Some connections between importance sampling and enhanced sampling methods in molecular dynamics

H. C. Lie[1,2,a)] and J. Quer[1,b)]
[1]*Zuse Institut Berlin, Takustrasse 7, 14195 Berlin, Germany*
[2]*Institut für Mathematik, Freie Universität Berlin, Arnimallee 6, 14195 Berlin, Germany*

(Received 9 June 2017; accepted 31 October 2017; published online 17 November 2017)

In molecular dynamics, enhanced sampling methods enable the collection of better statistics of rare events from a reference or target distribution. We show that a large class of these methods is based on the idea of importance sampling from mathematical statistics. We illustrate this connection by comparing the Hartmann-Schütte method for rare event simulation (J. Stat. Mech. Theor. Exp. **2012**, P11004) and the Valsson-Parrinello method of variationally enhanced sampling [Phys. Rev. Lett. **113**, 090601 (2014)]. We use this connection in order to discuss how recent results from the Monte Carlo methods literature can guide the development of enhanced sampling methods. *Published by AIP Publishing.* https://doi.org/10.1063/1.4989495

I. INTRODUCTION

The sampling problem of molecular dynamics[1] (MD) refers to the computational inefficiency of standard MD for the statistical estimation of certain properties of large or complex molecular systems. One cause of the sampling problem is the presence of rare events, which are often associated with high barriers on energy landscapes.

In this paper, by a "rare event," we shall refer to the event in which a molecule starting in some metastable state A transitions to another metastable state B within a period of time that is accessible by standard MD. Such rare events can lead to the sampling problem when a quantity of interest is given by the expected value of some path functional f (i.e., f takes as input a trajectory of the system originating from A and ending in B and returns as output a scalar), where the values of f tend to contribute less to the expected value as the duration of the transition path increases. This is the case for exponential work averages, which can be used in order to compute free energy differences.[2–4] Since the probability of observing a value of f that contributes significantly to the quantity of interest is small, the number of standard MD trajectories needed in order to get one such "good" sample value is high, and thus standard MD is computationally inefficient in this setting.

Many enhanced sampling techniques, such as metadynamics,[5] umbrella sampling,[6] replica exchange or parallel tempering,[7–12] and simulated annealing,[13] have been developed to tackle the sampling problem. A large class of methods tries to circumvent the sampling problem by increasing the probability that a MD trajectory is also a transition path from A to B. A subset of these methods is based on the fundamental principle that the energy landscape of a molecule plays an important role in determining its dynamics. Such methods increase the probability that a MD trajectory of feasible duration is also a transition path, by changing or "tilting" the energy landscape; the resulting dynamics are sometimes referred to as "accelerated dynamics,"[14] and the probability distribution on the trajectories of the (enhanced) molecular system is said to be "biased." The crucial idea that we wish to stress in this paper is that samples from the biased probability distribution are used to estimate statistical properties of the original, unbiased probability distribution.

In probability theory and mathematical statistics, the idea of using a "proposal" probability distribution to estimate some property of a "target" probability distribution is fundamental to the technique of importance sampling for Monte Carlo methods. Importance sampling is made possible by the Radon-Nikodym theorem, which states that if the target probability distribution is well behaved (i.e., absolutely continuous) with respect to the proposal distribution, then there exists a random variable called the *Radon-Nikodym derivative* of the target with respect to the proposal distribution, with which one can statistically reweight sample values drawn from the proposal in order to perform statistical estimation of properties of samples drawn from the target.

In this article, we aim to present some basic mathematical concepts that are frequently used in mathematical analyses of importance sampling methods and to explain the significance of these concepts for molecular dynamics. In particular, we highlight the connection between importance sampling and the class of enhanced sampling methods that are based on the idea of tilting the energy landscape. We illustrate this connection by comparing the rare event simulation method of Hartmann-Schütte[15] and the variationally enhanced sampling method of Valsson-Parrinello.[16] We show that this connection is useful, by discussing some recent results from the mathematical analysis of Monte Carlo methods and explaining their significance to developers of enhanced sampling methods.

The structure of this paper is as follows. In Sec. II, we describe the connections between enhanced sampling and importance sampling; Sec. II A provides a brief exposition of the enhanced sampling methods that we consider in this paper; Sec. II B provides an overview of importance sampling and

[a)]Electronic mail: hlie@math.fu-berlin.de
[b)]Electronic mail: quer@zib.de

describes how importance sampling and enhanced sampling methods are connected. In Sec. III, we illustrate the connection by comparing the Valsson-Parrinello and Hartmann-Schütte methods. Finally, in Sec. IV, we indicate how this connection may be significant for developers of enhanced sampling methods, by presenting some recent results from the mathematics community regarding importance sampling and discussing the implications of these results for developers of enhanced sampling methods.

II. MOLECULAR DYNAMICS AND PROBABILITY THEORY

A. The biasing function approach to the sampling problem

In molecular dynamics (MD), the Boltzmann-Gibbs distribution of the molecule determines many quantities of interest, such as mean first passage times and transition rates, in the sense that these quantities can be formulated as expected values of some function f of the trajectories of the molecule,

$$\mathbb{E}_\mu [f] = \int_{\mathcal{X}_{0:\tau}} f(x_{0:\tau})\mu(dx_{0:\tau}),$$

where $x_{0:\tau}$ denotes a given trajectory of finite duration $\tau > 0$, $\mathcal{X}_{0:\tau}$ denotes the set of all such trajectories, and μ denotes the "path measure," i.e., the probability distribution on the set $\mathcal{X}_{0:\tau}$ of trajectories associated with the Boltzmann-Gibbs distribution. More details on the connection between the probability distribution on state space and the distribution μ on path space can be found in Ref. 17. Since it is in practice impossible to calculate these expectations analytically, one resorts to a Monte Carlo approximation of the expected value by the empirical mean or sample mean,

$$\mathbb{E}_\mu[f] \approx \frac{1}{N}\sum_{i=1}^{N} f(X^i_{0:\tau}), \quad X^i_{0:\tau} \sim \mu \text{ i.i.d.}, \tag{1}$$

where $X^i_{0:\tau}$ denotes the ith random trajectory of duration τ out of N such trajectories, the notation $X^i_{0:\tau} \sim \mu$ emphasises that these trajectories are drawn from the path measure μ, and "i.i.d." means "independent and identically distributed."

