Extraction of self-diffusivity in systems with nondiffusive short-time behavior

Sachin Shanbhag*
Department of Scientific Computing, Florida State University, Tallahassee, Florida 32306, USA
(Received 3 July 2013; revised manuscript received 30 August 2013; published 25 October 2013)

We consider a toy model that captures the short-time nondiffusive behavior seen in many physical systems, to study the extraction of self-diffusivity from particle trajectories. We propose and evaluate a simple method to automatically detect the transition to diffusive behavior. We simulate the toy model to generate data sets of varying quality and test different methods of extracting the self-diffusion coefficient and characterizing its uncertainty. We find that weighted least-squares with statistical bootstrap is the most accurate and efficient means for analyzing the trajectory data. The analysis suggests an iterative recipe for designing simulations to conform to a specified level of accuracy.

DOI: 10.1103/PhysRevE.88.042816 PACS number(s): 82.20.Wt, 83.85.Ns

I. INTRODUCTION

Determination of the self-diffusivity D of a particle undergoing Brownian motion, from an analysis of particle motion, has a long and checkered past. The first of two popular strategies relies on Green-Kubo relationships; it computes D by integrating the velocity autocorrelation function. This method is widely used in molecular dynamics simulations [1,2]. The second strategy involves using particle positions, rather than velocities; it computes the mean-squared displacement (MSD) curve, and estimates D by appealing to the Einstein relation. This method, which forms the basis of this work, is more commonly employed in Monte Carlo simulations [3] and in single-particle tracking experiments [4], where accurate velocity data are not readily available. Formally, both these methods are equivalent, and any differences in estimated diffusivities can be traced to the numerical algorithms employed and the treatment of "noise" in the computation. A comparison of these two methods may be found, for example, in Ref. [5].

In its simplest, and perhaps most common, form, the self-diffusion coefficient can be extracted by performing an unweighted linear least-squares (LS) fit on the MSD curve. However, more careful analyses recognize and account for the fact that different points that make up the MSD curve are correlated and known with different degrees of certainty [4,6–9]. These sophisticated methods give us more accurate estimates of D and the associated uncertainty. In other words, these studies address the crucial question, "How much faith should I attach to a particular estimate of self-diffusivity?" The focus of these studies has primarily been the interpretation and resolution of localization uncertainty that characterizes single-particle microscopy data in experiments where the particle undergoes Brownian motion. Naturally, the assumption of Brownian motion is embedded deep within the machinery for extracting self-diffusivity and its confidence intervals.

On the other hand, in many physically important systems studied using computer simulation, a long-time diffusive behavior is preceded by an early-time nondiffusive behavior, which constitutes a non-negligible fraction of the observation window. This includes diffusion of the center of mass of polymer chains in melts and solutions [10–15], movement of a Brownian particle trapped in an inverted colloidal crystal matrix [16,17], technologically important confined fluid systems such as polyatomic molecules in nanoporous adsorbates like zeolites and metal organic frameworks [18], permeation of small molecules through polymeric membranes [19], and ions in solid conductors [20,21], or, more generally, the motion of a particle in an obstacle environment [22–24].

Unlike purely diffusive systems, the plot of MSD versus time swings from nonlinear at short time scales to linear at longer time scales. From a practical standpoint, both the localization uncertainty associated with data acquisition in experiments alluded to above and the nondiffusive early-time behavior in molecular simulations "corrupt" the MSD at short time scales, where the data happen to be well averaged and most reliable. Despite this superficial resemblance, the advanced methods developed to address the former cannot be directly refashioned to address the latter, due to their reliance on the assumption of Brownian motion.

A. Scope

Thus, in this work, we consider the extraction and uncertainty quantification of D from MSD data obtained from computer simulations, for systems with nondiffusive short-time behavior. We propose and simulate a minimal model, which reproduces the most relevant features of such systems. We develop and test an algorithm to infer D automatically, by analyzing the MSD data to estimate the transition from nondiffusive to diffusive behavior and the weights to use in linear regression. Further, we evaluate the validity of bootstrapping as a data-driven technique for characterizing the uncertainty in the computed diffusivities. Finally, we use lessons learned from a parametric analysis to suggest ways to improve the design of the particle simulation.

B. Background

The trajectory of a single particle, $\mathbf{r}(t)$, may be described by taking snapshots at fixed time intervals Δt and reporting its position $\mathbf{r}_i = \mathbf{r}(t = (i-1)\Delta t)$. Given such a discretized trajectory, $\{\mathbf{r}_1, \mathbf{r}_2, \ldots, \mathbf{r}_N\}$, the most common definition of the MSD $\hat{\rho}(t) = \langle(\mathbf{r}(t) - \mathbf{r}(0))^2\rangle$ is [4]

$$\hat{\rho}_n = \hat{\rho}(t = n\Delta t) = \frac{1}{N-n}\sum_{i=1}^{N-n}(\mathbf{r}_{i+n} - \mathbf{r}_i)^2, \quad 1 \leqslant n \leqslant N-1, \tag{1}$$

