of the time-dependent diffusion coefficient $D(t)$ as well as the VACF and the MSD. Besides the time-averaging procedure over a single sample trajectory, we analyze various averaging procedures, including averaging over independent sample trajectories and averaging over identical particles.

1.3. Comparison

In our computational approach, we confirm our theoretical results as follows. First, we consider the Langevin equations driven by Gaussian white noise (GWN) and Poissonian white shot noise (PWSN) [20,21], which allow closed-form expressions for the standard errors. Since the GWN-driven Langevin model corresponds to the case that the GPA is exactly valid, we *verify* our theoretical error estimates by comparing them with the simulation results. For the PWSN-driven Langevin model, where the GPA does not hold completely, we observe the deviation in our theoretical error estimates. By changing a parameter of PWSN so that it converges to GWN, we investigate the deviation and discuss a non-Gaussianity indicator for our theoretical predictions. Then, we *validate* our theoretical error estimates by performing MD simulations for the self-diffusion in a Lennard-Jones (LJ) fluid and the tracer diffusion of a large and massive colloidal particle in the fluid. In other words, for these molecular systems, we computationally investigate the validity of the GPA and the agreement of our theory with the simulation results. These systems have been extensively investigated and the diffusion coefficient has been reported for various fluid states and various tracer-particle parameters. For recent MD simulation studies on the self-diffusion of pure LJ fluids, we refer the reader to Refs. [22,23]. For the tracer diffusion in LJ fluids, we refer to Refs. [24–27]. We note that, despite extensive MD studies on these systems, systematic estimation of the standard errors in estimated transport coefficients has not been performed.

In order to accurately estimate the level of the statistical errors, we perform large-sized MD ensemble runs. With the help of parallel computing techniques, independent sample trajectories are concurrently simulated, MD simulation results are collected, and the mean values and standard errors are computed. We do not use the block averaging technique [28], which has been used for estimating the standard error of a simple (i.e., time-independent) equilibrium average [1,3,4]. However, we shall discuss a relevant concept, which is similar to the statistical inefficiency in correlated data [29]. The latter indicates the time difference of two time points in a time series which can be considered to be effectively uncorrelated.

The rest of the present paper is organized as follows. In Section 2, we present our theoretical results, including the statistical error estimates for the VACF, the MSD, and the time-dependent diffusion coefficient. In Section 3, we verify our theoretical results by using the Langevin models driven by GWN and PWSN. In Section 4, we perform two kinds of MD simulations and validate the GPA and our theoretical estimates. Finally, we conclude the paper by providing a brief summary and discussion in Section 5.

2. Theoretical error estimates

In this section, we obtain theoretical estimates for the level of the statistical errors present in the VACF, the MSD, and the time-dependent diffusion coefficient $D(t)$, which are summarized in [Main Results 1, 2, and 3](#) in Section 2.5.

We obtain our theoretical error estimates as follows. After setting up notations and defining the VACF, the MSD, and the time-dependent diffusion coefficient $D(t)$ in Section 2.1, we consider three types of averaging procedures and formulate our problem by defining estimators representing the VACF and MSD methods in Section 2.2. In Section 2.3, by using these estimators, we derive relations between the standard error of $D(t)$ and the error correlation functions of the VACF and the MSD. In Section 2.4, we discuss the assumptions of our theoretical approach, including the GPA. In Section 2.5, under the GPA, we obtain the error correlation functions of the VACF and the MSD and, by using the relations derived in Section 2.3, we finally obtain expressions for the standard errors of $D(t)$ for the VACF and MSD methods.

2.1. VACF, MSD, and time-dependent diffusion coefficient $D(t)$

For a tracer particle undergoing free Brownian motion, we introduce notations and define its VACF, MSD, and time-dependent diffusion coefficient $D(t)$. Then, we show that the two definitions of $D(t)$ given in terms of the VACF and the MSD are equivalent.

In the rest of the paper, we consider diffusion in one direction. In other words, rather than using the position and velocity vectors \mathbf{x} and \mathbf{v} , we use their components x and v in that direction. Then, the MSD and the VACF are denoted as $\langle [x(t) - x(0)]^2 \rangle$ and $\langle v(0)v(t) \rangle$, respectively, where the ensemble average over the equilibrium distribution of the system is denoted by the brackets. Later, we also use $\text{Var}[\]$ and $\text{Cov}[\ , \]$ to denote the variance and covariance over the equilibrium distribution. In a d -dimensional isotropic case, we have $\langle [\mathbf{x}(t) - \mathbf{x}(0)]^2 \rangle = d\langle [x(t) - x(0)]^2 \rangle$ and $\langle \mathbf{v}(0) \cdot \mathbf{v}(t) \rangle = d\langle v(0)v(t) \rangle$.

The time-dependent diffusion coefficient is defined as

$$D(t) = \frac{1}{2} \frac{d}{dt} \langle [x(t) - x(0)]^2 \rangle = \int_0^t \langle v(0)v(t') \rangle dt'. \quad (6)$$

The equality of the two expressions in Eq. (6) can be readily shown as follows. First, we note that

$$\dot{x}(t) = v(t), \quad x(t) - x(0) = \int_0^t v(t') dt'. \quad (7)$$

Table 1

Summary of the estimators to be analyzed in the present paper. For three types of averaging procedures, their definitions and physical meanings are listed. $x^{(k)}$ and $v^{(k)}$ denote the position and velocity of the k th sample, respectively.

Time-averaging	
$\hat{y}_{\text{time}}(t) = \frac{1}{T} \int_0^T v(t')v(t'+t)dt'$	VACF
$\hat{b}_{\text{time}}(t) = \frac{1}{T} \int_0^T [x(t'+t) - x(t')]^2 dt'$	MSD
$\hat{z}_{\text{time}}(t) = \int_0^t \hat{y}_{\text{time}}(t')dt'$	time integral of VACF
$\hat{a}_{\text{time}}(t_1, t_2) = \frac{\hat{b}_{\text{time}}(t_2) - \hat{b}_{\text{time}}(t_1)}{t_2 - t_1}$	average slope of MSD
$\hat{D}_{\text{time}}^{\text{VACF}}(t) = \hat{z}_{\text{time}}(t)$	$D(t)$ from VACF
$\hat{D}_{\text{time}}^{\text{MSD}}(t) = \frac{1}{2} \lim_{\Delta t \rightarrow 0} \hat{a}_{\text{time}}(t, t + \Delta t)$	$D(t)$ from MSD
Ensemble-averaging	
$\hat{y}_{\text{ens}}(t) = \frac{1}{N} \sum_{k=1}^N v^{(k)}(0)v^{(k)}(t)$	VACF
$\hat{b}_{\text{ens}}(t) = \frac{1}{N} \sum_{k=1}^N [x^{(k)}(t) - x^{(k)}(0)]^2$	MSD
$\hat{z}_{\text{ens}}(t) = \int_0^t \hat{y}_{\text{ens}}(t')dt'$	time integral of VACF
$\hat{a}_{\text{ens}}(t_1, t_2) = \frac{\hat{b}_{\text{ens}}(t_2) - \hat{b}_{\text{ens}}(t_1)}{t_2 - t_1}$	average slope of MSD
$\hat{D}_{\text{ens}}^{\text{VACF}}(t) = \hat{z}_{\text{ens}}(t)$	$D(t)$ from VACF
$\hat{D}_{\text{ens}}^{\text{MSD}}(t) = \frac{1}{2} \lim_{\Delta t \rightarrow 0} \hat{a}_{\text{ens}}(t, t + \Delta t)$	$D(t)$ from MSD
Hybrid ensemble/time-averaging	
$\hat{y}_{\text{hybrid}}(t) = \frac{1}{N^2 T} \sum_{k=1}^N \int_0^T v^{(k)}(t')v^{(k)}(t'+t)dt'$	VACF
$\hat{b}_{\text{hybrid}}(t) = \frac{1}{N^2 T} \sum_{k=1}^N \int_0^T [x^{(k)}(t'+t) - x^{(k)}(t')]^2 dt'$	MSD
$\hat{z}_{\text{hybrid}}(t) = \int_0^t \hat{y}_{\text{hybrid}}(t')dt'$	time integral of VACF
$\hat{a}_{\text{hybrid}}(t_1, t_2) = \frac{\hat{b}_{\text{hybrid}}(t_2) - \hat{b}_{\text{hybrid}}(t_1)}{t_2 - t_1}$	average slope of MSD
$\hat{D}_{\text{hybrid}}^{\text{VACF}}(t) = \hat{z}_{\text{hybrid}}(t)$	$D(t)$ from VACF
$\hat{D}_{\text{hybrid}}^{\text{MSD}}(t) = \frac{1}{2} \lim_{\Delta t \rightarrow 0} \hat{a}_{\text{hybrid}}(t, t + \Delta t)$	$D(t)$ from MSD

In addition, by the invariance of an equilibrium average under the shift of time origin, we have $\langle [x(t) - x(0)]^2 \rangle = \langle [x(0) - x(-t)]^2 \rangle$. Then, after differentiating both sides with respect to t and applying the time shift invariance again, we obtain

$$\langle [x(t) - x(0)] v(t) \rangle = \langle [x(t) - x(0)] v(0) \rangle. \tag{8}$$

By noting that putting the differential and integral operators inside the brackets on both sides of Eq. (6) leads to those of Eq. (8), we show the equality in Eq. (6).

2.2. Formulation: Estimators for VACF, MSD, and $D(t)$

Here, we formulate our problem by introducing estimators that correspond to various procedures performed in MD simulations, such as the calculation of the VACF and the MSD and the time integration and differentiation of these functions. Then, the standard errors in the MD simulation results become equal to the standard deviations of these estimators, which will be calculated in the subsequent subsections.

For three types of standard averaging methods, we first explain how the averaging procedures are performed in practice and then introduce corresponding estimators. After defining estimators for the time integral of the VACF and the average slope of the MSD, we finally introduce estimators for $D(t)$ and show that all estimators are unbiased. In Table 1, we summarize the estimators to be defined in this subsection. In Fig. 2, we present schematic diagrams for three types of averaging procedures to be considered in this subsection.

Time-averaging. In an MD simulation, the VACF and the MSD are usually estimated through time-averaging as follows:

$$\langle v(0)v(n\Delta t) \rangle \approx \frac{1}{N} \sum_{i=0}^{N-1} v(i\Delta t)v((i+n)\Delta t), \tag{9a}$$

$$\langle [x(n\Delta t) - x(0)]^2 \rangle \approx \frac{1}{N} \sum_{i=0}^{N-1} [x((i+n)\Delta t) - x(i\Delta t)]^2, \tag{9b}$$

where $0 \leq n \leq n_{\text{max}}$ and N is the number of sample pairs used for time-averaging. We note that N is usually much larger than n_{max} and the sample trajectory needs to be calculated up to $(N + n_{\text{max}} - 1)\Delta t = O(N\Delta t)$, see Fig. 2(a). In order to analyze these time-averaging procedures, we define the following estimators $\hat{y}_{\text{time}}(t)$ and $\hat{b}_{\text{time}}(t)$, respectively, for the VACF and the MSD:

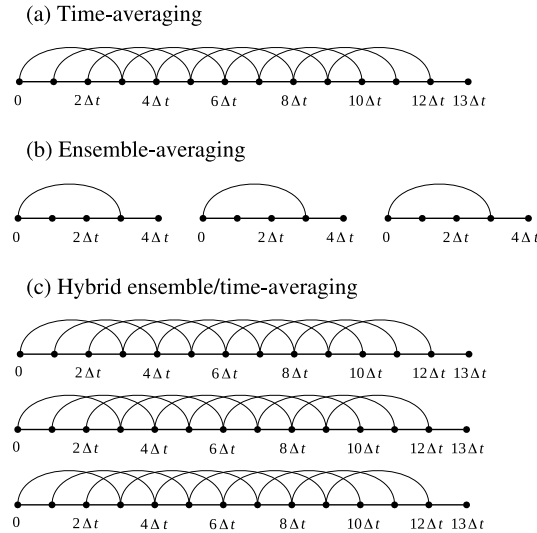


Fig. 2. Schematic diagrams for three types of averaging procedures in VACF/MSD evaluation. The straight lines represent MD sample trajectories and the points on them indicate discrete times $i\Delta t$ ($i = 0, 1, 2, \dots$). We depict the case that VACF/MSD for $t = n\Delta t$ (with $n = 3$) is calculated. The pairs of two time points connected by curly lines are used to calculate the average. In (a) and (c), the value of \mathcal{T} is chosen as $10\Delta t$. In (b) and (c), a total of \mathcal{N} (with $\mathcal{N} = 3$) line segments represent independent MD trajectories. The sample trajectories needed for the calculation of VACF/MSD up to time $n_{\max}\Delta t$ (with $n_{\max} = 4$) are drawn.

$$\hat{y}_{\text{time}}(t) \stackrel{\text{def}}{=} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} v(t')v(t'+t)dt', \quad (10a)$$

$$\hat{b}_{\text{time}}(t) \stackrel{\text{def}}{=} \frac{1}{\mathcal{T}} \int_0^{\mathcal{T}} [x(t'+t) - x(t')]^2 dt'. \quad (10b)$$

We note that, by taking $\Delta t \rightarrow 0$ in Eq. (9) with $\mathcal{T} = N\Delta t$ fixed, we retrieve Eq. (10). In practice, slightly different definitions may be used: for a sample trajectory of length \mathcal{T} , $\frac{1}{\mathcal{T}-t} \int_0^{\mathcal{T}-t} v(t')v(t'+t)dt'$ and $\frac{1}{\mathcal{T}-t} \int_0^{\mathcal{T}-t} [x(t'+t) - x(t')]^2 dt'$. However, under the usual assumption $t \ll \mathcal{T}$, these definitions produce identical asymptotic results for large \mathcal{T} .