In many cases, the quantity of interest $\mathbb{E}_\mu[f]$ is defined in terms of a rare event, such as the event in which the molecule transitions between two distinct, metastable conformations A and B in state space. By virtue of their rarity, for trajectories of moderate duration—e.g., durations accessible to standard MD methods—the probability of observing a transition from A to B will be small. Therefore, one will have to increase either the trajectory duration or the number of trajectories sampled, in order to obtain a reasonable approximation of the quantity of interest. Both these approaches are computationally expensive. For this reason, many methods have been developed to circumvent this sampling problem. The class of methods that we focus on in this paper approach the sampling problem in the following way:

1. Perturb the molecular system, by proposing a different path measure ν on the same set of trajectories $\mathcal{X}_{0:\tau}$ such that the ν-probability of the event of interest is higher than the μ-probability of the same event.
2. Draw independent, identically distributed (i.i.d.) trajectories $X^i_{0:\tau}$ from the proposal distribution ν.
3. Approximate the quantity of interest using the weighted Monte Carlo approximation

$$\mathbb{E}_\mu[f] \approx \frac{1}{N}\sum_{i=1}^{N} f(X^i_{0:\tau})w(X^i_{0:\tau}), \quad X^i_{0:\tau} \sim \nu \text{ i.i.d.}, \tag{2}$$

where w is a positive function on the set of trajectories that assigns a weight to each value $f(X^i_{0:\tau})$.

In molecular dynamics, the above approach is realized by exploiting the known correspondences between the energy function that determines the Boltzmann-Gibbs distribution and the associated distribution μ on path space (these correspondences hold, whether one considers dynamics in full state space or in the space determined by some collective variables). The class of enhanced sampling methods that we shall consider here are those that operate by changing the energy landscape of the molecule. In many situations, the negative gradient of the energy function dominates the other forces acting on the molecule, and thus many enhanced sampling methods involve adding a biasing term to the original energy function; this process is sometimes referred to as "tilting" or "biasing" the energy landscape. The idea is to change the energy landscape so that the molecule is more likely to undergo the desired transition, with the result that the Monte Carlo step (step 2 above) yields more occurrences of the desired event. However, since the quantity of interest is a statistical quantity with respect to μ and not ν, one needs to transform the empirical statistics of random variables drawn from ν, in order to estimate the corresponding statistics of random variables drawn from μ. In the approach mentioned above, this transformation is described by the statistical reweighting term $w(\cdot)$ in step 3.

Two significant challenges in developing enhanced sampling methods are those of finding biasing functions that increase the probability of observing the desired event and of designing statistical reweighting schemes that are both accurate and computationally efficient (in the sense of having low computational cost). The task of designing good biasing functions is known to be highly problem-specific, so we shall not discuss it here. Instead, we will consider the problem of statistical reweighting schemes since we can analyze this problem using existing mathematical theory.

B. Connections between importance sampling and enhanced sampling

The problem of importance sampling can be understood as a constrained optimisation problem. The problem involves estimating an expected value with respect to a target probability distribution, using samples drawn from an alternate "proposal" probability distribution, subject to a constraint involving statistically impossible events. We make this more precise below.

Let μ be a probability distribution on a set $\mathcal{X}$, let f be a real-valued function defined on $\mathcal{X}$, and let the quantity of interest be the expected value of f with respect to μ,

$$\mathbb{E}_\mu[f] = \int_{\mathcal{X}} f(x)\mu(dx).$$

Let ν be another probability distribution on $\mathcal{X}$. We define μ to be *absolutely continuous* with respect to ν if, for every subset $A \subset \mathcal{X}$ such that $\nu(A) = 0$, it also holds true that $\mu(A) = 0$. In the mathematical literature, absolute continuity of μ with respect to ν is written as $\mu \ll \nu$.

Absolute continuity of μ with respect to ν can be interpreted as the property that every event that is statistically impossible with respect to ν is also statistically impossible with respect to μ. Note that the notion of absolute continuity is not bidirectional, i.e., $\mu \ll \nu$ does not also imply that $\nu \ll \mu$. Indeed, if $\mu \ll \nu$, then the set of μ-statistically impossible events can be strictly larger than the set of ν-statistically impossible events. The following theorem establishes a significant property of pairs of absolutely continuous probability distributions.