*sshanbhag@fsu.edu

 ©2013 American Physical Society

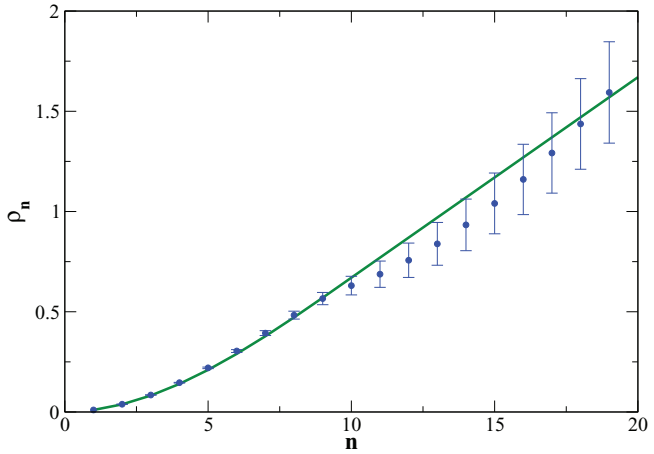

FIG. 1. (Color online) The theoretical MSD [solid (green) line] for the minimal model introduced in Sec. II A is nonlinear for $n < 10$, after which it becomes strictly linear. The same model is used to simulate $n_p = 50$ independent particle trajectories. The ensemble-average MSD of the simulated trajectories and the associated standard error, σ_n, are represented by filled circles.

where the hat on ρ is used to distinguish an MSD computed from a single particle. For $N \gg 1$ and $n \ll N$, the number of terms in the summation $(N-n)$ is large; this results in well-averaged values for $\hat{\rho}_n$. For a particle diffusing in d-dimensional space, the Einstein relation $\langle(\mathbf{r}(t)-\mathbf{r}(0))^2\rangle = 2dDt$ connects the self-diffusivity D and the MSD. If the MSDs of n_p independent particle trajectories, $\hat{\rho}_n^j$, are available, it is possible to define an ensemble-averaged MSD as

$$\rho_n = \frac{1}{n_p}\sum_{j=1}^{n_p}\hat{\rho}_n^j = \frac{1}{n_p}\frac{1}{N-n}\sum_{j=1}^{n_p}\sum_{i=1}^{N-n}\left(\mathbf{r}_{i+n}^j - \mathbf{r}_i^j\right)^2, \quad (2)$$

where the superscript $j, 1 \leqslant j \leqslant n_p$ is used to index a particular particle trajectory. In the molecular simulation literature, D is quite commonly inferred by a naive unweighted LS estimate of the MSD data, via $\rho_n = (2dD\Delta t)n$.

As mentioned earlier, many factors muddle such a straightforward analysis. Consider Fig. 1, which plots the theoretical and simulated MSD (averaged over $n_p = 50$ independent runs) of the minimal model discussed in Sec. II A. The shape of the simulated curve beyond the early nondiffusive regime is not unambiguously linear; it is also characterized by a relatively large uncertainty σ_n [4,6–8], which is defined as the standard error of mean of ρ_n via

$$\sigma_n^2 = \frac{1}{n_p}\left(\frac{1}{n_p}\sum_{j=1}^{n_p}\left(\hat{\rho}_n^j\right)^2 - \rho_n^2\right). \quad (3)$$

Similarly, for purely diffusive motion, it can be shown that if only the first few ρ_n values are used to estimate D, then statistically the LS estimate also corresponds to a maximum likelihood estimate, because the MSD is approximately normally distributed in this regime. However, a closer look at Fig. 1 clearly reveals the perils of using only the data at small n in this case.

Another subtle issue related to the variance in the MSD is the correlation in ρ_n. In LS fitting, even when σ_n is properly accounted for by weighting the points, it is tacitly assumed that ρ_n are uncorrelated. For pure diffusion, expressions for the correlation matrix have been derived and can be incorporated to properly appraise the uncertainty associated with the self-diffusion coefficient [7].

It may be noted that these definitions of the average ρ_n and σ_n are "trajectory-centric." It is also possible to define ρ_n and σ_n based purely on displacements, without regard to the trajectory from which they originate. It is relatively easy to show that such a definition of ρ_n is formally equivalent to Eq. (2), and while such a "displacement-centric" σ_n is not equivalent to Eq. (3), the primary results of this paper would remain mostly unchanged even if this alternative definition of σ_n were used.

II. METHODS

We first describe a simple one-dimensional (1D) minimal model which is diffusive at long time scales but nondiffusive at short time scales. Next, we propose a data-driven method to trace this transition, based on the curvature of the MSD plot. We then briefly describe weighted and unweighted LS and two methods to quantify the uncertainty in the estimated D.

A. Model

Consider a particle initially at $r = 0$ moving at a constant velocity of $v = +1$ units in a periodic lattice of "gates" separated by a distance of $L = 1$ unit, as shown in Fig. 2. When the particle "arrives" at a gate, it is reflected back instantaneously (the direction of v is reversed) with a probability of 0.5. If it is not reflected, it continues through the gate until it encounters another gate, where the same scenario unfolds again. The characteristic time between arrivals at gates is $\tau = L/v = 1$ unit.

At time scales much smaller than τ, the particle moves at a uniform velocity. At much larger time scales, the particle hops between gates, and its motion resembles simple 1D diffusion. Let us assume that we take snapshots of the particle at $\Delta t = 0.1\tau$ and compute the MSD according to Eq. (1). The theoretical MSD for this process can be computed exactly (see the Appendix) and is depicted by the line in Fig. 1. From the theory of 1D random walks, it can be shown that for $t \gg \tau$,

$$\frac{d\rho(t)}{dt} = 2D = \frac{L^2}{\tau}.$$

Since $\tau = L = 1$ here, $D = 0.5$.