Ensemble-averaging. By generating a total of N_{sample} independent equilibrated samples and simulating them up to time $n_{\max}\Delta t$, the VACF and the MSD can be estimated through ensemble-averaging as follows:

$$\langle v(0)v(n\Delta t) \rangle \approx \frac{1}{N_{\text{sample}}} \sum_{k=1}^{N_{\text{sample}}} v^{(k)}(0)v^{(k)}(n\Delta t), \quad (11a)$$

$$\langle [x(n\Delta t) - x(0)]^2 \rangle \approx \frac{1}{N_{\text{sample}}} \sum_{k=1}^{N_{\text{sample}}} [x^{(k)}(n\Delta t) - x^{(k)}(0)]^2, \quad (11b)$$

where the subscript $^{(k)}$ indicates that the labeled quantities belong to the k th sample, see Fig. 2(b). We analyze these ensemble averages by defining the following estimators $\hat{y}_{\text{ens}}(t)$ and $\hat{b}_{\text{ens}}(t)$:

$$\hat{y}_{\text{ens}}(t) \stackrel{\text{def}}{=} \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} v^{(k)}(0)v^{(k)}(t), \quad (12a)$$

$$\hat{b}_{\text{ens}}(t) \stackrel{\text{def}}{=} \frac{1}{\mathcal{N}} \sum_{k=1}^{\mathcal{N}} [x^{(k)}(t) - x^{(k)}(0)]^2. \quad (12b)$$

We have replaced N_{sample} with a general finite-sampling parameter \mathcal{N} in order to also consider averaging over identical particles later. But, in this section, we assume that the processes $\{x^{(k)}(t), v^{(k)}(t)\}_{k=1}^{\mathcal{N}}$ are independent of one another.

Hybrid ensemble/time-averaging. From an ensemble of sample trajectories, the VACF and the MSD can be estimated through the *ensemble averages of the time averages*, see Fig. 2(c). We call them the hybrid ensemble/time-averages. In this case, estimators for the VACF and the MSD are defined as follows:

$$\hat{y}_{\text{hybrid}}(t) \stackrel{\text{def}}{=} \frac{1}{\mathcal{N}\mathcal{T}} \sum_{k=1}^{\mathcal{N}} \int_0^{\mathcal{T}} v^{(k)}(t')v^{(k)}(t'+t)dt', \tag{13a}$$

$$\hat{b}_{\text{hybrid}}(t) \stackrel{\text{def}}{=} \frac{1}{\mathcal{N}\mathcal{T}} \sum_{k=1}^{\mathcal{N}} \int_0^{\mathcal{T}} [x^{(k)}(t'+t) - x^{(k)}(t')]^2 dt'. \tag{13b}$$

Note that if we set $\mathcal{N} = 1$, the hybrid ensemble/time-averaging case becomes the time-averaging case.

Estimators for $D(t)$. Now we define estimators for the time-dependent diffusion coefficient $D(t)$ by considering two types of operations on the VACF and MSD estimators. For estimators $\hat{y}_{\text{ens}}(t)$, $\hat{y}_{\text{time}}(t)$, and $\hat{y}_{\text{hybrid}}(t)$ for the VACF, we define their time integral estimators $\hat{z}_{\text{ens}}(t)$, $\hat{z}_{\text{time}}(t)$, and $\hat{z}_{\text{hybrid}}(t)$ as follows:

$$\hat{z}_{\bullet}(t) \stackrel{\text{def}}{=} \int_0^t \hat{y}_{\bullet}(t')dt', \tag{14}$$

where the bullets indicate that for any type of averaging procedure, including the three averages defined above, its time integral is correspondingly defined. Similarly, for the estimators $\hat{b}_{\bullet}(t)$ for the MSD, we define their average slope estimators over the interval $[t_1, t_2]$ as follows:

$$\hat{a}_{\bullet}(t_1, t_2) \stackrel{\text{def}}{=} \frac{\hat{b}_{\bullet}(t_2) - \hat{b}_{\bullet}(t_1)}{t_2 - t_1}. \tag{15}$$

Then, following Eq. (6), we define estimators for $D(t)$ from the VACF and MSD methods as follows:

$$\hat{D}_{\bullet}^{\text{VACF}}(t) \stackrel{\text{def}}{=} \hat{z}_{\bullet}(t), \tag{16a}$$

$$\hat{D}_{\bullet}^{\text{MSD}}(t) \stackrel{\text{def}}{=} \frac{1}{2} \lim_{\Delta t \rightarrow 0} \hat{a}_{\bullet}(t, t + \Delta t). \tag{16b}$$

Unbiasedness of estimators. Before closing this subsection, we show that all estimators are unbiased. In other words, for the three types of averaging procedures, we have

$$\langle \hat{y}_{\bullet}(t) \rangle = \langle v(0)v(t) \rangle, \tag{17}$$

$$\langle \hat{b}_{\bullet}(t) \rangle = \langle [x(t) - x(0)]^2 \rangle, \tag{18}$$

$$\langle \hat{D}_{\bullet}^{\text{VACF}}(t) \rangle = \langle \hat{D}_{\bullet}^{\text{MSD}}(t) \rangle = D(t). \tag{19}$$

All we need to do is taking the ensemble average on the definition of each estimator and using $\langle v(t')v(t'+t) \rangle = \langle v^{(k)}(0)v^{(k)}(t) \rangle = \langle v(0)v(t) \rangle$ for $\hat{y}_{\bullet}(t)$ and $\langle [x(t'+t) - x(t')]^2 \rangle = \langle [x^{(k)}(t) - x^{(k)}(0)]^2 \rangle = \langle [x(t) - x(0)]^2 \rangle$ for $\hat{z}_{\bullet}(t)$.

2.3. Relations between the standard error of $D(t)$ and the error correlation functions of the VACF and the MSD

Here, we investigate how the standard errors of $\hat{z}_{\bullet}(t)$ and $\hat{a}_{\bullet}(t)$ are expressed in terms of the error correlation functions of $\hat{y}_{\bullet}(t)$ and $\hat{b}_{\bullet}(t)$, respectively. In other words, we derive expressions indicating how the standard errors of $D(t)$ obtained from the VACF and MSD methods can be estimated from the error correlation functions of the VACF and the MSD, respectively. In fact, in Section 2.5, by using these expressions, we shall obtain the standard errors of $D(t)$ from the error correlation functions.

Since the resulting relations do not depend on which type of averaging procedure is employed, we drop the subscripts on the estimators in this subsection. After presenting the relations between $\hat{y}(t)$ and $\hat{z}(t)$ and between $\hat{b}(t)$ and $\hat{a}(t)$, we also consider a linear regression estimator \hat{A} , which generalizes the estimator \hat{a} . Then, we discuss the resulting relations.

Relation for time integration. For estimators $\hat{y}(t)$ and $\hat{z}(t)$ satisfying

$$\hat{z}(t) = \int_0^t \hat{y}(t')dt', \tag{20}$$

their statistical errors, defined as $\varepsilon_{\hat{y}}(t) = \hat{y}(t) - \langle \hat{y}(t) \rangle$ and $\varepsilon_{\hat{z}}(t) = \hat{z}(t) - \langle \hat{z}(t) \rangle$, satisfy the following relation:

$$\langle \varepsilon_{\hat{z}}^2(t) \rangle = \int_0^t dt' \int_0^t dt'' \langle \varepsilon_{\hat{y}}(t')\varepsilon_{\hat{y}}(t'') \rangle. \tag{21}$$

Hence, the standard error of $\hat{z}(t)$, $\sigma[\hat{z}(t)] = \sqrt{\langle \varepsilon_{\hat{z}}^2(t) \rangle}$, can be calculated from the double time integral of the error correlation function of $\hat{y}(t)$.

The derivation of Eq. (21) is straightforward. By substituting Eq. (20) into $\langle \varepsilon_{\hat{z}}^2(t) \rangle = \langle \hat{z}^2(t) \rangle - \langle \hat{z}(t) \rangle^2$, we obtain $\langle \varepsilon_{\hat{z}}^2(t) \rangle = \int_0^t dt' \int_0^t dt'' \text{Cov}[\hat{y}(t'), \hat{y}(t'')]$. Then, we obtain the relation from $\langle \varepsilon_{\hat{y}}(t') \varepsilon_{\hat{y}}(t'') \rangle = \text{Cov}[y(t'), y(t'')]$.

Relation for average slope. For estimators $\hat{b}(t)$ and $\hat{a}(t)$ satisfying

$$\hat{a}(t_1, t_2) = \frac{\hat{b}(t_2) - \hat{b}(t_1)}{t_2 - t_1}, \tag{22}$$

their statistical errors, defined as $\varepsilon_{\hat{b}}(t) = \hat{b}(t) - \langle \hat{b}(t) \rangle$ and $\varepsilon_{\hat{a}}(t_1, t_2) = \hat{a}(t_1, t_2) - \langle \hat{a}(t_1, t_2) \rangle$, satisfy the following relation:

$$\langle \varepsilon_{\hat{a}}^2(t_1, t_2) \rangle = \frac{\langle [\varepsilon_{\hat{b}}(t_2) - \varepsilon_{\hat{b}}(t_1)]^2 \rangle}{(t_2 - t_1)^2}. \tag{23}$$

Eq. (23) is obtained as follows. By substituting Eq. (22) into $\langle \varepsilon_{\hat{a}}^2(t_1, t_2) \rangle = \langle \hat{a}^2(t_1, t_2) \rangle - \langle \hat{a}(t_1, t_2) \rangle^2$, we obtain

$$\langle \varepsilon_{\hat{a}}^2(t_1, t_2) \rangle = \frac{\text{Var}[\hat{b}(t_1)] - 2\text{Cov}[\hat{b}(t_1), \hat{b}(t_2)] + \text{Var}[\hat{b}(t_2)]}{(t_2 - t_1)^2}. \tag{24}$$

Then, from $\text{Var}[\hat{b}(t_i)] = \langle \varepsilon_{\hat{b}}^2(t_i) \rangle$ ($i = 1, 2$) and $\text{Cov}[\hat{b}(t_1), \hat{b}(t_2)] = \langle \varepsilon_{\hat{b}}(t_1) \varepsilon_{\hat{b}}(t_2) \rangle$, we obtain the relation.

Relation for linear regression. The above result for the average slope can be generalized to the simple linear regression of an m -point data set $\{t_i, \hat{b}(t_i)\}_{i=1}^m$. For the least squares estimator for the slope, which is written as

$$\hat{A}(t_1, t_2, \dots, t_m) = \frac{\sum_{i=1}^m \sum_{j=i+1}^m [\hat{b}(t_j) - \hat{b}(t_i)] (t_j - t_i)}{\sum_{i=1}^m \sum_{j=i+1}^m (t_j - t_i)^2}, \tag{25}$$

the variance of its statistical error $\langle \varepsilon_{\hat{A}}^2(t_1, t_2, \dots, t_m) \rangle = \text{Var}[\hat{A}(t_1, t_2, \dots, t_m)]$ is expressed as

$$\frac{\sum_{i=1}^m \sum_{j=i+1}^m \sum_{k=1}^m \sum_{l=k+1}^m \langle [\varepsilon_{\hat{b}}(t_j) - \varepsilon_{\hat{b}}(t_i)] [\varepsilon_{\hat{b}}(t_l) - \varepsilon_{\hat{b}}(t_k)] \rangle (t_j - t_i)(t_l - t_k)}{\left[\sum_{i=1}^m \sum_{j=i+1}^m (t_j - t_i)^2 \right]^2}. \tag{26}$$

We note that, as expected, Eq. (26) reduces to Eq. (23) for $m = 2$.

Eq. (26) is obtained as follows. By using $\text{Var}[\sum_i X_i] = \sum_{i,j} \text{Cov}[X_i, X_j]$, the variance of \hat{A} is expressed as a weighted sum of $\text{Cov}[\hat{b}(t_j) - \hat{b}(t_i), \hat{b}(t_l) - \hat{b}(t_k)]$. Then, after replacing the latter by $\langle [\varepsilon_{\hat{b}}(t_j) - \varepsilon_{\hat{b}}(t_i)] [\varepsilon_{\hat{b}}(t_l) - \varepsilon_{\hat{b}}(t_k)] \rangle$, Eq. (26) is obtained.

Discussion. Before closing this subsection, we discuss the resulting relations, Eqs. (21), (23), and (26), in a more general context. We note that these relations are valid for any time integration and differentiation procedure performed in MD simulation. Most importantly, these relations demonstrate that the level of statistical errors in the processed data is determined by the error correlation function of the original data, i.e., $\langle \varepsilon_{\hat{y}}(t') \varepsilon_{\hat{y}}(t'') \rangle$ or $\langle \varepsilon_{\hat{b}}(t') \varepsilon_{\hat{b}}(t'') \rangle$. Since the statistical errors in the original data at different times are correlated, the levels of their statistical uncertainties at each time, i.e., $\langle \varepsilon_{\hat{y}}^2(t) \rangle$ or $\langle \varepsilon_{\hat{b}}^2(t) \rangle$, provide insufficient information. As mentioned in the Introduction, this implies that even if the error bars are available in the VACF or MSD data obtained from MD simulations, we cannot, in principle, estimate the error bars in the estimated values of $D(t)$ unless we estimate the error correlation function from MD simulations, which requires huge computational overhead. In Section 2.4, we shall see that this issue can be circumvented by introducing the GPA.

In the definition of $\hat{D}^{\text{MSD}}(t)$, the instantaneous slope of \hat{b} at time t , i.e., $\lim_{\Delta t \rightarrow 0} \hat{a}(t, t + \Delta t)$ is used. The level of statistical error in the instantaneous slope is expressed in terms of the error correlation function as follows. By applying L'Hôpital's rule twice to Eq. (23), we obtain

$$\lim_{\Delta t \rightarrow 0} \langle \varepsilon_{\hat{a}}^2(t, t + \Delta t) \rangle = \lim_{\Delta t \rightarrow 0} \frac{d^2}{d\Delta t^2} \left[\frac{1}{2} \langle \varepsilon_{\hat{b}}^2(t + \Delta t) \rangle - \langle \varepsilon_{\hat{b}}(t) \varepsilon_{\hat{b}}(t + \Delta t) \rangle \right]. \tag{27}$$

We use this later to calculate the standard errors of $\hat{D}^{\text{MSD}}(t)$.

2.4. Assumptions

Before presenting our theoretical error estimates in Section 2.5, we summarize the assumptions to be introduced. In Section 2.4.1, we discuss the GPA of the underlying process, which is the most important assumption in our theoretical approach. In Section 2.4.2, we clarify other assumptions to be made for the derivation of the theoretical error estimates.

2.4.1. Gaussian process approximation (GPA)

We shall see in Section 2.5 that the error correlation functions are expressed in terms of fourth-order cross moments of x and v , e.g., $\text{Cov}[v(0)v(t'), v(0)v(t'')] = \langle v(0)v(t')v(0)v(t'') \rangle - \langle v(0)v(t') \rangle \langle v(0)v(t'') \rangle$, see Eqs. (55), (56), (57), and (58). Compared to the VACF and the MSD, the calculation of these moment functions from MD simulation is very expensive. Hence, in order to approximate the fourth-order cross moments by using the VACF and the MSD, we introduce the GPA of x and v . We first define a Gaussian process and mention some relevant properties. Then, we explain how the GPA is used to approximate fourth-order cross moments of x and v in our theoretical approach.