Theorem 1 (Radon-Nikodym). *Let μ and ν be probability distributions on $\mathcal{X}$. If μ is absolutely continuous with respect to ν, then there exists an almost everywhere strictly positive function ρ on $\mathcal{X}$ such that for any function f for which $\mathbb{E}_\mu[f]$ exists and is finite,*

$$\mathbb{E}_\mu[f] = \mathbb{E}_\nu[f\rho], \tag{3}$$

where

$$\mathbb{E}_\nu[f\rho] = \int_{\mathcal{X}} f(x)\rho(x)\nu(dx).$$

The function ρ is called the "Radon-Nikodym derivative" of μ with respect to ν and is denoted by $\frac{d\mu}{d\nu}$. For the purposes of statistics, one can view the Radon-Nikodym derivative as being unique. Note that the notation $\frac{d\mu}{d\nu}$ does not mean that the function ρ can be expressed as the ratio of two quantities. However, the ratio notation does provide a mnemonic explanation for why absolute continuity implies the existence of a well-defined Radon-Nikodym derivative: since $\mu(A) = 0$ whenever $\nu(A) = 0$, it never happens that the numerator of the ratio $\mu(A)/\nu(A)$ is nonzero while the denominator is zero.

Given two different probability distributions μ and ν defined on a common set $\mathcal{X}$, there exist many quantitative descriptions that permit one to describe how different μ is from ν. A large class of such descriptions is given by f-divergences. Perhaps the most well-known divergence that one can compute for a pair of probability distributions μ and ν satisfying the property that $\mu \ll \nu$ is the Kullback-Leibler divergence or relative entropy,

$$D_{KL}(\mu\|\nu) = \mathbb{E}_\nu\left[-\log\frac{d\mu}{d\nu}\right]. \tag{4}$$

If μ is not absolutely continuous with respect to ν, the corresponding Kullback-Leibler divergence is defined to be $+\infty$. We will return to Kullback-Leibler divergences later.

To define the importance sampling problem, recall that the variance of the function f with respect to μ satisfies

$$\mathrm{Var}_\mu[f] = \mathbb{E}_\mu[(f - \mathbb{E}_\mu[f])^2] = \mathbb{E}_\mu[f^2] - (\mathbb{E}_\mu[f])^2.$$

Define the importance sampling estimator of $\mathbb{E}_\mu[f]$ with respect to a proposal distribution $\nu \ll \mu$ of sample size N by

$$I_N[f] = \frac{1}{N}\sum_{i=1}^{N} f(X^i)\rho(X^i), \quad X^i \sim \nu \text{ i.i.d.}. \tag{5}$$

The similarity between the right-hand sides of (5) and the approximation (2) in step 2 of the general scheme of enhanced sampling methods establishes the connection between importance sampling and the class of enhanced sampling methods that we consider here:

1. An importance sampling method on path space yields an enhanced sampling method because a Radon-Nikodym derivative is strictly positive and hence is a suitable statistical reweighting function. Moreover, by the law of large numbers and by (3), the importance sampling estimator $I_N[f]$ converges almost surely to the desired quantity $\mathbb{E}_\mu[f]$ in the large sample size limit.
2. Enhanced sampling methods can also be viewed as importance sampling methods if the strictly positive statistical reweighting function w in (2) satisfies

$$\mathbb{E}_\nu[w] = 1,$$

and this can be seen by setting f to be the constant function equal to one everywhere on $\mathcal{X}$ in (3) and using that $\mathbb{E}_\mu[1] = 1$ for any probability distribution μ. In this case, it follows that w is the Radon-Nikodym derivative of μ with respect to ν, and the random variable on the right-hand side of (2) is an importance sampling estimator.

Note that since X^i are i.i.d. draws from ν, it follows that

$$\mathbb{E}_\mu[f] = \mathbb{E}_\nu[I_n[f]], \tag{6}$$

i.e., the importance sampling estimator $I_N[f]$ is an unbiased estimator of the quantity of interest $\mathbb{E}_\mu[f]$. We can now formulate the importance sampling problem as the constrained optimization problem,

$$\text{minimize } \mathrm{Var}_\nu[f\rho] \text{ subject to } \mu \ll \nu.$$

We do not consider the variance of the estimator $I_N[f]$, but of the basic random variable $f\rho$, because the fact that the samples are independent and identically distributed implies that the ν-variance of $I_N[f]$ is simply the ν-variance of $f\rho$ divided by N.

Now we consider how the condition of absolute continuity in importance sampling may be related to enhanced sampling. By the Radon-Nikodym theorem, the condition of absolute continuity in importance sampling guarantees the existence of a unique statistical reweighting function, namely, the Radon-Nikodym derivative. In particular, if absolute continuity does not hold, then there is no statistical reweighting function w that one can use to transform sample statistics drawn from the proposal ν in order to estimate any expected value with respect to μ.

From the point of view of molecular dynamics, one can give absolute continuity a concrete interpretation, in terms of functions that perturb the energy landscape that defines the path measure. Recall that the target path measure μ is a probability distribution on the set of trajectories $\mathcal{X}_{0:\tau}$, that μ is defined by a canonical Boltzmann-Gibbs distribution, and that this density is defined by an energy function and inverse temperature β. If the proposal path measure ν is defined by perturbing the energy function—namely, by adding a biasing term to the original energy function—and keeping all

other factors such as temperature unchanged, and if the biasing term does not assume infinite values anywhere on its domain, then the resulting perturbed Boltzmann-Gibbs distribution is strictly positive at every point where the original Boltzmann-Gibbs distribution is strictly positive. This implies that the corresponding probability distributions—and hence the corresponding path measures μ and ν—are absolutely continuous with respect to each other. The significance of the finite-energy interpretation of absolute continuity is that any enhanced sampling method that biases the energy landscape by only finite amounts of energy will have a unique statistical reweighting function w given by the appropriate Radon-Nikodym derivative.