The model can also be directly simulated to produce n_p independent particle trajectories and, hence, MSD curves from Eq. (1). Given the n_p MSD curves, one can compute ρ_n and σ_n at each point [Eqs. (2) and (3)]; this is depicted in Fig. 1 for $n_p = 50$ particles, with $L = v = \tau = 1$, a simulation time of $\tau_{\text{sim}} = 2\tau$, and $\Delta t = 0.1\tau$. At small n, the simulated average has low variation and matches the theoretical curve quite well. At large n both the variation and the potential deviation from the theoretical $\rho(t)$ increase.

FIG. 2. (Color online) Schematic of the minimal 1D model.

TABLE I. Number of particles and simulation time considered for the three data sets analyzed in this work. Throughout this paper, $\tau = L = v = 1$ and $\Delta t = 0.1\tau$.

Quality	n_p	$\tau_{\rm sim}/\tau$
Low	50	2
Medium	100	5
High	500	10

As n_p and $\tau_{\rm sim}$ become large, the simulated average MSD curve asymptotically approaches the theoretical curve. In such cases, where well-averaged long simulation data are available, LS estimates of D converge, as demonstrated shortly. In such cases, the preponderance of the data is very forgiving; it even tolerates sloppy data analysis. Usually, however, we are confronted with much more limited data, and this luxury is not affordable.

In this work, we initially considered three data sets with different n_p and $\tau_{\rm sim}$, as reported in Table I. The guiding principle here was to construct data sets of varying "quality" and assess the performance of different methods of data analysis.

B. Detection of transition to a linear regime

To compute D, we first need to identify and discard the early-time MSD ($t \leqslant \tau$). In the minimal model described above, we set $\tau = 1$. In general, the characteristic time τ is not known and has to be inferred from the data. A straightforward method to check whether the infinite-time limit of the Einstein relation has been sufficiently sampled is to compute the exponent relating the MSD to time and retaining only the portion where the exponent is one. This method works well when high-quality data are available, and a log-log plot of ρ_n verus n reveals the characteristic time τ. In this paper, we propose and test an alternative algorithm for finding the transition point. It assumes that $\rho(t) \sim t^\alpha$ ($0 \leqslant \alpha \leqslant 2$ is a continuously varying parameter) transitions to $\rho(t) \sim t$ around $t \sim \tau$. This procedure offers advantages over the straightforward method for data sets of questionable quality, where detecting a transition in slopes is notoriously hard and prone to subjective bias.

Since the plot of ρ_n versus n is not linear at small n, we numerically compute the curvature of the MSD curve and seek to identify the point $n = p$ at which the curvature becomes 0. In practice, to be mindful of the error associated with the data points, we consider the absolute value of the curvature of ρ_n normalized by σ_n:

$$\text{curvature} = \frac{1}{\sigma_n}\left|\frac{d^2\rho_n}{dn^2}\right|. \tag{4}$$

This quantity has the advantage of being dimensionless and builds on our prior expectation that the nondiffusive regime has the highest quality data (small σ_n). For small n, $d^2\rho_n/dn^2$ is nonzero, and σ_n is small, making the dimensionless curvature [Eq. (4)] large. As n increases, $d^2\rho_n/dn^2$ decreases in magnitude, even as σ_n increases. The net effect is that the dimensionless curvature is a decreasing function of n.

We define the truncation point p to be the minimum n at which the curvature as defined above, numerically computed by the second-order difference formula $d^2\rho_n/dn^2 \approx \rho_{n+1} - 2\rho_n + \rho_{n-1}$, first drops below 0.01. Note that choosing a value much smaller than 0.01 as the cutoff results in a more conservative estimate for p, at the risk of unnecessarily throwing away high-quality data. On the other hand, choosing a much larger value for the cutoff results in a more aggressive estimate of p, at the risk of letting information from the nondiffusive regime taint the estimated D.

Note that this criterion assumes that the point at which the "early-time nonlinear regime ends" marks the transition to the diffusive regime. It does not explicitly test whether the data for $n > p$ are linear ($\rho_n \sim n$), since it is very difficult to do so for low-quality data sets, in any case. Presumably, the error bars for diffusion coefficients estimated will reflect the quality of the underlying data sets.

The overall procedure for inferring p may be summarized as follows: (i) simulate an ensemble of n_p particles for time $\tau_{\rm sim}$, (ii) compute the MSDs of the individual trajectories using Eq. (1), (iii) compute ρ_n and σ_n from Eqs. (2) and (3), (iv) compute the dimensionless curvature defined in Eq. (4) using a second-order difference scheme for ρ_n, and (v) find the $n = p$ at which this quantity first drops below 0.01.

We applied this simple criterion to 1000 independent simulations of ρ_n corresponding to the three data sets in Table I. Figure 1, for example, corresponds to one of these independent simulations of the low-quality data set. Histograms of p for the three data sets are reported in Fig. 3. For the high-quality data set, we obtained a mean value of $(p+1)\Delta t = 1.00 \pm 0.00$, where the error represents the standard deviation over the 1000 samples. Similarly, for the medium- and low-quality data sets the corresponding mean values were 0.95 ± 0.05 and 0.93 ± 0.15, respectively. These values all compare well with $\tau = 1$. The slightly smaller values for the medium- and low-quality data sets are acceptable, since the transition from nondiffusive to diffusive behavior is not abrupt; the empirical MSD curves often approach linearity well before τ. Changing the cutoff criterion between 0.005 and 0.02 for Eq. (4) did