A process X_t is Gaussian if and only if, for any finite set of time points $\{t_1, \dots, t_k\}$, the joint distribution of $\{X_{t_1}, \dots, X_{t_k}\}$ follows a multivariate normal distribution [30]. One useful property of a Gaussian process, which is relevant to our theoretical approach, is that all higher-order cross moments are expressed as the sum of the products of the first two lowest-order moments (i.e., $\langle X_t \rangle$ and $\langle X_t X_s \rangle$), which is referred to as Isserlis' theorem or Wick's theorem. For example, the fourth-order cross moment of a zero-mean Gaussian process X_t is decomposed as

$$\langle X_t X_s X_u X_v \rangle = \langle X_t X_s \rangle \langle X_u X_v \rangle + \langle X_t X_u \rangle \langle X_s X_v \rangle + \langle X_t X_v \rangle \langle X_s X_u \rangle. \tag{28}$$

A Gaussian process can also be characterized by its cumulants [31]: all joint cumulants of order greater than two are zero. For example, since the fourth-order joint cumulant of a zero-mean process X_t is defined as

$$\kappa[X_t, X_s, X_u, X_v] \stackrel{\text{def}}{=} \langle X_t X_s X_u X_v \rangle - \langle X_t X_s \rangle \langle X_u X_v \rangle - \langle X_t X_u \rangle \langle X_s X_v \rangle - \langle X_t X_v \rangle \langle X_s X_u \rangle, \tag{29}$$

it becomes zero if the process is Gaussian.

Based on these properties of a Gaussian process, we have the following identity for the fourth-order cross moment of a zero-mean process X_t :

$$\langle X_t X_s X_u X_v \rangle = \langle X_t X_s X_u X_v \rangle_{\text{GPA}} + \kappa[X_t, X_s, X_u, X_v], \tag{30}$$

where

$$\langle X_t X_s X_u X_v \rangle_{\text{GPA}} = \langle X_t X_s \rangle \langle X_u X_v \rangle + \langle X_t X_u \rangle \langle X_s X_v \rangle + \langle X_t X_v \rangle \langle X_s X_u \rangle. \tag{31}$$

In the present paper, under the GPA of X_t , we assume that $\langle X_t X_s X_u X_v \rangle \approx \langle X_t X_s X_u X_v \rangle_{\text{GPA}}$ or $\kappa[X_t, X_s, X_u, X_v] \approx 0$. Hence, under the GPA of x and v , we shall use the following approximations:

$$\langle v(0)v(t_1)v(t_2)v(t_3) \rangle \approx \langle v(0)v(t_1) \rangle \langle v(0)v(t_3 - t_2) \rangle + \langle v(0)v(t_2) \rangle \langle v(0)v(t_3 - t_1) \rangle + \langle v(0)v(t_3) \rangle \langle v(0)v(t_2 - t_1) \rangle, \tag{32}$$

$$\langle [x(t_1) - x(0)]^2 [x(t_3) - x(t_2)]^2 \rangle \approx \langle [x(t_1) - x(0)]^2 \rangle \langle [x(t_3 - t_2) - x(0)]^2 \rangle + 2 \langle [x(t_1) - x(0)][x(t_3) - x(t_2)] \rangle^2. \tag{33}$$

In addition, we note that the GPA of v implies that of x since a linear transform of a Gaussian process is also Gaussian.

It is important to note that the GPA is not always exactly valid even if the marginal distribution of $v(t_i)$ is Gaussian at each time t_i . Hence, the GPA needs to be validated for each case. One plausible explanation for the GPA in particle diffusion phenomena is given by the generalized Langevin equation [32], which is a formally exact equation for the dynamics of a tracer particle:

$$m\dot{v}(t) = - \int_0^t K(t-t')v(t')dt' + f^+(t), \tag{34}$$

where $f^+(t)$ is the fluctuating force and $K(t)$ is the memory function. In the case of a large tracer particle suspended in a dense fluid, where the distribution of the fluctuating force becomes Gaussian [33], the linearity of Eq. (34) implies that $v(t)$ also becomes a Gaussian process. The limiting case of this argument is the Langevin equation, where the fluctuating force becomes GWN and the memory function becomes a delta function.

Although analyzing the force on the tracer particle seems useful for investigating the validity of the GPA, it is also important to note that our theoretical framework does not rely on any assumption on the form of the force. For the dynamics of the system, we only assume the relation between x and v , see Eq. (7), and the GPA of v . Hence, our theoretical approach is applicable not only to MD systems where the force is fully deterministic but also to mesoscopic simulation methods such as the Langevin dynamics [32] and the dissipative particle dynamics [34], where stochastic forces are introduced.

2.4.2. Other assumptions

Assumptions on averaging procedures. For time-averaging, we assume that the VACF eventually decays to zero and the length \mathcal{T} of the time-averaging interval is much longer than the characteristic decay time scale of the VACF. We need these assumptions to obtain the \mathcal{T}^{-1} -dependence of the error correlation functions. We use the following approximation, which becomes asymptotically valid for large \mathcal{T} :

$$\frac{1}{\mathcal{T}^2} \int_0^{\mathcal{T}} dt' \int_0^{\mathcal{T}} dt'' h(t' - t'') \approx \frac{1}{\mathcal{T}} \int_{-\infty}^{\infty} h(\alpha) d\alpha, \quad (35)$$

where $h(t)$ is an integrable function decaying sufficiently fast. This approximation has been heuristically introduced in Refs. [3,16,18]. In Appendix A, we provide the error term of the approximation and a condition for the decay rate of $h(t)$.

For ensemble-averaging, we assume that $\{x^{(k)}(t), v^{(k)}(t)\}_{k=1}^N$ are independent of one another and each of them has the same statistical properties as $\{x(0), v(0)\}$. In other words, they are *independent replicas*. Since we are interested only in average values, we actually need a weaker condition: uncorrelatedness rather than independence. When we take particle-average in an MD simulation of a system consisting of identical particles, we shall investigate the uncorrelatedness assumption.

Finally, we assume that Δt is sufficiently small that any systematic error due to discretization is negligible. For the errors due to insufficient equilibration and early truncation of the time integral of the VACF, we refer the reader to Ref. [19].

2.5. Main theoretical results

Here, we present main theoretical results of the present paper, which are summarized below as [Main Results 1, 2, and 3](#). The definitions of the estimators employed in the present paper are summarized in [Table 1](#), and the three types of averaging procedures considered in the paper are schematically depicted in [Fig. 2](#).

In Sections 2.5.1 and 2.5.2, we explain how these results are obtained. As mentioned in the preceding sections, after obtaining the error correlation functions of the VACF and the MSD under the GPA, we express the standard error of $D(t)$ by using the relations obtained in Section 2.3, and conclude that the standard errors of $\hat{D}_{\bullet}^{\text{VACF}}(t)$ and $\hat{D}_{\bullet}^{\text{MSD}}(t)$ are identical. In Section 2.5.3, we compare our theoretical results with previous results. Finally, we discuss non-Gaussianity effect in Section 2.5.4.

Main Result 1 (Error correlation function of VACF). For the VACF estimators $\hat{y}_{\text{time}}(t)$, $\hat{y}_{\text{ens}}(t)$, and $\hat{y}_{\text{hybrid}}(t)$ defined in Eqs. (10a), (12a), and (13a), denote the sampling errors as

$$\varepsilon_{\hat{y}_{\bullet}}(t) = \hat{y}_{\bullet}(t) - \langle v(0)v(t) \rangle, \quad (36)$$

where $\bullet = \text{time, ens, and hybrid}$. If the velocity process of a diffusing particle is Gaussian, the error correlation functions of the estimators are expressed as

$$\langle \varepsilon_{\hat{y}_{\text{time}}}(t') \varepsilon_{\hat{y}_{\text{time}}}(t'') \rangle = \frac{1}{\mathcal{T}} \int_{-\infty}^{\infty} d\alpha [\langle v(0)v(\alpha) \rangle \langle v(0)v(\alpha + t'' - t') \rangle + \langle v(0)v(\alpha - t') \rangle \langle v(0)v(\alpha + t'') \rangle], \quad (37)$$

$$\langle \varepsilon_{\hat{y}_{\text{ens}}}(t') \varepsilon_{\hat{y}_{\text{ens}}}(t'') \rangle = \frac{1}{\mathcal{N}} [\langle v^2 \rangle \langle v(0)v(t'' - t') \rangle + \langle v(0)v(t') \rangle \langle v(0)v(t'') \rangle], \quad (38)$$

$$\langle \varepsilon_{\hat{y}_{\text{hybrid}}}(t') \varepsilon_{\hat{y}_{\text{hybrid}}}(t'') \rangle = \frac{1}{\mathcal{N}} \langle \varepsilon_{\hat{y}_{\text{time}}}(t') \varepsilon_{\hat{y}_{\text{time}}}(t'') \rangle. \quad (39)$$

Main Result 2 (Error correlation function of MSD). For the MSD estimators $\hat{b}_{\text{time}}(t)$, $\hat{b}_{\text{ens}}(t)$, and $\hat{b}_{\text{hybrid}}(t)$ defined in Eqs. (10b), (12b), and (13b), denote the sampling errors as

$$\varepsilon_{\hat{b}_{\bullet}}(t) = \hat{b}_{\bullet}(t) - \langle [x(t) - x(0)]^2 \rangle, \quad (40)$$

where $\bullet = \text{time, ens, and hybrid}$. If the velocity process of a diffusing particle is Gaussian, the error correlation functions of the estimators are expressed as

$$\langle \varepsilon_{\hat{b}_{\text{time}}}(t') \varepsilon_{\hat{b}_{\text{time}}}(t'') \rangle = \frac{2}{\mathcal{T}} \int_{-\infty}^{\infty} d\alpha \langle [x(t') - x(0)][x(\alpha + t'') - x(\alpha)]^2 \rangle, \quad (41)$$

$$\langle \varepsilon_{\hat{b}_{\text{ens}}}(t') \varepsilon_{\hat{b}_{\text{ens}}}(t'') \rangle = \frac{2}{\mathcal{N}} \langle [x(t') - x(0)][x(t'') - x(0)]^2 \rangle, \quad (42)$$

$$\langle \varepsilon_{\hat{b}_{\text{hybrid}}}(t') \varepsilon_{\hat{b}_{\text{hybrid}}}(t'') \rangle = \frac{1}{\mathcal{N}} \langle \varepsilon_{\hat{b}_{\text{time}}}(t') \varepsilon_{\hat{b}_{\text{time}}}(t'') \rangle. \quad (43)$$

Main Result 3 (Standard error of diffusion coefficient). For the estimators of the time-dependent diffusion coefficient, $\hat{D}_{\bullet}^{\text{VACF}}(t)$ and $\hat{D}_{\bullet}^{\text{MSD}}(t)$ ($\bullet = \text{time, ens, hybrid}$), defined through Eqs. (14), (15), and (16), denote the standard errors of the estimators as

$$\sigma[\hat{D}_{\bullet}^{\star}(t)] = \sqrt{\langle [\hat{D}_{\bullet}^{\star}(t) - D(t)]^2 \rangle}, \tag{44}$$

where $\star = \text{VACF or MSD}$. If the velocity process of a diffusing particle is Gaussian, the standard errors are expressed as

$$\sigma[\hat{D}_{\text{time}}^{\text{VACF}}(t)] = \sigma[\hat{D}_{\text{time}}^{\text{MSD}}(t)] = \sqrt{\frac{c_1(t)}{\mathcal{T}}}, \tag{45}$$

$$\sigma[\hat{D}_{\text{ens}}^{\text{VACF}}(t)] = \sigma[\hat{D}_{\text{ens}}^{\text{MSD}}(t)] = \sqrt{\frac{c_2(t)}{\mathcal{N}}}, \tag{46}$$

$$\sigma[\hat{D}_{\text{hybrid}}^{\text{VACF}}(t)] = \sigma[\hat{D}_{\text{hybrid}}^{\text{MSD}}(t)] = \sqrt{\frac{c_1(t)}{\mathcal{N}\mathcal{T}}}, \tag{47}$$

where

$$c_1(t) = \int_{-\infty}^{\infty} d\alpha \left[f(\alpha) \int_0^t dt' \int_{\alpha}^{\alpha+t} dt'' f(t' - t'') + \int_{\alpha}^{\alpha+t} f(t') dt' \int_{\alpha-t}^{\alpha} f(t'') dt'' \right], \tag{48}$$

$$c_2(t) = \int_0^t dt' \int_0^t dt'' [f(0)f(t'' - t') + f(t')f(t'')], \tag{49}$$

and $f(t) = \langle v(0)v(t) \rangle$.

Remark 1. All expressions containing \mathcal{T} (i.e., in the time-averaging and hybrid ensemble/time-averaging cases) are asymptotically valid for large \mathcal{T} . In fact, they are derived through Eq. (A.1). See also Section 2.4.2.

Remark 2. If we set $\mathcal{N} = 1$, all results of the hybrid ensemble/time-averaging case are reduced to the time-averaging case.

Remark 3. The standard error $\sigma(t)$ of an estimator at each time can be calculated from its error correlation function $\langle \varepsilon(t')\varepsilon(t'') \rangle$ by setting $t' = t'' = t$. In other words, $\sigma(t) = \sqrt{\langle \varepsilon^2(t) \rangle}$. From Main Results 1 and 2, we obtain the following expressions for the variance of sampling error:

$$\langle \varepsilon_{\hat{y}_{\text{time}}}^2(t) \rangle = \frac{1}{\mathcal{T}} \int_{-\infty}^{\infty} d\alpha \left[\langle v(0)v(\alpha) \rangle^2 + \langle v(0)v(\alpha - t) \rangle \langle v(0)v(\alpha + t) \rangle \right], \tag{50}$$

$$\langle \varepsilon_{\hat{y}_{\text{ens}}}^2(t) \rangle = \frac{1}{\mathcal{N}} \left[\langle v^2 \rangle^2 + \langle v(0)v(t) \rangle^2 \right], \tag{51}$$

$$\langle \varepsilon_{\hat{b}_{\text{time}}}^2(t) \rangle = \frac{2}{\mathcal{T}} \int_{-\infty}^{\infty} d\alpha \langle [x(t) - x(0)][x(\alpha + t) - x(\alpha)] \rangle^2, \tag{52}$$

$$\langle \varepsilon_{\hat{b}_{\text{ens}}}^2(t) \rangle = \frac{2}{\mathcal{N}} \langle [x(t) - x(0)]^2 \rangle. \tag{53}$$

For $\langle \varepsilon_{\hat{y}_{\text{hybrid}}}^2(t) \rangle$ and $\langle \varepsilon_{\hat{b}_{\text{hybrid}}}^2(t) \rangle$, replace \mathcal{T} with $\mathcal{N}\mathcal{T}$ in Eq. (50) and (52).