As is the case for enhanced sampling methods, the problem of designing an effective importance sampling bias is not straightforward and highly depends on the considered problem. The latter observation is suggested by the following well-known fact from the theory of importance sampling that one can verify by direct substitution: when the function f in the quantity of interest $\mathbb{E}_\mu[f]$ is strictly positive, the Radon-Nikodym derivative of μ with respect to the optimal proposal ν^* is given by $\frac{d\mu}{d\nu^*} = \mathbb{E}_\mu[f]/f$. In other words, the optimal proposal that yields a zero-variance estimator requires that one should know the desired expected value.

In some cases, the problem of finding an optimal importance sampling bias is related to finding a low-dimensional submanifold of $\mathcal{X}$ on which the function f contributes the most to the quantity of interest $\mathbb{E}_\mu[f]$. In these cases, finding good approximations of the coordinates that describe this low-dimensional submanifold is a necessary prerequisite for finding a good biasing function. The task of finding such coordinates coincides with the prerequisite of finding good collective variables for enhanced sampling methods.

We note that statistical reweighting in the context of importance sampling is fully justified, provided that the condition of absolute continuity is fulfilled. Equivalently, provided that the biasing function satisfies the finite-energy property, importance sampling proceeds without any additional conditions. In contrast, some enhanced sampling methods for obtaining kinetics from biased sampling are presented with the constraint that one should not apply a biasing function in the transition region or equivalently that the biasing function is zero in the transition region.[9,18]

III. ILLUSTRATION: THE VALSSON-PARRINELLO AND HARTMANN-SCHÜTTE METHODS

The material in this section takes place in some finite-dimensional space $\mathcal{S}$ of collective variables (CVs). We will write $s \in \mathcal{S}$ to denote a vector of CVs and $s_{0:\tau}$ to denote a trajectory of duration $\tau > 0$ in CV space. We assume that the reader is familiar with how CVs and free energy landscapes are defined.

A. The Valsson-Parrinello method

The Valsson-Parrinello method[16] involves a variational approach to enhanced sampling for complex systems in CV space, with the aim of computing the free energy landscape of the molecule in CV space. Some applications of the Valsson-Parrinello method in MD are given in Refs. 19–22.

Suppose that the target distribution on $\mathcal{S}$ is given by $Z^{-1}\exp(-\beta F(s))ds$, where F is an energy function on $\mathcal{S}$ and $Z = \int_{\mathcal{S}} \exp(-\beta F(s))ds$, and suppose that one wishes to sample from some proposal distribution $h(s)ds$, where $h(s)$ is positive everywhere on $\mathcal{S}$. We assume that h is a proper probability density so that $\int_{\mathcal{S}} h(s)ds = 1$. The authors define the following functional of a biasing function F_{bias}:

$$\phi(F_{\text{bias}}) = \frac{1}{\beta}\log\frac{\int e^{-\beta(F(s)+F_{\text{bias}}(s))}ds}{\int e^{-\beta F(s)}ds} + \int h(s)F_{\text{bias}}(s)ds \quad (7)$$

and show that the biasing function given by

$$F^*_{\text{bias}}(s) = -F(s) - \frac{1}{\beta}\log h(s) \quad (8)$$

extremises the functional ϕ. In particular, one can define an optimization problem, where the objective function is given by ϕ, and the solution is given by F^*_{bias}. In Ref. 22, a formula is given that expresses the functional ϕ as the difference of two Kullback-Leibler divergences.

To solve the optimization problem, Valsson and Parrinello used the fact that the functional ϕ is midpoint convex in order to find F^*_{bias}, by parametrizing the biasing function and using an optimization scheme to find the optimal parameters. The parametrization is done by an expansion of F_{bias} into a linear set of basis functions, and the optimization is done by stochastic gradient descent.

From the point of view of probability theory and importance sampling, it follows from the two equations above that if the function h is strictly positive everywhere on $\mathcal{S}$, then the target distribution $Z^{-1}\exp(-\beta F(s))ds$ is absolutely continuous with respect to the proposal $h(s)ds$ since $\exp(-\beta F)$ is strictly positive on $\mathcal{S}$. Therefore, by the Radon-Nikodym theorem, the corresponding Radon-Nikodym derivative exists and is given by $(h(s))^{-1}Z^{-1}e^{-\beta F(s)}$ since we have

$$\int_{\mathcal{S}} \psi(s)\frac{1}{h(s)}\frac{e^{-\beta F(s)}}{Z}h(s)ds = \int_{\mathcal{S}} \psi(s)\frac{e^{-\beta F(s)}}{Z}ds,$$

in agreement with (3).

B. Hartmann-Schütte method

In Ref. 15, the Hartmann-Schütte method considered the overdamped Langevin or Brownian dynamics,

$$\dot{s}(t) = -\nabla F(s(t)) + \sqrt{2\beta^{-1}}\dot{\xi}(t), \quad 0 \le t \le \tau. \quad (9)$$

Some applications of the Hartmann-Schütte method are given in Refs. 23–26. In (9), $\xi(t) = (\xi_1(t), \ldots, \xi_s(t))$ is the Brownian motion in $\mathcal{S}$, which implies $\mathbb{E}[\xi_j(t)\xi_i(s)] = \delta(i-j)\delta(t-s)t$. Let μ denote the target probability distribution of solutions of (9) associated with F. The quantity of interest in the Hartmann-Schütte method is the graph of the function

$$g(\beta; s_0) = -\beta^{-1}\log\mathbb{E}_\mu[\exp(-\beta W)|s(0) = s_0]$$

as a function of the initial state $s_0 \in \mathcal{S}$. If W is the work done along the trajectory $s_{0:\tau}$, then $g(\beta; s_0)$ is the logarithm of the exponential work average at the inverse temperature β,

conditioned on the initial distribution being the Dirac distribution at s_0. Hartmann and Schütte considered W to be the path functional that assigns to any solution of (9) the value $W(s_{0:\tau}) = \int_0^\tau \phi_1(s(t))dt + \phi_2(s_\tau)$, for some functions ϕ_1 and ϕ_2 on CV space that are bounded from below. Thus, $g(\cdot; s_0)$ is the conditional cumulant generating function of W.