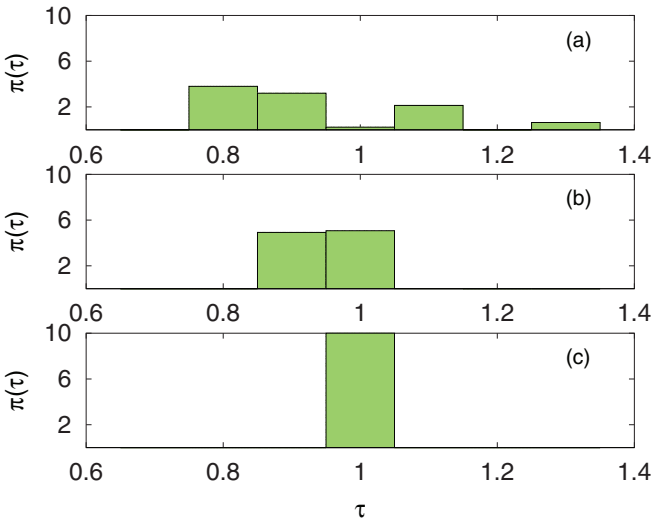

FIG. 3. (Color online) Histograms of $\tau = (p+1)\Delta t$ computed for the (a) low-, (b) medium-, and (c) high-quality data sets reported in Table I using 1000 independent ensembles.

not change these results significantly, and hence we stuck with 0.01 throughout this paper.

C. Least-squares

After identifying and discarding the nonlinear portion of the MSD, we fit the rest of the data to the equation $\rho_i = 2Di(\Delta t) + c$, where $p < i \leqslant N$ and c is the intercept. We begin by building the matrix $\mathbf{A}$,

$$\mathbf{A} = \begin{bmatrix} 1 & (p+1)\Delta t \\ 1 & (p+2)\Delta t \\ \vdots & \vdots \\ 1 & N\Delta t \end{bmatrix}. \tag{5}$$

We also construct a diagonal "weighting" matrix $\mathbf{w}$, whose elements are inversely proportional to the variance $\mathbf{w}_{ii} = 1/\sigma_i^2$, and a column vector $\mathbf{b} = [\rho_{p+1}, \ldots, \rho_N]^T$, which contains the linear portion of the MSD.

We then solve the weighted linear LS problem for D using normal equations or QR factorization [25],

$$\mathbf{A}^T\mathbf{w}\mathbf{A}\begin{bmatrix} c \\ 2D \end{bmatrix} = \mathbf{A}^T\mathbf{w}\mathbf{b}. \tag{6}$$

Setting $\mathbf{w} = \mathbf{I}$ in the equation above gives us the unweighted estimate for D, which is also considered here because of its prevalence in the literature. Note that we use all the points $n > p$ in the LS estimation; for particular subproblems where a detailed model for ρ_n is available, it is possible to explore the optimization problem of using only a subset of these points. We have not done so in this paper, to align the treatment more closely with standard procedures currently used in molecular simulation research.

D. Distribution of estimated diffusivities

For a given n_p and τ_{sim} in Table I, we consider two alternative methods to quantify the uncertainty in the estimated diffusivity. In the first method, we perform 1000 independent replicas at each n_p and τ_{sim} and, hence, compute a value of D for each replica. For comparison, we compute both the weighted, D_w, and the unweighted, D_u, LS estimate for each replica. We assume that the histograms of D_w and D_u approximate the true distribution of these quantities. While this is a reasonable technique for analyzing the minimal model, it is somewhat impractical in realistic simulations; the generation of a single replica is often a computationally challenging task, in itself (e.g., diffusion of long polymers).

Hence we consider an alternative method, based on simple statistical bootstrap. In this method, we perform a *single independent* ensemble simulation at each setting of n_p and τ_{sim} in Table I. We compute individual MSDs for all the n_p particles in this "original" data set and average them to estimate D. We then generate a new "bootstrap" sample from the original data set by randomly resampling n_p MSDs, *with replacement*. Thus, unlike the original sample, the n_p MSDs in the bootstrap sample are not all different, and in general, we expect copies of MSDs to be present in the latter. It is easy to see that the total number of bootstrap samples that can be constructed from n_p independent particles in the original data set is $n_p^{n_p}$, which is a huge number to sample exhaustively. In this work we consider 10 000 bootstrap samples derived from the original sample.

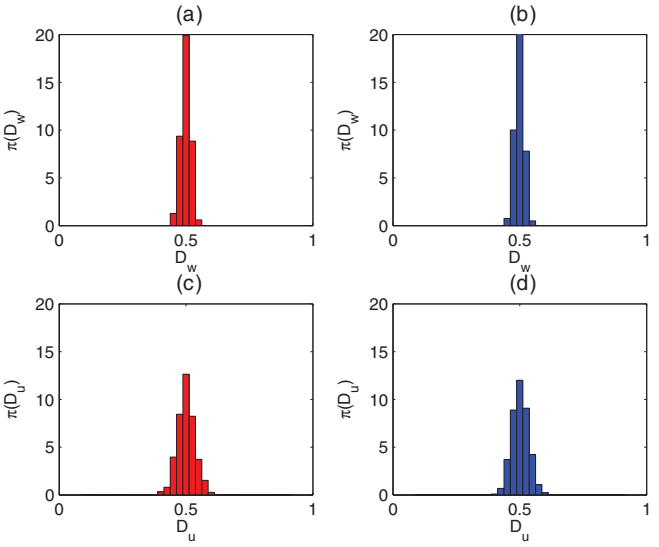

FIG. 4. (Color online) Histograms of D computed for the high-quality data set with $n_p = 500$ and $\tau_{\text{sim}}/\tau = 10$. Weighted and unweighted LS were used on 1000 independent samples to generate the (red) histograms (a) and (c), respectively. Similarly, weighted and unweighted LS were used on 10 000 (blue) bootstrap samples to generate histograms (b) and (d), respectively.