2.5.1. Derivation: Main Results 1 and 2

We divide the derivation into two steps. First, we express the error correlation functions in terms of covariance. Then, we apply the GPA of x and v , see Eqs. (32) and (33), to the covariance expressions. For the time-averaging case, we also apply the asymptotic relation (A.1) for large \mathcal{T} . Since the scaling behavior with respect to \mathcal{N} in Eqs. (39) and (43) is trivial, we omit the hybrid ensemble/time-averaging case.

In the first step, we use the following basic properties of covariance:

$$\text{Cov}[X, Y] = \langle [X - \langle X \rangle][Y - \langle Y \rangle] \rangle, \tag{54a}$$

$$\text{Cov}\left[\int_0^t X(t')dt', \int_0^t Y(t'')dt''\right] = \int_0^t dt' \int_0^t dt'' \text{Cov}[X(t'), Y(t'')], \tag{54b}$$

$$\text{Cov}\left[\sum_i X_i, \sum_j Y_j\right] = \sum_{i,j} \text{Cov}[X_i, Y_j], \tag{54c}$$

$$\text{Cov}[X, Y] = 0, \quad \text{if } X \text{ and } Y \text{ are uncorrelated.} \tag{54d}$$

By using Eqs. (54a) and (54b), we obtain

$$\begin{aligned} \langle \varepsilon_{\hat{y}_{\text{time}}}(t_1) \varepsilon_{\hat{y}_{\text{time}}}(t_2) \rangle &= \text{Cov}[\hat{y}_{\text{time}}(t_1), \hat{y}_{\text{time}}(t_2)] \\ &= \frac{1}{\mathcal{T}^2} \int_0^{\mathcal{T}} dt' \int_0^{\mathcal{T}} dt'' \text{Cov}[v(t')v(t'+t_1), v(t'')v(t''+t_2)], \end{aligned} \tag{55}$$

$$\begin{aligned} \langle \varepsilon_{\hat{b}_{\text{time}}}(t_1) \varepsilon_{\hat{b}_{\text{time}}}(t_2) \rangle &= \text{Cov}[\hat{b}_{\text{time}}(t_1), \hat{b}_{\text{time}}(t_2)] \\ &= \frac{1}{\mathcal{T}^2} \int_0^{\mathcal{T}} dt' \int_0^{\mathcal{T}} dt'' \text{Cov}[(x(t'+t_1) - x(t'))^2, (x(t''+t_2) - x(t''))^2]. \end{aligned} \tag{56}$$

Similarly, by using Eqs. (54a), (54c), and (54d), we obtain

$$\langle \varepsilon_{\hat{y}_{\text{ens}}}(t') \varepsilon_{\hat{y}_{\text{ens}}}(t'') \rangle = \frac{1}{\mathcal{N}} \text{Cov}[v(0)v(t'), v(0)v(t'')], \tag{57}$$

$$\langle \varepsilon_{\hat{b}_{\text{ens}}}(t') \varepsilon_{\hat{b}_{\text{ens}}}(t'') \rangle = \frac{1}{\mathcal{N}} \text{Cov}[(x(t') - x(0))^2, (x(t'') - x(0))^2]. \tag{58}$$

In the second step, we replace the covariance terms in Eqs. (55), (56), (57), and (58) with terms obtained from the GPA of x and v by using Eqs. (32) and (33). Then, we obtain Eqs. (38) and (42). As mentioned above, by further applying Eq. (A.1), we obtain Eqs. (37) and (41).

2.5.2. Derivation: Main Result 3

We first derive the expressions for the VACF method, which are more easily obtained through Eq. (21). Since

$$\sigma[\hat{D}_{\bullet}^{\text{VACF}}(t)] = \sqrt{\langle \varepsilon_{\hat{z}_{\bullet}}^2(t) \rangle}, \tag{59}$$

by substituting Eqs. (37), (38), and (39) into Eq. (21), we obtain Eqs. (45), (46), (47), respectively, with Eqs. (48) and (49).

For the expressions for the MSD method, we use Eq. (27) and

$$\sigma[\hat{D}_{\bullet}^{\text{MSD}}(t)] = \frac{1}{2} \sqrt{\lim_{\Delta t \rightarrow 0} \langle \varepsilon_{\hat{a}_{\bullet}}^2(t, t + \Delta t) \rangle}. \tag{60}$$

In the ensemble-averaging case, by substituting Eq. (42) to Eq. (27) and simplifying the result using Eqs. (7) and (8), we obtain

$$\lim_{\Delta t \rightarrow 0} \langle \varepsilon_{\hat{a}_{\text{ens}}}^2(t, t + \Delta t) \rangle = \frac{4}{\mathcal{N}} \left[\langle v^2 \rangle \langle [x(t) - x(0)]^2 \rangle + \langle [x(t) - x(0)]v(0) \rangle^2 \right]. \tag{61}$$

Then, by expressing $\langle [x(t) - x(0)]^2 \rangle$ and $\langle v(0)[x(t) - x(0)] \rangle$ in terms of the VACF (i.e., $\int_0^t dt' \int_0^t dt'' \langle v(t')v(t'') \rangle$ and $\int_0^t \langle v(0)v(t') \rangle dt'$, respectively), we obtain Eq. (46) with Eq. (49). Following the same procedure, for the time-averaging case, we obtain Eq. (45) with Eq. (48). Then, the result for the hybrid ensemble/time-averaging case also follows.

2.5.3. Comparison with previous results

We compare our results for the time-averaging case with previous results by Zwanzig and Ailawadi [16], by Frenkel [18], and by Jones and Mandadapu [19]. Only Zwanzig and Ailawadi have considered the error correlation function of \hat{y}_{time} rather than directly calculating the mean-squared error $\langle \varepsilon_{\hat{y}_{\text{time}}}^2(t) \rangle$. However, they have estimated the error correlation function very roughly as a constant: $\langle \varepsilon_{\hat{y}_{\text{time}}}(t') \varepsilon_{\hat{y}_{\text{time}}}(t'') \rangle \approx \frac{2}{\mathcal{T}} \int_{-\infty}^{\infty} \langle v(0)v(t) \rangle^2 dt$. Frenkel and also Jones and Mandadapu have shown that

$$\langle \varepsilon_{\hat{y}_{\text{time}}}^2(t) \rangle = \frac{a(t)}{\mathcal{T}} \int_{-\infty}^{\infty} \langle v(0)v(\alpha) \rangle^2 d\alpha, \tag{62}$$

where $a(0) = 2$ and $\lim_{t \rightarrow \infty} a(t) = 1$. By using Eq. (50), we can easily show Eq. (62).

We can also re-derive the following result obtained by Jones and Mandadapu:

$$\lim_{t \rightarrow 0} \frac{d}{dt} \langle \varepsilon_{\hat{z}_{\text{time}}}^2(t) \rangle = 0, \quad \lim_{t \rightarrow \infty} \frac{d}{dt} \langle \varepsilon_{\hat{z}_{\text{time}}}^2(t) \rangle = \frac{4D^2}{\mathcal{T}}. \tag{63}$$

By differentiating $\langle \varepsilon_{\hat{z}_{\text{time}}}^2(t) \rangle = c_1(t)/\mathcal{T}$ with respect to t , we have

$$\begin{aligned} \frac{d}{dt} \langle \varepsilon_{\hat{z}_{\text{time}}}^2(t) \rangle &= \frac{2}{\mathcal{T}} \int_{-\infty}^{\infty} d\alpha \int_0^t dt' [f(\alpha)f(\alpha+t-t') + f(\alpha)f(\alpha+t+t')] \\ &= \frac{2}{\mathcal{T}} \int_{-\infty}^{\infty} d\alpha \int_0^{2t} dt' f(\alpha)f(\alpha+t'), \end{aligned} \tag{64}$$

where $f(t) = \langle v(0)v(t) \rangle$. Hence, the first result in Eq. (63) is easily obtained and the second one is obtained from

$$\lim_{t \rightarrow \infty} \frac{d}{dt} \langle \varepsilon_{\hat{z}_{\text{time}}}^2(t) \rangle = \frac{4}{\mathcal{T}} \left(\int_0^\infty f(t') dt' \right)^2. \tag{65}$$

In addition, from Eq. (64), we can also obtain the second derivative of the mean-squared error of \hat{z}_{time} as follows:

$$\frac{d^2}{dt^2} \langle \varepsilon_{\hat{z}_{\text{time}}}^2(t) \rangle = \frac{4}{\mathcal{T}} \int_{-\infty}^\infty f(\alpha) f(\alpha + 2t) d\alpha. \tag{66}$$

2.5.4. Effect of non-Gaussianity

Here, we discuss the deviation of our theoretical error estimates from the true error estimates. Based on Eq. (30), when the covariance terms in Eqs. (55), (56), (57), (56) are approximated by the GPA, the resulted errors are expressed as the fourth-order cumulants of x and v . Since the error correlation functions are obtained from some linear transforms of the covariance terms, the deviation $\langle \varepsilon(t_1)\varepsilon(t_2) \rangle_{\text{true}} - \langle \varepsilon(t_1)\varepsilon(t_2) \rangle_{\text{GPA}}$ is linearly proportional to the corresponding cumulant function. Following a similar argument, we can show that the deviation in the standard error estimates of $D(t)$, which is characterized by $\sigma^2[\hat{D}(t)]_{\text{true}} - \sigma^2[\hat{D}(t)]_{\text{GPA}}$, is linearly proportional to the cumulant function.

3. Verification: Langevin equation

In this section, we investigate the statistical errors occurring in the simulation of the following Langevin equation:

$$\begin{aligned} \dot{x}(t) &= v(t), \\ m\dot{v}(t) &= -\gamma v(t) + \xi(t), \end{aligned} \tag{67}$$

where m and γ are the mass and friction coefficient of the particle, respectively. We consider two types of noise process $\xi(t)$: Gaussian white noise (GWN) $\xi_G(t)$ and Poissonian white shot noise (PWSN) $\xi_P(t)$. For the definition of PWSN, see Section 3.2. Both noise processes are delta-correlated and we impose the same noise intensity

$$\langle \xi(t')\xi(t'') \rangle = 2k_B T \gamma \delta(t' - t''), \tag{68}$$

where T is the temperature of the system.

Then, the GWN and PWSN cases have the same VACF, MSD, $D(t)$, which are expressed as

$$\langle v(0)v(t) \rangle = \frac{k_B T}{m} e^{-\frac{\gamma}{m}|t|}, \tag{69}$$

$$\langle [x(t) - x(0)]^2 \rangle = \frac{2k_B T}{\gamma} |t| - \frac{2mk_B T}{\gamma^2} \left(1 - e^{-\frac{\gamma}{m}|t|} \right), \tag{70}$$

$$D(t) = \frac{k_B T}{\gamma} \left(1 - e^{-\frac{\gamma}{m}t} \right), \quad \text{for } t \geq 0, \tag{71}$$

respectively. This is because the two-time noise correlation function $\langle \xi(0)\xi(t) \rangle$ completely determines the VACF, and thus the MSD and $D(t)$ as well. More specifically, since $v(t)$ is a stationary process expressed as

$$v(t) = \frac{1}{m} \int_{-\infty}^t e^{-\frac{\gamma}{m}(t-s)} \xi(s) ds, \tag{72}$$

the VACF is determined from Eq. (68) and the MSD and $D(t)$ are determined from the VACF by using Eq. (6).

When we investigate the sampling uncertainty on these quantities, however, the two cases exhibit different behaviors. Since the GPA is satisfied exactly in the GWN case, it is expected that our theoretical error estimates obtained in Section 2 hold perfectly. On the other hand, in the PWSN case, some deviations are expected. In Section 3.1, we verify our theoretical error estimates by comparing simulation results. In Section 3.2, we observe non-Gaussianity effect, which is discussed in Section 2.5.4.

3.1. Langevin equation driven by Gaussian white noise

In Section 3.1.1, we summarize closed-form expressions for the theoretical error estimates, which are obtained by substituting Eqs. (69) and (70). In Section 3.1.2, we present simulation results and compare numerical error estimates with the theoretical ones. Since the GPA is exactly valid, they are expected to match, from which we verify our theoretical results.

3.1.1. Closed-form expressions for the error estimates

Here, we assume $0 \leq t_1 \leq t_2$. Since the explicit expression of the error correlation function of \hat{b}_{time} is complicated, we only present its mean-squared error. For the hybrid ensemble/time-averaging case, replace \mathcal{T} with $\mathcal{N}\mathcal{T}$ in the time-averaging case.