The Hartmann-Schütte method involves replacing F with $F + F_{\text{bias}}$ in (9) in order to obtain a proposal probability distribution ν on path space determined by F_{bias} and constructing a constrained optimization problem where the objective is given by

$$\begin{aligned}\phi(F_{\text{bias}}) &= \mathbb{E}_\nu[W] + \mathbb{E}_\nu[-\beta^{-1}\log\tfrac{d\mu}{d\nu}] \\ &= \mathbb{E}_\nu[W] + \beta^{-1}D_{KL}(\mu\|\nu),\end{aligned} \tag{10}$$

where $D_{KL}(\mu\|\nu)$ is the Kullback-Leibler divergence given in (4). Under the stated conditions, the objective ϕ is strictly convex[27] so that there exists a unique F^*_{bias} that minimises ϕ. The connection between the constrained optimization problem and variance minimization is as follows: for any initial condition s_0, $W(s_{0:\tau})$ is a zero-variance estimator of $g(\beta; s_0)$ whenever $s_{0:\tau}$ solves (9) with F replaced by $F + F^*_{\text{bias}}$.[15]

Like the Valsson-Parrinello method, the Hartmann-Schütte method finds the best approximation to F^*_{bias} within the set of all linear combinations of a predefined collection of basis functions; it uses stochastic gradient descent to find the best parametrization. Since the basis functions are assumed to be finite on their domains, μ is absolutely continuous with respect to ν, and the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ in (10) exists and is well defined. Thus the Hartmann-Schütte method is an importance sampling method, in the sense that it searches from a suitable class of proposal distributions for an optimal, variance-minimizing one.

C. Comparing the methods

As described in Secs. III A and III B, the Valsson-Parrinello and Hartmann-Schütte methods share some common features. Both methods

1. use a proposal distribution—obtained by adding a biasing function to some energy landscape—in order to estimate an expected value with respect to the target distribution,
2. use the correspondence between the proposal distribution and the energy function in order to set up convex optimization problems defined on the space of biasing functions,
3. search for the best biasing function in a parametric class of biasing functions, and
4. are related to the minimization of the Kullback-Leibler divergence of the target distribution with respect to the proposal distribution.

We argue that these similarities are natural to enhanced sampling methods. The first common feature is a significant component of importance sampling, as we showed in Sec. II. The second common feature is relevant because it is well known that convex optimization problems are easier to solve than nonconvex ones. The third common feature makes the convex optimization problem easier to solve computationally. We shall discuss the fourth common feature in Sec. IV A.

We also note that in Ref. 15, Hartmann and Schütte suggested that the biasing function be a sum of suitably scaled Gaussians that are chosen to "fill" the basins in the energy landscape that are associated with metastable conformations. This is fundamentally the same idea as that of metadynamics.[5] Several different aspects of the Hartmann-Schütte approach have been studied, e.g., the question of whether Gaussians are a good choice of ansatz function,[24] a method for placing such ansatz functions automatically,[28] and the convergence of the gradient descent approach.[27] Analyses of similar questions have also been done for metadynamics.[29–31]

With regards to the connection between importance sampling and enhanced sampling, the two methods are similar in the following way. In traditional MD, one can perform reweighting using standard umbrella sampling and importance sampling techniques to obtain the free energy landscape for arbitrary collections of CVs, regardless of whether or not the CVs in these collections are involved in the biased dynamics. A significant feature of the Valsson-Parrinello method is that one can use the optimal biasing term from (8) in order to directly obtain the free energy landscape using the Valsson-Parrinello method, without a need for reweighting. Likewise, the Hartmann-Schütte method does not perform statistical reweighting in order to calculate the landscape of the function g over $\mathcal{S}$ for a fixed β, although one can perform statistical reweighting using Girsanov's theorem (see Sec. III D) for any path functional W', even if W' differs from the functional W that defines the quantity of interest g.

The Hartmann-Schütte and Valsson-Parrinello methods also differ in certain ways, the most prominent difference being that the quantity of interest in the Valsson-Parrinello method is the free energy landscape over some CV space $\mathcal{S}$, while the quantity of interest in the Hartmann-Schütte method is the conditional cumulant generating function g of some path functional W.

D. Girsanov's theorem

Recall that one of the stated aims of this article is to present some basic mathematical concepts that are frequently used in mathematical analyses of importance sampling methods. In this section, we present Girsanov's theorem, a mathematical result that is often used for statistical reweighting of path functionals, whenever these paths are solutions of the overdamped Langevin dynamics equation (9). Girsanov's theorem provides an explicit formula for the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ of the target probability distribution μ on path space with respect to the proposal ν in terms of the biasing function F_{bias}.