III. RESULTS

Figure 4 depicts the histograms, $\pi(D)$, obtained for the high-quality data set with $n_p = 500$ and $\tau_{\text{sim}}/\tau = 10$. The top and bottom rows show the weighted and unweighted LS estimates [Eq. (6)], respectively; the left and right columns show the histograms using 1000 independent runs and 10 000 bootstrap samples, respectively. In all cases, $\pi(D)$ shows a narrow, approximately symmetric, distribution around the true theoretical value of 0.5.

The agreement between the histograms deduced from independent and bootstrapped samples, for both D_w and D_u, is quite remarkable. This observation is practically useful. Typically, the computational effort required to simulate particle trajectories far exceeds that required for their analysis. In other words, generating independent samples to model the uncertainty in D escalates the already substantial computational cost. The use of bootstrap samples avoids this escalation, since only a single ensemble simulation is carried out; it shifts the computational load from the generation of trajectories, which is expensive, to their analysis, which is much cheaper. This is apparent even in the present case: despite the simplicity of the minimal model, the amount of effort required to generate the distribution in Fig. 4(a) was 2.5 h, versus 13 min for that in Fig. 4(b), on a single modern processor.

In Fig. 4, the mean (and standard deviation) for the independent runs was $\bar{D}_w = 0.499 \pm 0.019$ and $\bar{D}_u = 0.501 \pm 0.035$. The larger standard deviation of the latter is visually apparent from the width of $\pi(D)$ in Figs. 4(a) and 4(c). Similarly, the mean for the bootstrapped samples was $\bar{D}_w = 0.498 \pm 0.017$ and $\bar{D}_u = 0.502 \pm 0.033$.

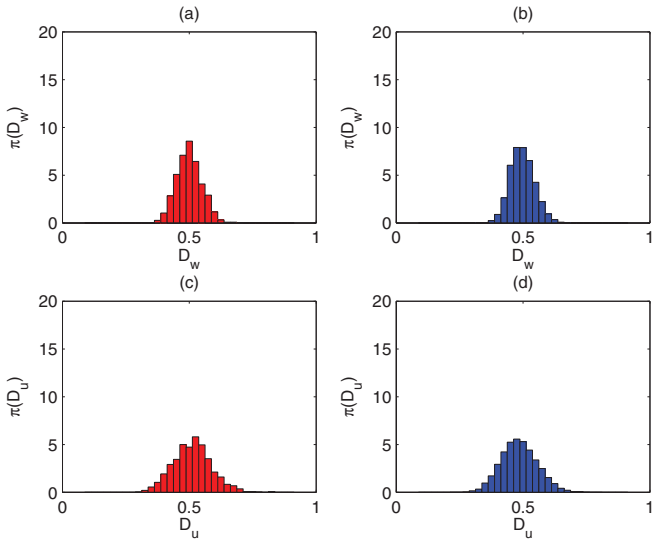

FIG. 5. (Color online) Histograms of D computed for $n_p = 100$ and $\tau_{\rm sim}/\tau = 5$. Weighted and unweighted LS were used on 1000 independent samples to generate the (red) histograms (a) and (c), respectively. Similarly, weighted and unweighted LS were used on 10 000 (blue) bootstrap samples to generate histograms (b) and (d), respectively.

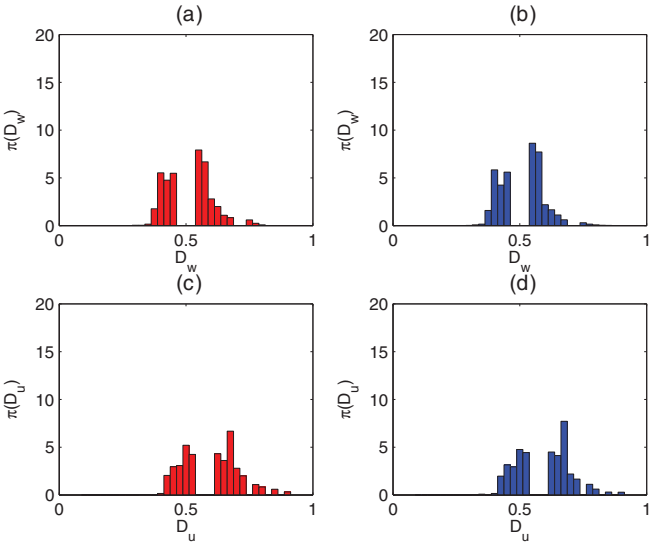

FIG. 6. (Color online) Histograms of D computed for $n_p = 50$ and $\tau_{\rm sim}/\tau = 2$. Weighted and unweighted LS were used on 1000 independent samples to generate the (red) histograms (a) and (c), respectively. Similarly, weighted and unweighted LS were used on 10 000 (blue) bootstrap samples to generate histograms (b) and (d), respectively.

Thus, on the basis of the high-quality data set, we can draw two tentative conclusions: (i) bootstrapping is a convenient and accurate alternative to uncertainty quantification via independent simulations, and (ii) given the same trajectory data, analysis using weighted LS is preferable to that using unweighted LS because of the smaller variance. Next, we examine how these tentative conclusions hold up for the medium- and low-quality data sets.