Ensemble-averaging case.

$$\langle \varepsilon_{\hat{y}_{\text{ens}}}(t_1) \varepsilon_{\hat{y}_{\text{ens}}}(t_2) \rangle = \frac{1}{\mathcal{N}} \left(\frac{k_B T}{m} \right)^2 \left(e^{-\frac{\gamma}{m}(t_2-t_1)} + e^{-\frac{\gamma}{m}(t_1+t_2)} \right) \tag{73}$$

$$\langle \varepsilon_{\hat{b}_{\text{ens}}}(t_1) \varepsilon_{\hat{b}_{\text{ens}}}(t_2) \rangle = \frac{2}{\mathcal{N}} \left(\frac{k_B T}{\gamma} \right)^2 \left[2t_1 - \frac{m}{\gamma} \left(1 - e^{-\frac{\gamma t_1}{m}} - e^{-\frac{\gamma t_2}{m}} + e^{-\frac{\gamma(t_2-t_1)}{m}} \right) \right]^2 \tag{74}$$

$$\sigma^2[\hat{D}_{\text{ens}}(t)] = \frac{1}{\mathcal{N}} \left(\frac{k_B T}{\gamma} \right)^2 \left[\frac{2\gamma t}{m} - 1 + e^{-\frac{2\gamma t}{m}} \right] \tag{75}$$

Time-averaging case.

$$\langle \varepsilon_{\hat{y}_{\text{time}}}(t_1) \varepsilon_{\hat{y}_{\text{time}}}(t_2) \rangle = \frac{1}{\mathcal{T}} \left(\frac{k_B T}{m} \right)^2 e^{-\frac{\gamma(t_1+t_2)}{m}} \left[\frac{m}{\gamma} (1 + e^{\frac{2\gamma t_1}{m}}) + t_1 (1 - e^{\frac{2\gamma t_1}{m}}) + t_2 (1 + e^{\frac{2\gamma t_1}{m}}) \right] \tag{76}$$

$$\langle \varepsilon_{\hat{b}_{\text{time}}}^2(t) \rangle = \frac{1}{\mathcal{T}} \cdot \frac{4(k_B T)^2}{3\gamma^5} \left[15m^3 (3 - 4e^{-\frac{\gamma t}{m}} + e^{-\frac{2\gamma t}{m}}) - 6m^2 \gamma t (4 + 2e^{-\frac{\gamma t}{m}} - e^{-\frac{2\gamma t}{m}}) + 4\gamma^3 t^3 \right] \tag{77}$$

$$\sigma^2[\hat{D}_{\text{time}}(t)] = \frac{1}{\mathcal{T}} \left(\frac{k_B T}{\gamma} \right)^2 \left[(4 + 2e^{-\frac{2\gamma t}{m}})t - \frac{3m}{\gamma} (1 - e^{-\frac{2\gamma t}{m}}) \right] \tag{78}$$

3.1.2. Simulation results and discussion

By numerically generating sample trajectories of the Langevin equation (67), we calculate the VACF, the MSD, and time-dependent diffusion coefficient $D(t)$ through ensemble-averaging and time-averaging, and estimate the standard errors in the VACF, the MSD, and $D(t)$. The system parameters are chosen as $m = \gamma = k_B T = 1$. Sample trajectories are generated by the Euler scheme [35] with $\Delta t_{\text{LE}} = 0.001$. The initial condition of v is sampled from the equilibrium distribution of v ,

$$\rho_{\text{eq}}(v) = \sqrt{\frac{m}{2\pi k_B T}} \exp\left(-\frac{mv^2}{2k_B T}\right), \tag{79}$$

by a standard Gaussian random number generator. The VACF, MSD, and $D(t)$ are calculated up to time $T_{\text{max}} = 10$ with $\Delta t = 10\Delta t_{\text{LE}} = 0.01$. The finite-sampling size parameters are chosen as $\mathcal{N} = 1000$ and $\mathcal{T} = 1000$ for ensemble-averaging and time-averaging, respectively. In order to estimate the standard errors of the estimators \hat{y}_\bullet , \hat{D}^{VACF} , \hat{b}_\bullet , and \hat{D}^{MSD} , a total of $N_r = 32768$ realizations for each estimator are obtained. In other words, a total of $\mathcal{N}N_r$ trajectories of length T_{max} are calculated for ensemble-averaging, whereas a total of N_r trajectories of length \mathcal{T} are calculated for time-averaging. For each quantity, the mean-squared error $\langle \varepsilon^2(t) \rangle$ is estimated from N_r realizations and the *normalized* standard error $\bar{\sigma}$ is calculated as follows:

$$\bar{\sigma}_{\text{ens}} = \sqrt{\mathcal{N} \langle \varepsilon_{\text{ens}}^2(t) \rangle}, \quad \bar{\sigma}_{\text{time}} = \sqrt{\mathcal{T} \langle \varepsilon_{\text{time}}^2(t) \rangle}. \tag{80}$$

Then, the normalized standard error is expected to have little dependence on the sampling size parameters \mathcal{N} and \mathcal{T} and the mean value obtained from averaging over N_r realizations has standard error $\bar{\sigma}_{\text{ens}}/\sqrt{N_r \mathcal{N}}$ in the ensemble-averaging case and $\bar{\sigma}_{\text{time}}/\sqrt{N_r \mathcal{T}}$ in the time-averaging case. Simulation results are compared with theoretical results in Fig. 3. The agreement is excellent, so we verify our results in this case.

We make several observations on the time profile of the standard errors of the estimators. As expected from Eq. (62), the initial value of the standard error of \hat{y}_{time} eventually decreases by factor of $\frac{1}{\sqrt{2}}$. Although the same ratio of the long-time value to the initial value is observed for \hat{y}_{ens} , the time profile at short times is different from the former case and the transition is faster than the former case. This is because the time average uses correlated sample points and the statistical errors at different times are more correlated. By comparing Eqs. (37) and (38), we see that the same expression is obtained as the coefficient of \mathcal{N}^{-1} dependence in Eq. (38), if $\alpha = 0$ is substituted into the integrand in Eq. (37). This implies that the integration with respect to α in the time-averaging case *blurs* the temporal dependence in the error correlation function. In the MSD calculation, the standard error asymptotically increases like t in the ensemble-averaging case and $t^{3/2}$ in the time-averaging case. For the evaluation of $D(t)$, even after $D(t)$ attains a plateau, the standard deviations of the estimators increase asymptotically as \sqrt{t} in both averaging cases.

Based on these observations, we discuss the long-time behavior of the relative errors. Since the VACF decays rapidly as $t \rightarrow \infty$ but its standard error decreases only by factor $\frac{1}{\sqrt{2}}$, the relative error increases rapidly. Hence, in order to guarantee statistical reliability of the long-time tail of the VACF, one needs to severely increase the values of sampling size parameters such as \mathcal{N} and \mathcal{T} . For the MSD, the long-time behaviors of the relative errors are very different depending on the type of averaging procedure. Since the MSD asymptotically becomes proportional to t , the relative error becomes constant in the ensemble-averaging case, but it asymptotically increases like \sqrt{t} in the time-averaging case. Since the standard errors of $D(t)$ asymptotically increase like \sqrt{t} in both cases, the relative errors also increase like \sqrt{t} .

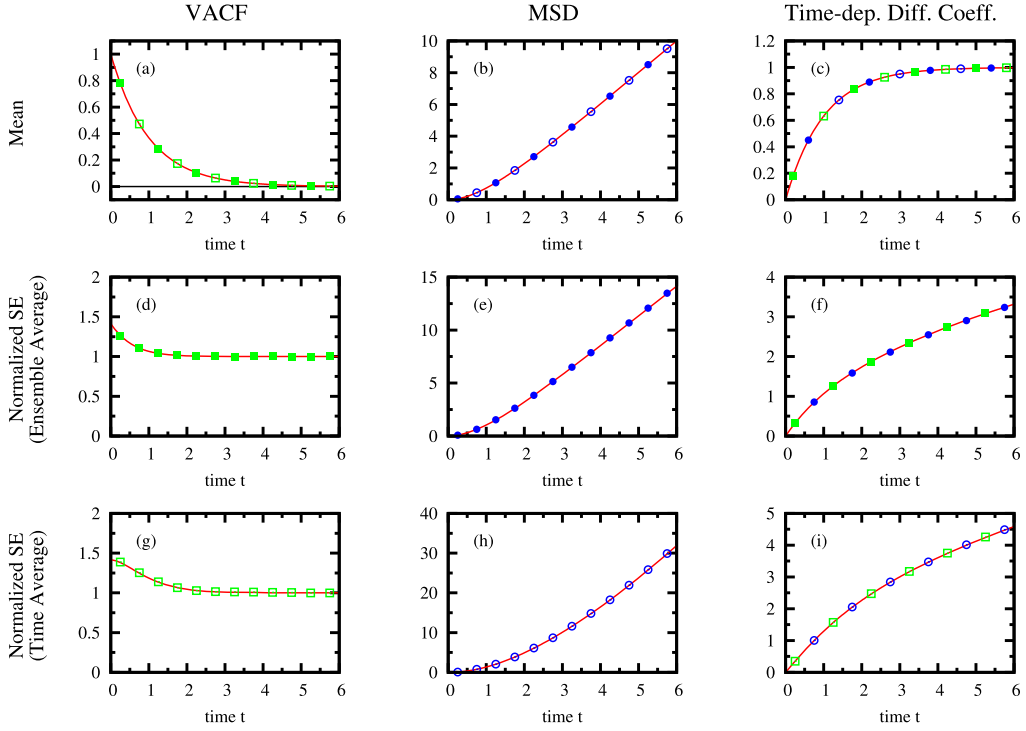


Fig. 3. Numerical simulation results of the Langevin equation (67) with $m = \gamma = k_B T = 1$. The VACF, MSD, and time-dependent coefficient $D(t)$ and their standard errors are calculated through either ensemble-averaging with $\mathcal{N} = 1000$ or time-averaging with $\mathcal{T} = 1000$ by generating a total of $N_r = 32768$ realizations. The time profiles of the three quantities are shown in the plots in the top row. The normalized standard errors (SE) are presented in the middle row for ensemble-averaging and in the bottom row for time-averaging. The standard errors are normalized by $\sqrt{\mathcal{N}}$ for the ensemble-averaging case, whereas they are normalized by $\sqrt{\mathcal{T}}$ for the time-averaging case, see Eq. (80). Hence, the standard errors of the mean values in the top row are estimated from the normalized standard errors $\bar{\sigma} \approx \bar{\sigma} / \sqrt{N_r \mathcal{N}}$ in the ensemble-averaging case and $\bar{\sigma} / \sqrt{N_r \mathcal{T}}$. The red lines depict analytic results, whereas four different symbols depict simulation results: filled green squares (VACF method, ensemble-averaging), empty green squares (VACF method, time-averaging), filled blue circles (MSD method, ensemble-averaging), and empty blue circles (MSD method, time-averaging). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

Before closing this section, we consider a time scale τ^* such that a sample point after time τ^* can be regarded as an uncorrelated sample point. Then, we may consider a time average over a trajectory of length \mathcal{T} as an ensemble average over a total of $\mathcal{N}^* = \mathcal{T} / \tau^*$ uncorrelated samples. By denoting the coefficients of \mathcal{N}^{-1} - and \mathcal{T}^{-1} -dependence for the ensemble- and time-averaging, respectively, as $I_{\text{ens}}(t)$ and $I_{\text{time}}(t)$ (see, e.g., Eqs. (50) and (51)), we equate $\frac{1}{\mathcal{T}} I_{\text{time}}(t)$ with $\frac{1}{\mathcal{N}^*} I_{\text{ens}}(t)$, which yields

$$\tau^* = \frac{I_{\text{time}}(t)}{I_{\text{ens}}(t)}. \tag{81}$$

For the VACF, the MSD, and $D(t)$, the time profiles of τ^* calculated from Eq. (81) are shown in Fig. 4. For the VACF and $D(t)$, τ^* maintains consistent values over the time interval, meaning that τ^* is well-defined for these quantities. On the other hand, τ^* increases linearly for the MSD, so τ^* is not well-defined in this case. From the closed-form expressions for the error estimates, we obtain the long-time limits of the ratio $I_{\text{time}}(t) / I_{\text{ens}}(t)$ for the three quantities:

$$\frac{I_{\text{time}}^{\text{VACF}}(t)}{I_{\text{ens}}^{\text{VACF}}(t)} \rightarrow \frac{m}{\gamma}, \quad \frac{I_{\text{time}}^{\text{MSD}}(t)}{I_{\text{ens}}^{\text{MSD}}(t)} \rightarrow \infty, \quad \frac{I_{\text{time}}^D(t)}{I_{\text{ens}}^D(t)} \rightarrow \frac{2m}{\gamma}. \tag{82}$$

We note that the time scale for the exponential decay of the VACF (see Eq. (69)) is m / γ . Hence, Eq. (82) shows that the time scale τ^* for uncorrelated sample points depends on the quantity to be averaged and may not always be well-defined.

3.2. Langevin equation driven by Poissonian white shot noise

We consider the following PWSN [20,30] $\xi_P(t)$:

$$\xi_P(t) = \sum_{i=1}^{N(t)} z_i \delta(t - \tau_i), \tag{83}$$

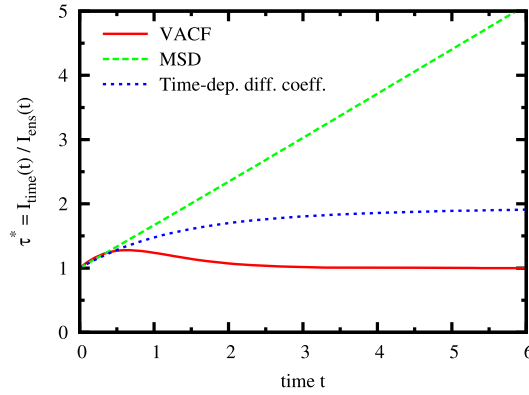


Fig. 4. For the VACF, MSD, and time-dependent diffusion coefficient, the ratio of $I_{\text{time}}(t)$ to $I_{\text{ens}}(t)$ is plotted. For the definition of $I_{\text{time}}(t)$ and $I_{\text{ens}}(t)$, see the text. The ratio can be considered as a time scale τ^* such that two points in a sample trajectory are effectively uncorrelated if their time difference is greater than τ^* .

where $N(t)$ is a homogeneous Poisson process with rate λ (i.e., $N(t)$ obeys a Poisson distribution with mean λt), and τ_i and z_i denote the arrival time and amplitude of the i th delta peak, respectively. We also assume that $\{z_i\}$ are independent and identically distributed with distribution $\rho_z(z)$. Note that PWSN is the (formal) time derivative of a compound Poisson process $\Gamma_P(t) = \sum_{i=1}^{N(t)} z_i$.

By increasing the rate λ and decreasing the magnitude of the delta peaks accordingly, PWSN converges to GWN [20,36]. More specifically, if $\rho_z(z)$ depends on λ and satisfies

$$\langle z \rangle = 0, \quad \lambda \langle z^2 \rangle = 2k_B T \gamma, \quad \lim_{\lambda \rightarrow \infty} \lambda \langle z^n \rangle = 0 \quad \text{for } n \geq 3, \quad (84)$$

ξ_P converges to ξ_G as $\lambda \rightarrow \infty$. This can be shown by calculating the joint cumulants of $\xi_P(t)$ as follows [20]. For all $\lambda > 0$, the first two cumulants are $\langle \xi_P(t) \rangle = 0$ and $\langle \xi_P(t_1) \xi_P(t_2) \rangle = 2k_B T \gamma \delta(t_1 - t_2)$. On the other hand, as $\lambda \rightarrow \infty$, all higher cumulants become zero. For example, the fourth-order joint cumulant, which is expressed as

$$\kappa[\xi_P(t_1), \xi_P(t_2), \xi_P(t_3), \xi_P(t_4)] = \lambda \langle z^4 \rangle \delta(t_1 - t_2) \delta(t_2 - t_3) \delta(t_3 - t_4), \quad (85)$$

shrinks to zero in the limit.