Theorem 2 (Girsanov). *Let μ be the probability distribution of solutions of (9) and ν be the probability distribution of solutions of (9) for F replaced by $F + F_{bias}$. If F_{bias} is finite on $\mathcal{S}$, then the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ on the space of paths of duration τ satisfies*

$$\frac{d\mu}{d\nu} = \exp\left(\sqrt{\frac{\beta}{2}}\int_0^\tau F_{bias}(s(t))\cdot d\xi(t) - \frac{\beta}{4}\int_0^\tau |F_{bias}(s(t))|^2\,dt\right). \tag{11}$$

For a complete statement and proof of Girsanov's theorem, see Theorem 7.2 of Ref. 32. Girsanov's theorem is particularly useful for two reasons. First, we can use it to interpret the Kullback-Leibler divergence term on the right-hand side of (10): given the properties of the Brownian motion process ξ, we have

$$\mathbb{E}_\nu\left[-\beta^{-1}\log\frac{d\mu}{d\nu}\right] = \mathbb{E}_\nu\left[\frac{1}{4}\int_0^\tau |F_{\text{bias}}(s(t))|^2 dt\right]$$

so that the Kullback-Leibler divergence is proportional to the average energy dissipated along the overdamped Langevin trajectories $s_{0:\tau}$. Second, the Radon-Nikodym derivative $\frac{d\mu}{d\nu}$ can be computed on the fly along any one trajectory because it is expressed in terms of quantities that are computed at every step in the dynamics. We note that, while Girsanov's formula (11) may appear similar to certain expressions in stochastic path integral hyperdynamics[33,34] [see, e.g., Eqs. (5) and (6) in Ref. 34], the latter expressions differ from (11) because they do not contain a stochastic integral term of the form $\int_0^\tau F(s(t))\cdot d\xi(t)$. Thus statistical reweighting based on Girsanov's theorem and the reweighting described in path integral hyperdynamics are different. A statistical reweighting scheme based on Girsanov's theorem has only very recently been applied in the MD context, e.g., for Markov state models.[17]

IV. DISCUSSION

In this section, we show the significance of the connections between the class of enhanced sampling methods that we consider in this paper and importance sampling methods, by presenting some results from the mathematical analysis of Monte Carlo and importance sampling methods and describing their significance to the development of enhanced sampling methods.

A. Bounds on error in terms of Kullback-Leibler divergence

Chatterjee and Diaconis[35] considered the question of the sample size n required in order to obtain a good estimate of some quantity of interest $\mathbb{E}_\mu[f]$, under the assumption that $\mathbb{E}_\mu[f^2]$ is finite. The first case they consider is as follows: Let μ and ν be probability distributions on some set $\mathcal{X}$, and suppose that μ is absolutely continuous to ν, so that the Kullback-Leibler divergence $D_{KL}(\mu\|\nu)$ of μ with respect to ν given by (4) exists and is finite. Chatterjee and Diaconis showed that, for the importance sampling estimator $I_N[f]$ defined by (5), the number of samples required such that $I_N[f]$ is close to the quantity of interest with high probability is approximately equal to $\exp(D_{KL}(\mu\|\nu))$. In particular, if $s\in\mathbb{R}$ is such that the standard deviation of $\log\frac{d\mu}{d\nu}$ about its mean is $O(10^s)$, then a sample of size $\exp(D_{KL}(\mu\|\nu)+O(10^s))$ is sufficient, and a sample of size $\exp(D_{KL}(\mu\|\nu)-O(10^s))$ is necessary, for the error of the importance sampling estimator to be close to zero with high probability.

The preceding result applies to the additive error $|\mathbb{E}_\mu[f]-I_N[f]|$ and hence is not useful when the quantity of interest itself is small, e.g., when $\mathbb{E}_\mu[f]=\mu(A)$ is a rare event probability. In the latter case, the multiplicative error $I_N(f)/\mu(A)$ is more appropriate, and the analogous result asserts that the required sample size is exponential in the Kullback-Leibler divergence $D_{KL}(\mu_A\|\nu)$, where μ_A denotes the target measure μ conditioned on the event A.

The results of the work by Chatterjee and Diaconis show that the sample size needed for accurate importance sampling scales exponentially in the Kullback-Leibler divergence, provided that the Kullback-Leibler divergence term dominates the fluctuations in the logarithm of the Radon-Nikodym derivative. Their result also provides theoretical support to the convex optimization problems of the Valsson-Parrinello and Hartmann-Schütte methods since these convex optimization problems can be reformulated as problems of minimizing the Kullback-Leibler divergence of the target with respect to the proposal.

A caveat that Chatterjee and Diaconis noted is that, although one can use a sample mean estimate of the Kullback-Leibler divergence as a diagnostic for convergence, there is a fundamental flaw in using sample mean estimates of the Kullback-Leibler divergence in doing so. They suggest another scalar quantity as a diagnostic for convergence but also provide an example in which this diagnostic fails. Their work suggests that any convergence diagnostic will have a weakness that renders it uninformative under certain conditions.

We note that, in general, one might not have *a priori* information on the order of magnitude of the fluctuations in $\log\frac{d\mu}{d\nu}$, as is required for the result of the work by Chatterjee and Diaconis. Fortunately, their analysis is complemented by the analysis[36] of Agapiou *et al.*, which shows that, under the assumption of bounded random variables, the sample size required for a wide class of importance sampling methods to yield an accurate estimate scales linearly with the χ^2 divergence,

$$D_{\chi^2}(\mu\|\nu) := \mathbb{E}_\nu\left[\left(\frac{d\mu}{d\nu}-1\right)^2\right].$$

Agapiou *et al.* showed that the worst-case bias and mean-square error in a particle approximation of the target distribution with N particles for bounded quantities of interest are bounded from above by a constant times $N^{-1}(D_{\chi^2}(\mu^N\|\nu)+1)$. The analysis of Agapiou *et al.* confirms that the exponential scaling in the Kullback-Leibler divergence is correct because the χ^2-divergence is related to the Kullback-Leibler divergence according to

$$\exp\left(D_{KL}(\mu\|\nu)\right) \le D_{\chi^2}(\mu\|\nu)+1.$$

Thus, the fact that the required sample size scales linearly with respect to the χ^2 divergence is consistent with the results of the work by Chatterjee and Diaconis that require the required sample size to scale exponentially with respect to the Kullback-Leibler divergence. While the assumption of boundedness mentioned above is rigorously met for some quantities of interest, e.g., rare event probabilities, it might not be for others, e.g., first passage times. Therefore, as in the case of the Chatterjee-Diaconis result, the theoretical results of Agapiou *et al.* must be applied with care.