Figure 5 depicts the histograms, $\pi(D)$, obtained for the medium-quality data set with $n_p = 100$ and $\tau_{\rm sim}/\tau = 5$. In all cases, the distribution is roughly symmetric around the true value of D. The mean for the independent runs was $\bar{D}_w = 0.500 \pm 0.050$ and $\bar{D}_u = 0.512 \pm 0.076$ [Figs. 5(a) and 5(c)]. While the means are comparable to the high-quality data sets, we note that the width of the corresponding distributions has increased. This is not surprising, since the uncertainty associated with the estimated D depends on the quality of the underlying data set. Similarly, the mean for the bootstrapped samples was $\bar{D}_w = 0.497 \pm 0.047$ and $\bar{D}_u = 0.485 \pm 0.071$ [Figs. 5(b) and 5(d)].

The low-quality data set is shown in Fig. 6. The distribution $\pi(D)$ has a spurious "void" near 0.5, which makes it qualitatively different from the $\pi(D)$ for the medium- and high-quality data sets. This void is a direct consequence of the simplicity of the minimal model for $\tau_{\rm sim}/\tau = 2$. An intuitive, albeit incomplete, explanation stems from the realization that a single particle, with initial conditions $r = 0$ and $v = +1$, tracks one of only two possible trajectories. It either passes through the gate at $r = L$ at $t = \tau$, to end up at $r = 2L$ at $t = \tau_{\rm sim} = 2\tau$, or is reflected back at the gate to end up at $r = 0$. Estimates of diffusivity based on the former (latter) trajectory overestimate (underestimate) the true diffusivity of 0.5 for this toy model, leading to a void near the true diffusivity. In the actual simulation, instead of a single trajectory, we have an ensemble of particles, which results in a cluster with $D > 0.5$ if the ensemble is rich in particles that pass through the gate, and vice versa. In other words, the void is a "feature" of the toy model itself.

The mean and standard deviation for the independent runs was $\bar{D}_w = 0.514 \pm 0.0920$ and $\bar{D}_u = 0.598 \pm 0.112$ [Figs. 6(a) and 6(c)]. For the bootstrap runs, $\bar{D}_w = 0.512 \pm 0.088$ and $\bar{D}_u = 0.595 \pm 0.108$. Note that this low-quality dataset underscores the importance of proper data analysis.

It is clear that the quality of the estimated D depends on the quality of the underlying data set; the standard deviation, σ_D, varies inversely with n_p and $\tau_{\rm sim}$. In any case, it should be noted that σ_D provides a reasonable estimate of the uncertainty in the estimated D, regardless of the method of analysis chosen.

In summary, we find that the two tentative conclusions drawn from the high-quality data set appear to hold remarkably well for the medium- and low-quality data sets. This provides strong support for the use of weighted LS with bootstrapping to determine confidence intervals, which are applied for all the computations that follow. As we do not use unweighted LS in the rest of the paper, we drop the subscript in D_w and simply use the symbol D.

A. Design of good-quality data sets

We now turn our attention to the issue of efficient simulation design. It is obvious that increasing both n_p and $\tau_{\rm sim}$ improves the quality of the data set (Table I), as reflected in the more accurate estimation of $\bar{D}$ via a smaller σ_D. Here, we decouple n_p and $\tau_{\rm sim}$ and ask the question, "If I hold one of these quantities constant, and change the other, how does my estimate of $\bar{D}$ and σ_D change?"

We generalize the three cases in Table I, by considering a range of $\tau_{\rm sim}$ and n_p. It should be pointed out that system-specific considerations may impose hard constraints on the

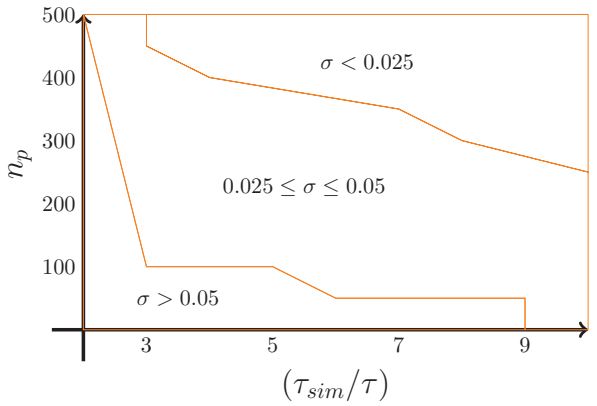

FIG. 7. (Color online) The uncertainty in the estimated self-diffusivity is characterized by σ_D for a range of $\tau_{\rm sim}/\tau$ and n_p. For simplicity, we partition the sampled parameter space into regions of low ($\sigma_D < 0.025$), medium ($0.025 \leqslant \sigma_D \leqslant 0.05$), and high uncertainty ($\sigma_D > 0.05$).