In this subsection, we assume that $\rho_z(z)$ is the uniform distribution on $[-a, a]$. Then, from $\lambda \langle z^2 \rangle = 2k_B T \gamma$, we have $a = \sqrt{\frac{6k_B T \gamma}{\lambda}}$. From Eq. (85), the fourth-order cumulant of $\xi_P(t)$ is proportional to λ^{-1} . By substituting Eq. (72) into Eq. (30) (with $X_t = v(t)$), we obtain the following expression:

$$\kappa[v(t_1), v(t_2), v(t_3), v(t_4)] = \frac{1}{m^4} \left(\prod_{i=1}^4 \int_{-\infty}^{t_i} ds_i e^{-\frac{\gamma}{m}(t_i - s_i)} \right) \kappa[\xi_P(s_1), \xi_P(s_2), \xi_P(s_3), \xi_P(s_4)]. \quad (86)$$

Hence, the fourth-order joint cumulant of $v(t)$ is also proportional to λ^{-1} . Based on the discussion in Section 2.5.4, we can use λ^{-1} as a non-Gaussianity parameter and expect that the deviation between the true error estimate and the GPA-based error estimate satisfies

$$\langle \mathcal{E}^2 \rangle_{\text{true}} - \langle \mathcal{E}^2 \rangle_{\text{GPA}} \propto \frac{1}{\lambda}. \quad (87)$$

To confirm this prediction, we perform numerical simulation of Eq. (67) with Eq. (83). A sample trajectory is calculated as follows [21]. First, we generate the arrival times $\{t_i\}$ and between two time points solve the deterministic equation analytically. The sampling procedure is the same as the GWN case, which is described in Section 3.1.2. The only exception is that after sampling the initial velocity from the Maxwell-Boltzmann distribution (i.e., Eq. (79)), the initial trajectory up to $t = 10$ is discarded. This is because the equilibrium distribution is no more Gaussian in the PWSN case and hence the trajectory has to be equilibrated.

In fact, the equilibrium distribution of v demonstrates non-Gaussianity of $v(t)$. As shown in Fig. 5, in the limit $\lambda \rightarrow \infty$, the non-Gaussian distribution converges to the Gaussian distribution corresponding to the GWN case. This behavior can also be confirmed by observing $\langle v^4 \rangle$. From Eq. (86), we obtain

$$\langle v^4 \rangle = \langle v^4 \rangle_{\text{GPA}} + \kappa[v, v, v, v] = \frac{3(k_B T)^2}{m^2} + \frac{9\gamma(k_B T)^2}{5m^3} \cdot \frac{1}{\lambda}. \quad (88)$$

As expected, the non-Gaussianity of $\langle v^4 \rangle$ is proportional to λ^{-1} .

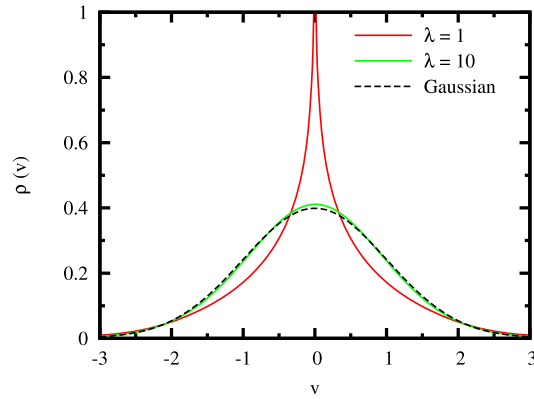


Fig. 5. The equilibrium velocity distribution $\rho(v)$ of the PWSN-driven Langevin model (see Eqs. (67) and (83)) is plotted for $\lambda = 1$ and 10. For comparison, the GWN case with the same noise intensity is plotted by the dashed line. For the larger value of λ , the distribution becomes similar to the GWN case.

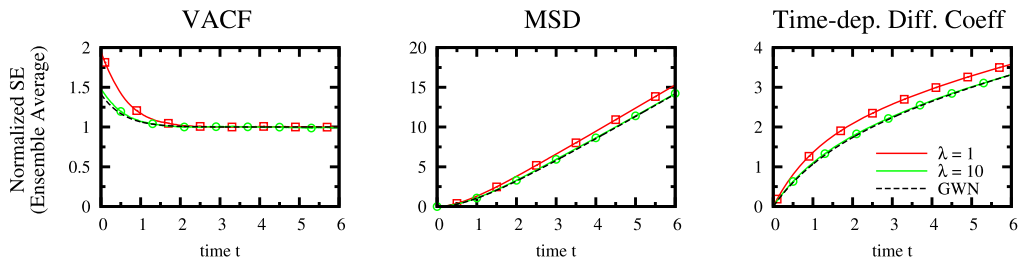


Fig. 6. For the PWSN-driven Langevin model, the time profiles of the normalized standard errors (SE) for ensemble-averaging are plotted for $\lambda = 1$ and 10. For comparison, the GWN case with the same noise intensity is plotted by the dashed lines. For the larger value of λ , the time profiles become similar to the GWN case. For the interpretation of the normalized standard errors, see the caption of Fig. 3.

We finally confirm our theoretical prediction on the non-Gaussianity effect, which is given in Eq. (87). As shown in Fig. 6, the standard errors of the VACF, the MSD, and $D(t)$ converge to the GWN case in the limit $\lambda \rightarrow \infty$. We observe that Eq. (87) is perfectly satisfied.

4. Validation: MD simulation

In this section, we investigate the error statistics involved in the estimation of the diffusion coefficient for two kinds of MD systems: a pure LJ fluid and a colloidal particle suspended in a LJ fluid. We observe the self-diffusion of a fluid particle in the former case and the tracer diffusion of a colloidal particle in the latter case. We call the two systems as the pure LJ fluid system vs. the colloidal system or the self-diffusion case vs. the tracer-diffusion case. Based on the discussion given in Section 2.4.1, see Eq. (34), the GPA is expected to be better satisfied in the tracer diffusion of a larger and heavier colloidal particle. We investigate the validity of the GPA in the MD systems by calculating the covariance function $\text{Cov}[v(0)v(t_1), v(0)v(t_2)]$ and comparing both sides of the following relation, which holds under the GPA:

$$\text{Cov}[v(0)v(t_1), v(0)v(t_2)] \stackrel{\text{GPA}}{=} \langle v^2 \rangle \langle v(0)v(t_1 - t_2) \rangle + \langle v(0)v(t_1) \rangle \langle v(0)v(t_2) \rangle. \tag{89}$$

Then, we assess the agreement of our theoretical predictions with the actual standard errors estimated from MD simulations and relate the level of agreement with the level of validity of the GPA. In addition, for the pure LJ fluid system, we consider averaging over identical particles and investigate the scaling behavior of the standard errors on the number of averaged particles. For accurate estimation, all calculations are carried out with a large-sized MD ensemble of 32 768 independent sample trajectories.

4.1. Simulation details

We report all simulation quantities in the LJ reduced units. In other words, we choose the mass of a fluid particle m_{fl} , the diameter of a fluid particle σ_{fl} , and the energy parameter of the LJ potential ϵ_{fl} as the units of mass, length, and energy, respectively.

System parameters. The pure LJ fluid system is defined as follows. The number density and temperature of the system are chosen as $\rho_{fl} = 0.8442$ and $T = 0.722$, which is close to the triple point of the fluid. This state has been frequently used

as a reference state after investigation in Ref. [37]. The number of fluid particles is set $N_{\text{fl}} = 2048$, so the side of the cubic simulation box L_{box} is 13.44. Periodic boundary conditions are imposed for all directions. A sufficiently long cutoff radius $r_{\text{cut}} = 5$ is used.

For the colloidal system, a colloidal particle of mass $M_{\text{tracer}} = 60$ is placed in the LJ fluid system described above. The interaction potential between a colloidal particle and a fluid particle is given as the Week–Chandler–Andersen potential with $\sigma_{\text{fl-tracer}} = 2.5$ and $\epsilon_{\text{fl-tracer}} = 1$. In other words, the interaction potential is a shifted LJ potential with cutoff $2^{1/6}\sigma_{\text{fl-tracer}}$, which contains only a repulsive contribution. From the values of σ_{fl} and $\sigma_{\text{fl-tracer}}$, we can infer that the radius of the colloidal particle is $R_{\text{tracer}} = 2$ and the density ratio is $\rho_{\text{tracer}}/\rho_{\text{fl}} = 2.1$.

MD simulation procedures. The same simulation procedures are performed for the pure LJ fluid system and the colloidal system. A total of $N_r = 32768$ sample trajectories are calculated by the velocity Verlet algorithm with time step size $\Delta t = 0.0025$. The initial configurations for the samples in equilibrium are prepared as follows. After the position and velocity of each particle are randomly sampled from the uniform distribution and the Maxwell–Boltzmann distribution, respectively, a total of 20 velocity scaling procedures are performed after every run of $10^4 \Delta t = 25$. So, the whole equilibration procedure is performed through the period of $2 \times 10^5 \Delta t = 500$.

The evaluation procedures of the VACF and the MSD are performed as follows. They are calculated up to time 10 through the ensemble-averaging and time-averaging procedures. In the ensemble-averaging case, the normalized standard error (see Eq. (80)) is calculated for $\mathcal{N} = 1$ with $N_r = 32768$ realizations. This is because the generation of equilibrium samples through long equilibration is expensive. In the time-averaging case, the normalized standard error is calculated for $\mathcal{T} = 500$ with the same number of realizations. We note that all simulation parameters are chosen so that all assumptions mentioned in Section 2.4.2 are satisfied. Hence, the deviation in our theoretical error estimates, which will be observed in this section, is due to the GPA.

For the calculation of the covariance function $\text{Cov}[v(0)v(t_1), v(0)v(t_2)]$, the correlation function $\langle v(0)v(t_1)v(0)v(t_2) \rangle$ ($t_1, t_2 \in [0, t_{\text{max}}]$) is additionally calculated in an ensemble-averaging manner as follows:

$$\frac{1}{mN_r} \sum_{k=1}^{N_r} \sum_{i=0}^{m-1} v^{(k)}(it_{\text{max}})v^{(k)}(it_{\text{max}} + t_1)v^{(k)}(it_{\text{max}})v^{(k)}(it_{\text{max}} + t_2). \quad (90)$$

For both systems, it is roughly calculated with $m = 1$ and $t_{\text{max}} = 10$ and. Then, it is more accurately calculated with $m = 64$ and $t_{\text{max}} = 1$ for the pure LJ fluid and $t_{\text{max}} = 4$ for the colloidal system.

For averaging over identical particles in the pure LJ fluid, the number of averaged particles n varies from 1 to 2048. Since the calculation of the VACF and the MSD for each fluid particle requires large memory and computational time in the time-averaging case, we reduce the computation size by simulating a smaller number of sample trajectories, $N_r = 512$.

Mean-squared velocity $\langle v^2 \rangle$. In order to confirm that the equilibration is well performed and the temperature is well maintained during production runs, the mean-squared velocity $\langle v^2 \rangle$ is monitored. For the pure LJ fluid system, we obtain $\langle v^2 \rangle = 0.7218$ with standard error 9.7×10^{-5} from time-averaging, which is consistent with the equipartition theorem $\langle v^2 \rangle = k_B T/m_{\text{fl}}$ with the target temperature $T = 0.722$. For the colloidal particle system, we obtain $\langle v_{\text{tracer}}^2 \rangle = 1.1684 \times 10^{-2}$ with standard error 3.7×10^{-6} from time-averaging. In this case, we need to compare this value with the value estimated from the following expression for NVEPG ensemble [38]:

$$\left\langle v_{\text{tracer}}^2 \right\rangle = \frac{k_B T}{M_{\text{tracer}}} \frac{N_{\text{fl}} m_{\text{fl}}}{M_{\text{tracer}} + N_{\text{fl}} m_{\text{fl}}}. \quad (91)$$

In other words, for a finite multi-component system, if a microcanonical-ensemble (i.e., NVE) MD simulation is performed under periodic boundary conditions, the equipartition theorem $\langle v_{\text{tracer}}^2 \rangle = k_B T/M_{\text{tracer}}$ does not exactly hold. This is attributed to constraints on \mathbf{P} (total linear momentum) and \mathbf{G} (generator of infinitesimal Galilean boosts), which are additionally imposed on the system. Since the values of the mean-squared velocity predicted by Eq. (91) and the equipartition theorem are 1.1691×10^{-2} and 1.2033×10^{-2} , respectively, the simulation result accurately confirms Eq. (91).

4.2. Simulation results

Validity of GPA. As mentioned above, the validity of the GPA is investigated by the covariance function $\text{Cov}[v(0)v(t_1), v(0)v(t_2)]$. For each system, the MD result calculated from the direct calculation of $\langle v(0)v(t_1)v(0)v(t_2) \rangle$ is compared with the GPA result calculated through Eq. (89) by using the VACF. Fig. 7 shows the time profile of the VACF and the comparison of the MD and GPA results for the covariance function. Although the VACF of the tracer particle decays more slowly due to its large mass than the VACF of a fluid particle, the time profiles of the two VACFs have similar shapes containing a negative tail. While the agreement of the MD and GPA results for the covariance function is overall good in the self-diffusion case, the agreement is excellent in the tracer-diffusion case. This observation is consistent with the physical explanation of the GPA based on the generalized Langevin equation given in Section 2.4.1, see Eq. (34). We also note that the tracer-diffusion case, where the GPA holds very accurately, is different from the Langevin model system since its VACF has a negative tail.

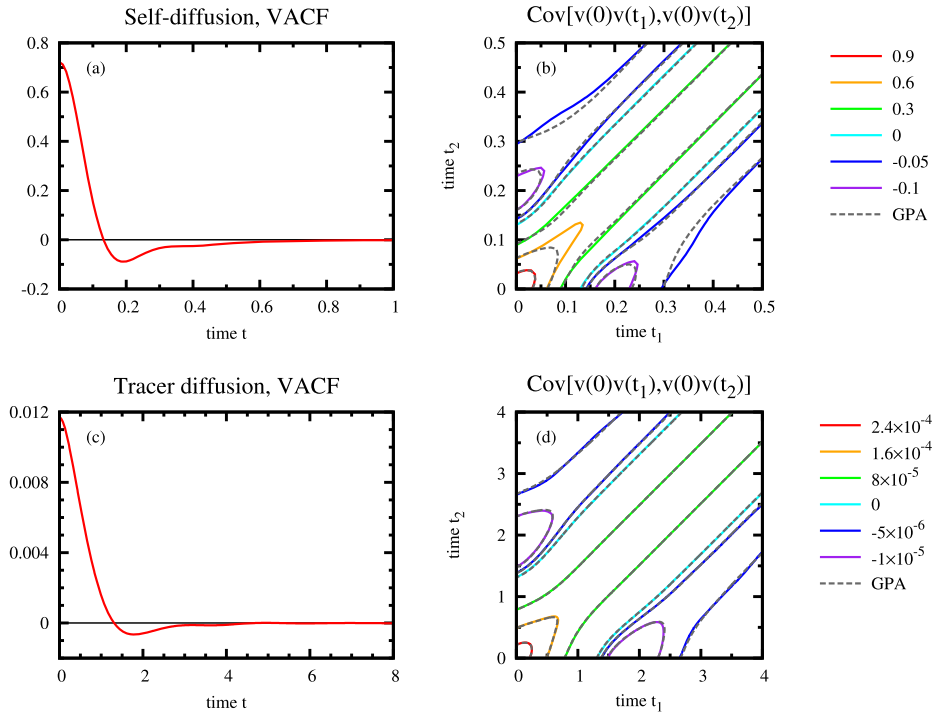


Fig. 7. For the self-diffusion of the pure LJ fluid and the tracer diffusion of the colloidal system, the time profiles of the VACF and the contour plots of covariance function $\text{Cov}[v(0)v(t_1), v(0)v(t_2)]$ are presented. In the contour plots, the GPA predictions from Eq. (89) are also plotted by the dashed lines.