B. The curse of dimension via concentration of measure

Many methods for enhanced sampling are built around the assumption that one has identified a low-dimensional collective variable space that accurately captures much of the dynamics in the full state space. Using the connection between enhanced sampling and importance sampling methods in this article, we can shed some light on this assumption.

Recent work[37] of Polyak and Shcherbakov showed that Monte Carlo methods fail for simple high-dimensional optimization problems involving linear objective functions defined over regular domains such as hypercubes and balls. They show that this failure arises due to the *concentration of measure* phenomenon. Roughly speaking, the concentration of measure phenomenon refers to the situation when probabilities of fixed sets in a space of dimension n decrease exponentially with the dimension.

The intuition for the results of the work by Polyak and Shcherbakov is that, as the dimension of the space increases, the probability of the set containing the solution of the optimization problem decreases exponentially so that the probability that a Monte Carlo method will draw samples from this set becomes extremely small. Polyak and Shcherbakov demonstrated that this phenomenon occurs even if one does not perform random sampling but uses instead quasirandom sampling, e.g., Sobol sequences. For more sophisticated particle filter-based importance sampling schemes, Bengtsson, Bickel, and Li showed[38] a decade earlier that the concentration of measure phenomenon is responsible for the failure of such importance sampling methods in the context of numerical weather prediction, where the dimension of the state space in question is frequently on the order of 10^6 or higher.

We note that the work of Polyak and Shcherbakov is significant in light of the suggestion of Chatterjee and Diaconis that importance sampling methods can be designed with a view to minimizing the Kullback-Leibler divergence. In the context of enhanced sampling methods, the work of Polyak and Shcherbakov suggests that if one formulates an enhanced sampling method that seeks to solve a high-dimensional optimization problem where the objective function involves the Kullback-Leibler divergence, then the probability that a naive Monte Carlo method will yield a good approximation of the optimizer decreases exponentially with the dimension of the constraint set.

Given that it is ideal to search for optimal biasing functions in a structured manner and given that such structured search methods are built around the idea that an optimal biasing function solves some constrained optimization problem, the work of Polyak and Shcherbakov provides rigorous support for the notion that dimension reduction plays an essential role in enhanced sampling.

C. The active subspaces method

In order to emphasise that the connection between importance sampling methods (more broadly, Monte Carlo methods) and enhanced sampling methods is significant in the context of molecular dynamics, we now briefly describe the method of active subspaces,[39] which has gained traction in a wide range of applications; we refer the interested reader to activesubspaces.org/applications.

Active subspaces were developed for the purpose of studying the dependence of a function defined on a high-dimensional space on its parameters. In such studies, a common goal is to find those parameters which have the largest influence on the values on the function or in other words, to perform sensitivity analysis of the output of the function with respect to its inputs.

The main idea of active subspaces is to replace an experiment whose computational cost renders it unfeasible, with surrogate experiments whose computational cost is sufficiently low as to guarantee feasibility. The idea is to encode the derivatives of the input-to-output function into a symmetric, positive semidefinite matrix and then to study the eigenvalue-eigenvector pairs of this matrix. Relevant parameters—i.e., those inputs for which the corresponding derivatives assume large values—will be encoded by the dominant eigenvectors of this matrix. If the input-to-output map depends only on a few relevant parameters, then there will be a spectral gap in the matrix encoding the derivatives, and thus, the number of dominant eigenvectors will be much smaller than the dimension of the space over which the function is defined. An active subspace method uses Monte Carlo to approximate the eigenpairs and approximates the otherwise prohibitively expensive experiment by investing computational effort to perform expensive experiments on the low-dimensional span of the dominant eigenvectors and using cheaper simulation techniques to sample from the higher-dimensional subspace spanned by the non-dominant eigenvectors.

To the best of our knowledge, active subspace methods have not been applied in the context of molecular dynamics. The method based on spectral gap optimization of order parameters (SGOOP)[40] and a method based on time-lagged independent component analysis (TICA)[41] tackled a similar problem but in the context of Markov state models. These methods find a spectral gap in certain transition probability matrices in order to identify a small number of "slow" variables. Although the inherently linear nature of the active subspaces method may appear to limit its applicability to molecular dynamics (since many collective variables are nonlinear functions of the state), we believe that active subspaces may be useful. Indeed, the successful application of active subspace methods to a wide range of applications in engineering indicates that the problem of finding suitable "dominant" variables remains a significant challenge in many areas of applied mathematics and suggests that there is scope for cooperation between mathematicians and chemical physicists.

ACKNOWLEDGMENTS

The research of J.Q. is funded by the German Research Foundation (Deutsche Forschungsgemeinschaft) through Grant No. CRC 1114 "Scaling Cascades in Complex Systems," Project No. A05 "Probing scales in equilibrated systems by optimal nonequilibrium forcing." H.C.L. is supported by the Free University of Berlin within the Excellence Initiative of the German Research Foundation (DFG). This article is based upon the work partially supported by the National Science Foundation (NSF) under Grant No. DMS-1127914

to the Statistical and Applied Mathematical Sciences Institute (SAMSI). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the NSF or SAMSI.