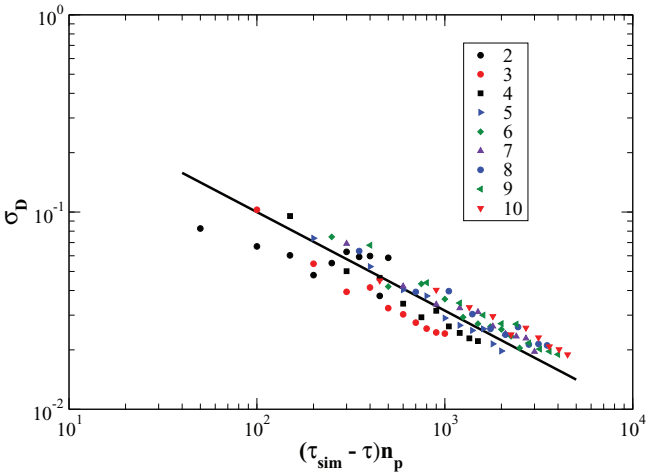

FIG. 8. (Color online) The data considered in Fig. 7 are replotted to show the variation of the uncertainty σ_D as a function of the "useful" computational effort, which is quantified by $n_p(\tau_{\rm sim} - \tau)$ instead of $n_p\tau_{\rm sim}$ to reflect the lack of information in the early nondiffusive part of the simulation. Different symbols correspond to different values of $\tau_{\rm sim}/\tau$ reported in the legend. The solid line is drawn after Eq. (7).

minimum n_p and $\tau_{\rm sim}$ allowed. For example, in simulations of polymers, the radius of gyration may determine the minimum simulation box size and, hence, the minimum n_p, and the reptation or relaxation time may determine an approximate τ and, hence, a minimum $\tau_{\rm sim}$.

Here, we varied $\tau_{\rm sim}$ between 2τ and 10τ, in increments of τ ($\tau = 1$ was held fixed), and n_p between 50 and 500, in increments of 50. We used weighted LS to estimate $\bar{D}$ and σ_D from 10 000 bootstrap samples. In Fig. 7, we present a bird's-eye view of the results by dividing the n_p versus $\tau_{\rm sim}/\tau$ space into regions of low ($\sigma_D < 0.025$), medium ($0.025 \leqslant \sigma_D \leqslant 0.05$), and high uncertainty ($\sigma_D > 0.05$). As n_p and $\tau_{\rm sim}$ increase, σ_D becomes smaller.

We can interpret the data more quantitatively to address an important practical concern: "Given a fixed computational budget, are certain choices of n_p and τ_{sim} better than others?" Note that the computational complexity of simulating the trajectories of n_p particles for a simulation time $\tau_{\rm sim}$ is $\mathcal{O}(n_p\tau_{\rm sim})$. For example, in molecular dynamics simulations, the use of cell lists and neighbor lists reduces the complexity of particles that interact via short-range pairwise potentials from $\mathcal{O}(n_p^2\tau_{\rm sim})$ to $\mathcal{O}(n_p\tau_{\rm sim})$ [3]. Thus, the computational cost of doubling n_p while holding $\tau_{\rm sim}$ constant is *approximately* the same as that of doubling $\tau_{\rm sim}$ while holding n_p constant ["approximately" because prefactors like $\log(n_p)$ arise in certain implementations of Ewald summation, for example].

For the systems we are interested in here, the early-time nondiffusive behavior is not particularly useful in the determination of the long-time self-diffusivity. Roughly, the fraction $\tau/\tau_{\rm sim}$ of the computational effort expended is "thrown away" during the analysis. Since τ is a constitutive property of the material being studied, this revelation suggests that, at the margins, increasing $\tau_{\rm sim}$ is preferable to increasing n_p, as it leads to less waste.

In other words, the amount of information in the trajectories is proportional to $n_p(\tau_{\rm sim} - \tau)$, although the computational cost is proportional to $n_p\tau_{\rm sim}$. Figure 8 plots σ_D as a function of the amount of information used to estimate $\bar{D}$, $n_p(\tau_{\rm sim} - \tau)$, on a log-log scale. The data approximately appear to obey a relationship implied by the central-limit theorem,

$$\sigma_D \sim \frac{1}{\sqrt{n_p(\tau_{\rm sim} - \tau)}}, \tag{7}$$

as indicated by the solid line. If we assume that σ_D is a good measure of the accuracy of the estimated D, this provides a recipe for choosing n_p and $\tau_{\rm sim}$ based on accuracy requirements.

(1) Carry out a simulation with a reasonable guess for n_p and $\tau_{\rm sim}$.

(2) Compute the MSD using Eq. (1).

(3) Estimate $\tau = p\Delta t$ by analyzing the transition of the MSD to linearity.

(4) Use weighted LS with statistical bootstrap to estimate D and σ_D.

(5) If σ_D is larger than required, extrapolate Eq. (7) to increase $\tau_{\rm sim}$ (or n_p) to the required level of accuracy and repeat the process.

IV. SUMMARY AND PERSPECTIVE

Physical systems in which a long-time diffusive behavior is preceded by an early-time nondiffusive behavior are quite common. Extracting the self-diffusivity of such systems using the MSD requires special care. In this paper, we have considered a toy model, which captures these essential features, and presented a simple method to recognize the crossover to diffusive behavior based on the curvature of the MSD. We then apply weighted and unweighted linear LS to estimate self-diffusivity in the toy problem and conclude that weighted LS was clearly superior.

Inferring confidence intervals on the estimated self-diffusion coefficients is important, because in many physical simulations the characteristic time τ is quite long and not known precisely; there is always a danger of having too much confidence in a value of self-diffusivity from relatively "short" simulations. We found statistical bootstrap to be a valid and

cheap alternative for characterizing the uncertainty in the estimates.

We found that the quality of the estimated diffusivity covaries with the quality of the underlying data set. While the computational cost increases with system size and simulation time as $\mathcal{O}(n_p\tau_{\rm sim})$, the uncertainty in the estimated D measured by its standard deviation decreases according to $\sigma_D \sim (n_p(\tau_{\rm sim}-\tau))^{-0.5}$. Collectively, these findings suggest a recipe for designing simulations to a prescribed level of uncertainty in the estimated self-diffusivity.