Time profiles of VACF, MSD, and $D(t)$. Before analyzing the error statistics, we observe the time profiles of the VACF, the MSD, and $D(t)$, which are shown in the top row of Fig. 8 for the self-diffusion case and in Fig. 9 for the tracer-diffusion case. As mentioned above, the time scale for the decay of the VACF is larger in the tracer-diffusion case due to large mass of the tracer particle. This time scale can be also found in the asymptotic linear increase of the MSD and the long-time saturation of $D(t)$. However, except different magnitudes of the time scales, the overall shapes of the three quantities are more or less similar for the two cases. Hence, we may not predict the level of validity of the GPA by investigating the VACF or the MSD.

Since the time profiles of the VACF, the MSD and $D(t)$ presented in Fig. 8 and Fig. 9 are obtained from the time-averaging, their statistical errors are estimated as $\bar{\sigma}/\sqrt{N_r T}$ from the normalized standard errors $\bar{\sigma}$. From the plateau values of $D(t)$, the diffusion coefficients are estimated from time interval [9, 10] as $3.303 \times 10^{-2} \pm 5.3 \times 10^{-5}$ in the self-diffusion and $6.083 \times 10^{-3} \pm 9.6 \times 10^{-6}$ in the tracer-diffusion case. The value of the self-diffusion coefficient is consistent with the value 3.2619×10^{-2} obtained by using 1372 fluid particles in Ref. [22].

Validity of theoretical error estimates. In Fig. 8 and Fig. 9, the standard errors estimated from the MD ensemble runs are compared with our theoretical error estimates derived in Section 2.5. While the agreement is overall good in the self-diffusion case, the agreement is perfect in the tracer-diffusion case. This is consistent with the observation for the level of validity of the GPA in the two systems. Considering that even rough error estimates are useful in practice, our theoretical error estimates provide fairly accurate information for the level of the statistical errors even if the GPA does not hold exactly.

Scaling behavior of standard errors for particle-averaging. Finally, we consider averaging over identical particles and investigate the scaling behavior of the statistical errors for the number of averaged particle n . If n is much smaller than the total number N of identical particles in the system, then we expect that the velocity processes of the averaged particles are almost uncorrelated and thus the statistical errors decrease by factor of $n^{-1/2}$. However, if n is comparable to N , then we may not neglect the correlation among the velocity processes and the scaling behavior $n^{-1/2}$ may fail to hold. In Appendix B, under the GPA of the velocity processes, we show that this scaling behavior still holds for the NVEPG ensemble of a pure fluid system. We test this prediction for the pure LJ fluid system. For various values of n from 1 to $N_{fl} = 2048$, we calculate the normalized standard errors by additionally multiplying the factor \sqrt{n} . As shown in Fig. 10, the scaling behavior of the standard errors is well maintained up to $n = 256$ for all quantities. However, beyond this value of n , the scaling behavior fails in some cases. In the ensemble-averaging case, it is well maintained for the VACF and $D(t)$. Even for the MSD, it is still valid for short times. In the time-averaging case, it is roughly valid for the VACF and only valid at short times for $D(t)$.

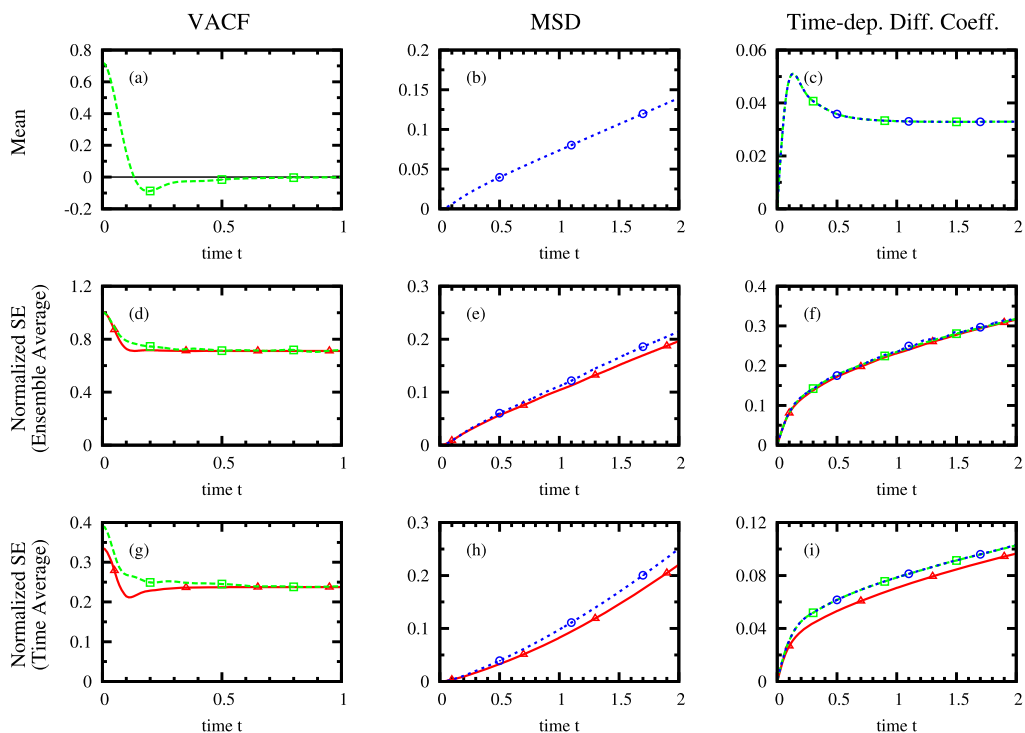


Fig. 8. For the pure LJ fluid system, the time profiles of the VACF, the MSD, and the time-dependent diffusion coefficient and their normalized standard errors (SE) are presented. The MD simulation data obtained from the time-averaging procedure with $\mathcal{T} = 500$ and $N_r = 32\,768$ realizations are plotted in the top row and their normalized standard errors $\bar{\sigma}$ (see Eq. (80)) are plotted in the middle row. The standard errors of the three quantities plotted in the top row are estimated by $\bar{\sigma}/\sqrt{N_r\mathcal{T}}$. In the bottom row, the normalized standard errors estimated from the ensemble-averaging procedure with $\mathcal{N} = 1$ and the same number of realizations are plotted. The red lines with the triangles indicates the GPA prediction estimated from the VACF and the MSD through our theoretical error estimates derived in Section 2.5. The green lines with the squares and the blue lines with the circles indicate the MD simulation results obtained by the VACF and MSD methods, respectively. (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

We guess that the failure of the prediction given in Appendix B is due to the failure of the GPA for the velocity processes, which is even stronger than the GPA for a single velocity process.

5. Summary and discussion

For the evaluation of the diffusion coefficient D from MD simulations, we have investigated the uncertainty due to finite sampling (i.e., statistical errors). We have analyzed two standard methods, which use the time integral of the VACF and the long-time slope of the MSD. By using theoretical and computational approaches, we have presented a complete systematic analysis on the standard errors of all quantities involved in both methods. We have concluded that the VACF and MSD methods are equivalent in the sense that they provide the same mean values with the same standard errors. The equivalence of the VACF and MSD methods can be understood as follows: although the quality of the MSD data is much better than that of the VACF data, the former becomes worse through time differentiation and the latter becomes better through time integration.

In our theoretical approach, we have derived explicit expressions for the standard errors of the VACF, the MSD, and the time-dependent diffusion coefficient $D(t)$, see Section 2.5. It is the error correlation functions of the VACF and the MSD that play a crucial role. Compared with the standard errors of the VACF and the MSD, they contain additional information on the error statistics between different times. We have derived relations (21) and (23), which have enabled us to calculate the standard errors of $D(t)$ from the error correlation functions. However, since the direct calculation of the error correlation functions from MD simulations is impractical due to huge computational overhead, we have introduced the GPA of the velocity process. By using the property that all higher-order statistics of a Gaussian process can be expressed in terms of the second-order statistics, we have expressed all theoretical error estimates in terms of the VACF. In fact, since the fourth-order cross moments of x and v are needed to be approximated by the GPA, we have shown that the deviation in our theoretical error estimates due to the GPA is proportional to the fourth-order cumulant of v . Since the error correlation functions depend on the type of averaging procedure, we have defined estimators representing several averaging procedures and calculated their error correlation functions. We have considered time-averaging, ensemble-averaging, hybrid

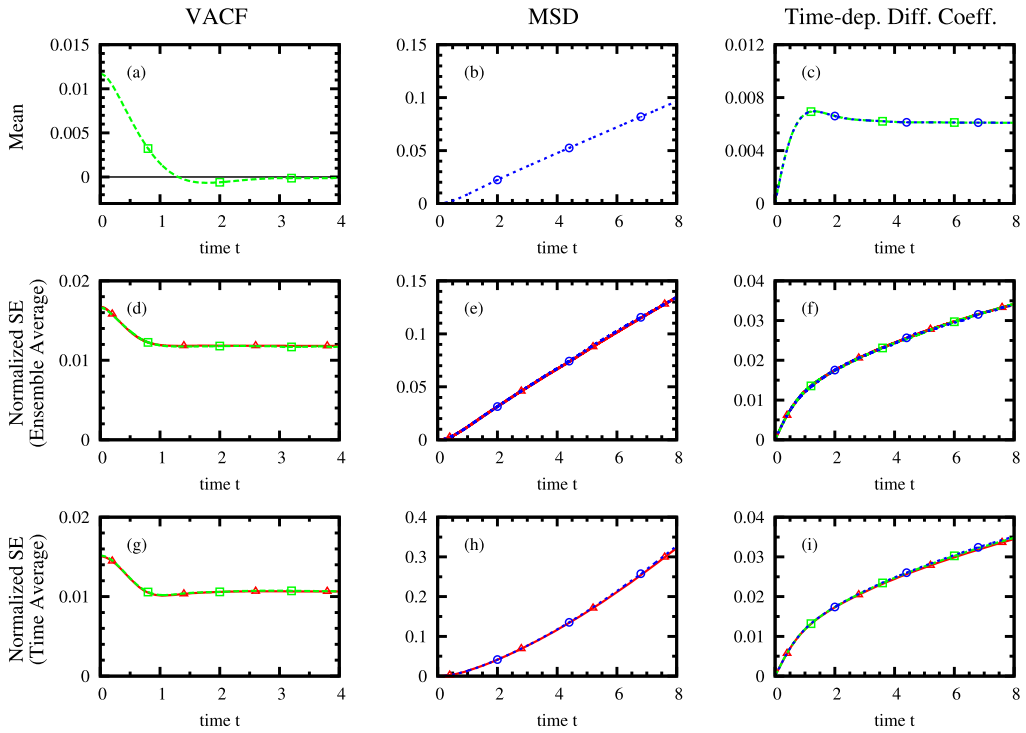


Fig. 9. For the colloidal system, the time profiles of the VACF, the MSD, and the time-dependent diffusion coefficient and their normalized standard errors (SE) are presented. Since the plots are presented in the same manner as in Fig. 8, see the caption of Fig. 8.

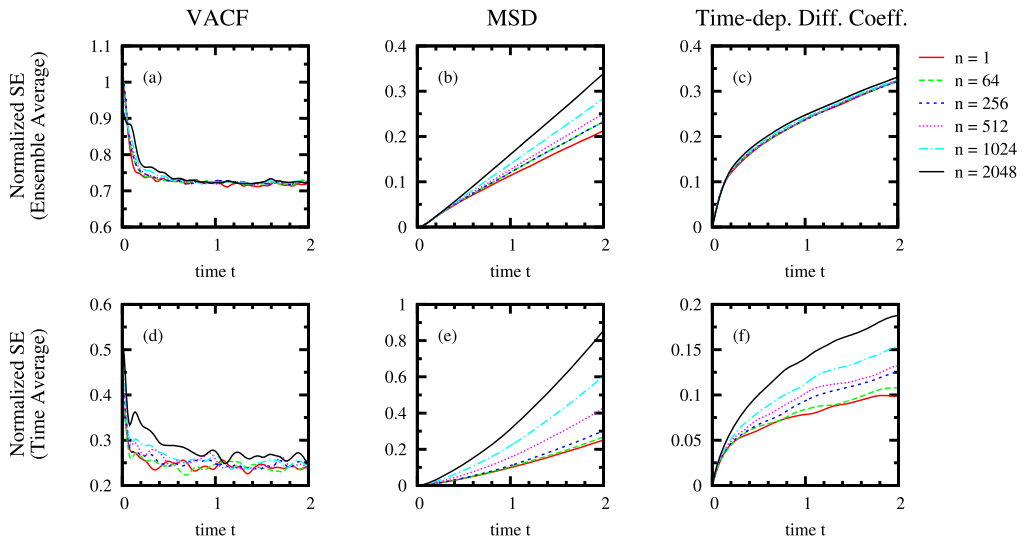


Fig. 10. For the pure LJ fluid system, the normalized standard errors (SE) estimated from the ensemble-particle and time-particle averaging procedures are plotted. Due to additional averaging over identical particles, the standard errors are also normalized by \sqrt{n} , where n is the number of averaged particles. The accordance of the curves indicates the scaling behavior of the statistical error as $n^{-1/2}$. Due to expensive computation, a smaller number of realizations $N_r = 512$ is used.

ensemble/time-averaging as well as averaging over identical particles. By considering the Langevin models driven by GWN and PWSN, we have verified our theoretical predictions.