[1]R. Bernardi, M. Melo, and K. Schulten, Biochim. Biophys. Acta **1850**, 872 (2015), recent developments of molecular dynamics.
[2]C. Jarzynski, Phys. Rev. Lett. **78**, 2690 (1997).
[3]G. E. Crooks, Phys. Rev. E **60**, 2721 (1999).
[4]C. Jarzynski, Phys. Rev. E **73**, 046105 (2006).
[5]A. Laio and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **99**, 12562 (2002).
[6]G. Torrie and J. Valleau, J. Comput. Phys. **23**, 187 (1977).
[7]R. Swendsen and J. Wang, Phys. Rev. Lett. **57**, 2607 (1986).
[8]U. H. Hansmann, Chem. Phys. Lett. **281**, 140 (1997).
[9]A. F. Voter, J. Chem. Phys. **106**, 4665 (1997).
[10]A. F. Voter, Phys. Rev. **57**, R13985 (1998).
[11]M. R. Sorensen and A. F. Voter, J. Chem. Phys. **112**, 9599 (2000).
[12]D. Roe, C. Bergonzo, and T. Cheatham, J. Phys. Chem. B **118**, 3543 (2014).
[13]S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, Science **220**, 671 (1983).
[14]T. Leliévre, Eur. Phys. J. Spec. Top. **224**, 2429 (2015).
[15]C. Hartmann and C. Schütte, J. Stat. Mech. Theor. Exp. **2012**, P11004.
[16]O. Valsson and M. Parrinello, Phys. Rev. Lett. **113**, 090601 (2014).
[17]L. Donati, C. Hartmann, and B. G. Keller, J. Chem. Phys. **146**, 244112 (2017).
[18]P. Tiwary and M. Parrinello, Phys. Rev. Lett. **111**, 230602 (2013).
[19]J. McCarty, O. Valsson, P. Tiwary, and M. Parrinello, Phys. Rev. Lett. **115**, 070601 (2015).
[20]P. Piaggi, O. Valsson, and M. Parrinello, Faraday Discuss. **195**, 557 (2016).
[21]P. Shaffer, O. Valsson, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **113**, 1150 (2016).
[22]M. Invernizzi, O. Valsson, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **114**, 3370 (2017).
[23]H. Wang, C. Hartmann, and C. Schütte, Mol. Phys. **111**, 3555 (2013).
[24]W. Zhang, H. Wang, C. Hartmann, M. Weber, and C. Schütte, SIAM J. Sci. Comput. **36**, A2654 (2014).
[25]C. Hartmann, J. C. Latorre, W. Zhang, and G. A. Pavliotis, J. Comput. Dyn. **1**, 279 (2014).
[26]C. Hartmann, C. Schütte, M. Weber, and W. Zhang, Probab. Theory Related Fields **29**, 1 (2017).
[27]H. C. Lie, "Convexity of a stochastic control functional related to importance sampling of Itô diffusions," e-print arXiv:1603.05900 (2016).
[28]J. Quer, L. Donati, K. Keller, and M. Weber, "An automatic adaptive importance sampling algorithm for molecular dynamics in reaction coordinates," ZIB-Report 17–09, Zuse Institut Berlin, 2017.
[29]J. F. Dama, M. Parrinello, and G. A. Voth, Phys. Rev. Lett. **112**, 240602 (2014).
[30]A. Barducci, G. Bussi, and M. Parrinello, Phys. Rev. Lett. **100**, 020603 (2008).
[31]G. A. Tribello, M. Ceriotti, and M. Parrinello, Proc. Natl. Acad. Sci. U. S. A. **107**, 17509 (2010), http://www.pnas.org/content/107/41/17509.full.pdf.
[32]R. Robert Liptser and A. Shiryaev, *Statistics of Random Processes: I. General Theory, Stochastic Modelling and Applied Probability* (Springer-Verlag Berlin Heidelberg, Berlin, Heidelberg, 2001).
[33]L. Y. Chen and N. J. M. Horing, J. Chem. Phys. **126**, 224103 (2007).
[34]T. Ikonen, M. D. Khandkar, L. Y. Chen, S. C. Ying, and T. Ala-Nissila, Phys. Rev. E **84**, 026703 (2011).
[35]S. Chatterjee and P. Diaconis, "The sample size required in importance sampling," Ann. Appl. Probab. e-print arXiv:1511.01437 (2017).
[36]S. Agapiou, O. Papaspiliopoulos, D. Sanz-Alonso, and A. M. Stuart, Stat. Sci. **32**, 405 (2017).
[37]B. Polyak and P. Shcherbakov, J. Optim. Theory Appl. **173**, 612 (2017).
[38]T. Bengtsson, P. Bickel, and B. Li, "Curse-of-dimensionality revisited: Collapse of the particle filter in very large scale systems," in *Probability and Statistics: Essays in Honor of David A. Freedman, Collections*, edited by D. Nolan and T. Speed (Institute of Mathematical Statistics, Beachwood, Ohio, USA, 2008), Vol. 2, pp. 316–334.
[39]P. G. Constantine, *Active Subspaces: Emerging Ideas for Dimension Reduction in Parameter Studies* (Society for Industrial and Applied Mathematics, Philadelphia, PA, USA, 2015).
[40]P. Tiwary and B. J. Berne, Proc. Natl. Acad. Sci. U. S. A. **113**, 2839 (2016).
[41]G. Pérez-Hernández, F. Paul, T. Giorgino, G. De Fabritiis, and F. Noé, J. Chem. Phys. **139**, 015102 (2013).