ACKNOWLEDGMENT

This work is based in part upon work supported by the National Science Foundation under Grant No. NSF CAREER DMR-0953002.

APPENDIX: THEORETICAL MSD OF THE TOY MODEL

To theoretically evaluate the MSD of the toy model, we can carry out an exact enumeration. For example, let us consider the evaluation of the first few ρ_n. For concreteness, let us assume $L = 1$, $\Delta t = 0.1$, $\Delta = L\Delta t = 0.1$, and $v = +1$. Note that no real loss of generality is incurred by making these assumptions. To compute ρ_n, we have to consider all the trajectories of n steps starting at $r = 0, \Delta, 2\Delta, \ldots, 9\Delta$ and average over them.

For $n = 1$, all 10 different starting points lead to a displacement of Δ. Thus,

$$\rho_1 = \frac{\sum_{i=0}^{9}\Delta^2}{10} = \Delta^2. \tag{A1}$$

For $n = 2$, nine of the trajectories starting at $r = 0, \Delta, \ldots, 8\Delta$ lead to a displacement of 2Δ. Trajectories starting at $r = 9\Delta$ can either go past the gate to $r = 11\Delta$ (displacement of 2Δ) or be reflected back at the gate with probability 1/2 for a net displacement of 0. Thus,

$$\rho_2 = \frac{\sum_{i=0}^{8}(2\Delta)^2 + \left(\frac{1}{2}(2\Delta)^2 + \frac{1}{2}0\right)}{10} = 3.8\Delta^2. \tag{A2}$$

A similar calculation for $n = 3$ yields

$$\rho_3 = \frac{\sum_{i=0}^{7}(3\Delta)^2 + \left(\frac{1}{2}(3\Delta)^2 + \frac{1}{2}\Delta^2\right)\left(\frac{1}{2}(3\Delta)^2 + \frac{1}{2}\Delta^2\right)}{10} = 8.2\Delta^2. \tag{A3}$$

As n increases, the exact enumeration becomes cumbersome, as the number of trajectories to consider scale approximately as $2^{\text{floor}(n\Delta/L)}$. However, the simplicity of the model allows us to carry out the exact enumeration on a computer in a reasonable amount of time. Computation of the theoretical curve in Fig. 1 required less than a minute on a single modern processor.

[1] M. Allen and D. Tildesley, *Computer Simulation of Liquids* (Oxford University Press, New York, 1989).
[2] D. C. Rapaport, *The Art of Molecular Dynamics Simulation*, 2nd ed. (Cambridge University Press, Cambridge, UK, 2004).
[3] D. Frenkel and B. Smit, *Understanding Molecular Simulation: From Algorithms to Applications*, 2nd ed. (Academic Press, Orlando, FL, 2002).
[4] H. Qian, M. P. Sheetz, and E. L. Elson, Biophys. J. **60**, 910 (1991).
[5] D. J. Keffer, B. J. Edwards, and P. Adhangale, J. Non-Newton. Fluid Mech. **120**, 41 (2004).
[6] M. J. Saxton, Biophys. J. **72**, 1744 (1997).
[7] X. Michalet, Phys. Rev. E **82**, 041914 (2010).
[8] A. J. Berglund, Phys. Rev. E **82**, 011917 (2010).
[9] X. Michalet and A. J. Berglund, Phys. Rev. E **85**, 061916 (2012).
[10] G. Subramanian and S. Shanbhag, Macromolecules **41**, 7239 (2008).
[11] S. Wang, S.-Q. Wang, A. Halasa, and W.-L. Hsu, Macromolecules **36**, 5355 (2003).
[12] A. De Cecca and J. J. Freire, Polymer **44**, 2589 (2003).
[13] V. A. Harmandaris, V. G. Mavrantzas, D. N. Theodorou, M. Kröger, J. Ramrez, H. C. Öttinger, and D. Vlassopoulos, Macromolecules **36**, 1376 (2003).
[14] K. Hur, C. Jeong, R. G. Winkler, N. Lacevic, R. H. Gee, and D. Y. Yoon, Macromolecules **44**, 2311 (2011).
[15] C. D. Chapman, S. Shanbhag, D. E. Smith, and R. M. Robertson-Anderson, Soft Matter **8**, 9177 (2012).
[16] S. Shanbhag, J. Lee, and N. A. Kotov, Biomaterials **26**, 5581 (2005).
[17] H. Zhou, S. B. Chen, J. Peng, and C-H. Wang, J. Colloid Interf. Sci. **342**, 620 (2010).
[18] D. S. Sholl, Acc. Chem.l Res. **39**, 403 (2006).
[19] F. Muller-Plathe, Acta Polymer. **45**, 259 (1994).
[20] M. Saiful Islam, J. Mater. Chem. **10**, 1027 (2000).
[21] E. M. Calvo-Muñoz, M. E. Selvan, R. Xiong, M. Ojha, D. J. Keffer, D. M. Nicholson, and T. Egami, Phys. Rev. E **83**, 011120 (2011).
[22] A. M. Berezhkovskii, V. Y. Zitserman, and S. Y. Shvartsman, J. Chem. Phys. **119**, 6991 (2003).
[23] A. M. Berezhkovskii, V. Y. Zitserman, and S. Y. Shvartsman, J. Chem. Phys. **118**, 7146 (2003).
[24] H. Zhou and S. B. Chen, Phys. Rev. E **79**, 021801 (2009).
[25] G. H. Golub and C. F. Van Loan, *Matrix Computations*, 3rd ed. (Johns Hopkins University Press, Baltimore, MD, 1996).