In our computational approach, we have investigated the validity of our theoretical results. Most importantly, since the accuracy of our theoretical error estimates largely relies on the validity of the GPA, we have investigated the latter for two kinds of molecular systems: the pure LJ fluid system and the colloidal system. In both cases, we have observed that the GPA holds very well. As expected from the physical explanation of the GPA based on the generalized Langevin equation, see Eq. (34), we have also observed that the GPA holds almost perfectly for the tracer diffusion of a large and heavy colloidal

particle. Then, by comparing the standard errors estimated from MD simulations with our theoretical error estimates, we have confirmed the validity of our theoretical results. In order to accurately estimate the standard errors in a direct way from MD simulations, we have performed large-sized MD ensemble runs of size $N_r = 32678$.

The practical implications of the present work are as follows. First of all, once the VACF is roughly estimated, the order of magnitude for the standard errors of all relevant quantities can be readily estimated from our theoretical estimates. This is useful especially when MD simulation is so expensive that the standard errors cannot be directly estimated from a large-sized ensemble run. Second, for a pure fluid system, averaging over *all* identical particles is favorable for ensemble-averaging but not so practical for time-averaging. While the all-particle-averaging is not so expensive for the ensemble-averaging procedure, it requires huge memory and computational time for the time-averaging. Furthermore, it has been observed that the $n^{-1/2}$ scaling behavior of the standard errors on the number of averaged particle n quickly fails when n increases in the time-averaging case.

Our theoretical approach has the following advantageous features. First, it is applicable to particle-based methods such as the Langevin dynamics and the dissipative particle dynamics. As discussed in Section 2.4, it basically requires only the condition $\dot{x}(t) = v(t)$ and the GPA of v . Second, it can be generalized to the evaluation of other transport coefficients such as the shear viscosity and the thermal diffusivity, which are computable from the Green–Kubo relations or the Einstein–Kubo–Helfand relations. The most important issue may be the validity of the GPA for the underlying process.

Acknowledgements

The authors would like to thank Professor Mike Kirby of the University of Utah for discussions. This research was sponsored by the U.S. Army Research Laboratory and was accomplished under Cooperative Agreement Number W911NF-12-2-0023.

Appendix A. Asymptotic approximation for large \mathcal{T}

For an arbitrary integrable function $h(t)$, the following relation holds:

$$\frac{1}{\mathcal{T}^2} \int_0^{\mathcal{T}} dt' \int_0^{\mathcal{T}} dt'' h(t' - t'') = \frac{1}{\mathcal{T}} \int_{-\infty}^{\infty} h(\alpha) d\alpha + \varepsilon \quad (\text{A.1})$$

where

$$\varepsilon = -\frac{1}{\mathcal{T}^2} \int_0^{\mathcal{T}} dt \int_t^{\infty} d\alpha [h(\alpha) + h(-\alpha)]. \quad (\text{A.2})$$

We obtain Eq. (A.1) as follows. By introducing a new variable $\alpha = t' - t''$ to the inner integral of the left-hand side of Eq. (A.1), we have $\frac{1}{\mathcal{T}^2} \int_0^{\mathcal{T}} dt' \int_{t'-\mathcal{T}}^{t'} d\alpha h(\alpha)$. By approximating the latter by $\frac{1}{\mathcal{T}^2} \int_0^{\mathcal{T}} dt' \int_{-\infty}^{\infty} d\alpha h(\alpha)$ (i.e., by extending the integration range of its inner integral), we obtain the first term in the right-hand side of Eq. (A.1). The error of the approximation becomes the term ε defined in Eq. (A.2).

If $h(t)$ becomes zero rapidly as $t \rightarrow \pm\infty$, ε becomes negligible in Eq. (A.1) as $\mathcal{T} \rightarrow \infty$. More precisely, if $I_\varepsilon = -\int_0^{\infty} dt \int_t^{\infty} d\alpha [h(\alpha) + h(-\alpha)]$ is finite, then we have $\mathcal{T}^2 \varepsilon \rightarrow I_\varepsilon$ as $\mathcal{T} \rightarrow \infty$ and, thus, $\mathcal{T} \varepsilon \rightarrow 0$. Hence, for large \mathcal{T} , we can neglect the ε term in the right-hand side of Eq. (A.1), which is equal to $\frac{1}{\mathcal{T}} (\int_{-\infty}^{\infty} h(\alpha) d\alpha + \mathcal{T} \varepsilon)$. Two simple examples of $h(t)$ having finite I_ε are as follows: (1) If there is a time scale τ such that $h(t) = 0$ for $|t| > \tau$, then $I_\varepsilon = -\int_0^{\tau} dt \int_t^{\tau} d\alpha [h(\alpha) + h(-\alpha)]$; (2) If $h(t)$ decays exponentially, i.e., $h(t) = Ae^{-|t|/\tau}$, then $I_\varepsilon = -2A\tau^2$.

Appendix B. Scaling behavior of standard errors for particle-averaging in a pure liquid

For a pure fluid system consisting of N identical fluid particles, we first investigate a scaling behavior of the standard errors for a particle-averaging estimator for the VACF and then discuss a general scaling behavior for an estimator including particle-averaging. The particle-averaging estimator for the VACF is defined as follows:

$$\hat{w}(t) = \frac{1}{n} \sum_{i=1}^n v_i(0) v_i(t) \quad (\text{B.1})$$

where n is the number of identical particles to be included for averaging and v_i is the velocity of the i th particle. Then, we have $\langle \hat{w}(t) \rangle = \langle v(0)v(t) \rangle$ and

$$\text{Var}[\hat{w}(t)] = \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}[v_i(0)v_i(t), v_j(0)v_j(t)]. \quad (\text{B.2})$$

Under the GPA of the processes $\{v_i(t)\}_{i=1}^N$, we have

$$\text{Var}[\hat{w}(t)] = \frac{1}{n} \left\{ \langle v^2 \rangle^2 + \langle v(0)v(t) \rangle^2 \right\} + \frac{1}{n^2} \sum_{i=1}^n \sum_{j \neq i} \left\{ \langle v_i v_j \rangle^2 + \langle v_i(0)v_j(t) \rangle \langle v_j(0)v_i(t) \rangle \right\}. \quad (\text{B.3})$$

Hence, if the processes $\{v_i(t)\}_{i=1}^n$ are uncorrelated, then we obtain the same result as the ensemble-averaging case, see Eq. (51). However, in the NVEPG ensemble, we have $\sum_{j=1}^N v_j(t) = 0$. By multiplying $v_i(0)$ and taking average, we obtain

$$\langle v_i(0)v_j(t) \rangle = -\frac{1}{N-1} \langle v(0)v(t) \rangle, \quad \text{for } i \neq j. \quad (\text{B.4})$$

By substituting Eq. (B.4) into Eq. (B.3), we obtain

$$\text{Var}[\hat{w}(t)] = \frac{1}{n} \left[1 + \frac{n-1}{(N-1)^2} \right] \left\{ \langle v^2 \rangle^2 + \langle v(0)v(t) \rangle^2 \right\}. \quad (\text{B.5})$$

The scaling factor behaves almost like $\frac{1}{n}$ for $1 \leq n \leq N$; even in the case $n = N$, the scaling factor becomes $\frac{1}{N-1}$.

We note that the same argument is applicable to the error correlation function of \hat{w} , and hybrid estimators including particle-averaging for the VACF. Hence, almost-like $\frac{1}{\sqrt{n}}$ scaling behavior of the standard errors is expected for the VACF method. Similarly, it is applicable to the MSD method. However, we note that the GPA of the processes $\{v_i(t)\}_{i=1}^N$ is even stronger than that of a single process $v(t)$.

References

- [1] M.P. Allen, D.J. Tildesley, *Computer Simulation of Liquids*, Clarendon Press, Oxford, 1987.
- [2] J.M. Haile, *Molecular Dynamics Simulation: Elementary Methods*, Wiley-Interscience, New York, 1997.
- [3] D. Frenkel, B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd edition, Academic Press, San Diego, 2002.
- [4] D.C. Rapaport, *The Art of Molecular Dynamics Simulation*, 2nd edition, Cambridge University Press, New York, 2004.
- [5] D.J. Evans, G.P. Morriss, *Statistical Mechanics of Nonequilibrium Liquids*, Academic Press, New York, 1990.
- [6] M.S. Green, Markoff random processes and the statistical mechanics of time-dependent phenomena. II. Irreversible processes in fluids, *J. Chem. Phys.* 22 (1954) 398–413.
- [7] R. Kubo, Statistical-mechanical theory of irreversible processes. I. General theory and simple applications to magnetic and conduction problems, *J. Phys. Soc. Jpn.* 12 (1957) 570–586.
- [8] E. Helfand, Transport coefficients from dissipation in a canonical ensemble, *Phys. Rev.* 119 (1960) 1–9.
- [9] S. Viscardy, J. Servantie, P. Gaspard, Transport and Helfand moments in the Lennard-Jones fluid. I. Shear viscosity, *J. Chem. Phys.* 126 (2007) 184512.
- [10] S. Viscardy, J. Servantie, P. Gaspard, Transport and Helfand moments in the Lennard-Jones fluid. II. Thermal conductivity, *J. Chem. Phys.* 126 (2007) 184513.
- [11] D. Frenkel, Simulations: the dark side, in: *Proceedings of the International School of Physics "Enrico Fermi"*, vol. 184, Elsevier, 2013, pp. 195–227, also published in *Eur. Phys. J. Plus* 128 (2013) 10.
- [12] P. Angelikopoulos, C. Papadimitriou, P. Koumoutsakos, Bayesian uncertainty quantification and propagation in molecular dynamics simulations: a high performance computing framework, *J. Chem. Phys.* 137 (2012) 144103.
- [13] I.-C. Yeh, G. Hummer, System-size dependence of diffusion coefficients and viscosities from molecular dynamics simulations with periodic boundary conditions, *J. Phys. Chem. B* 108 (2004) 15873–15879.
- [14] D.M. Heyes, System size dependence of the transport coefficients and Stokes–Einstein relationship of hard sphere and Weeks–Chandler–Andersen fluids, *J. Phys. Condens. Matter* 19 (2007) 376106.
- [15] A. Papoulis, *Probability, Random Variables, and Stochastic Processes*, McGraw–Hill, New York, 1965.
- [16] R. Zwanzig, N.K. Ailawadi, Statistical error due to finite time averaging in computer experiments, *Phys. Rev.* 182 (1969) 280–283.
- [17] O.P.L. Maître, O.M. Knio, PC analysis of stochastic differential equations driven by Wiener noise, *Reliab. Eng. Syst. Saf.* 135 (2015) 107–124.
- [18] D. Frenkel, Intermolecular spectroscopy and computer simulations, in: *Proceedings of the International School of Physics "Enrico Fermi"*, vol. 75, Elsevier, 1980, pp. 156–201.
- [19] R.E. Jones, K.K. Mandadapu, Adaptive Green–Kubo estimates of transport coefficients from molecular dynamics based on robust error analysis, *J. Chem. Phys.* 136 (2012) 154102.
- [20] P. Hänggi, Langevin description of Markovian integro-differential master equations, *Z. Phys. B* 36 (1980) 271–282.
- [21] C. Kim, E.K. Lee, P. Hänggi, P. Talkner, Numerical method for solving stochastic differential equations with Poissonian white shot noise, *Phys. Rev. E* 76 (2007) 011109.
- [22] K. Meier, A. Laesecke, S. Kabelac, Transport coefficients of the Lennard-Jones model fluid. II Self-diffusion, *J. Chem. Phys.* 121 (2004) 9526–9535, and references therein.
- [23] M. Cappezzo, C.A. Capellari, S.H. Pezzin, L.A.F. Coelho, Stokes–Einstein relation for pure simple fluids, *J. Chem. Phys.* 126 (2007) 224516.
- [24] F. Ould-Kaddour, D. Levesque, Molecular-dynamics investigation of tracer diffusion in a simple fluid: test of the Stokes–Einstein law, *Phys. Rev. E* 63 (2000) 011205.
- [25] J.R. Schmidt, J.L. Skinner, Hydrodynamic boundary conditions, the Stokes–Einstein law, and long-time tails in the Brownian limit, *J. Chem. Phys.* 119 (2003) 8062–8068.
- [26] J.R. Schmidt, J.L. Skinner, Brownian motion of a rough sphere and the Stokes–Einstein law, *J. Phys. Chem. B* 108 (2004) 6767–6771.
- [27] F. Ould-Kaddour, D. Levesque, Diffusion of nanoparticles in dense fluids, *J. Chem. Phys.* 127 (2007) 154514.
- [28] H. Flyvbjerg, H.G. Petersen, Error estimates on averages of correlated data, *J. Chem. Phys.* 91 (1989) 461–466.
- [29] R. Friedberg, J.E. Cameron, Test of the Monte Carlo method: fast simulation of a small Ising lattice, *J. Chem. Phys.* 52 (1970) 6049–6058.
- [30] N.G. van Kampen, *Stochastic Processes in Physics and Chemistry*, 3rd edition, Elsevier, North-Holland, Boston, 2007.
- [31] C. Gardiner, *Stochastic Methods: A Handbook for the Natural and Social Sciences*, 4th edition, Springer, Berlin, 2009.
- [32] I. Snook, *The Langevin and Generalized Langevin Approach to the Dynamics of Atomic, Polymeric and Colloidal Systems*, Elsevier, Boston, 2007.
- [33] H.K. Shin, C. Kim, P. Talkner, E.K. Lee, Brownian motion from molecular dynamics, *Chem. Phys.* 375 (2010) 316–326.

- [34] R.D. Groot, P.B. Warren, Dissipative particle dynamics: bridging the gap between atomistic and mesoscopic simulation, *J. Chem. Phys.* 107 (1997) 4423–4435.
- [35] P.E. Kloeden, E. Platen, *Numerical Solution of Stochastic Differential Equations*, Springer-Verlag, Berlin, 1992.
- [36] C. van den Broeck, On the relation between white shot noise, Gaussian white noise, and the dichotomic Markov process, *J. Stat. Phys.* 31 (1983) 467–483.
- [37] D. Levesque, L. Verlet, J. Kürkijarvi, Computer "experiments" on classical fluids. IV. Transport properties and time-correlation functions of the Lennard-Jones liquid near its triple point, *Phys. Rev. A* 7 (1973) 1690–1700.
- [38] M.J. Uline, D.W. Siderius, D.S. Corti, On the generalized equipartition theorem in molecular dynamics ensembles and the microcanonical thermodynamics of small systems, *J. Chem. Phys.* 128 (2008) 124301.