Reduced variance analysis of molecular dynamics simulations by linear combination of estimators

S. W Coles, E. Mangaud, D. Frenkel, B. Rotenberg

Reduced variance analysis of molecular dynamics simulations by linear combination of estimators

S. W. Coles,[1, 2] E. Mangaud,[3] D. Frenkel,[4] and B. Rotenberg[5, a)]

[1)]*Department of Chemistry, University of Bath, Claverton Down BA2 7AY, United Kingdom*
[2)]*The Faraday Institution, Quad One, Harwell Science and Innovation Campus, Didcot OX11 0RA, United Kingdom*
[3)]*MSME, Université Gustave Eiffel, UPEC, CNRS, F-77454 Marne-la-Vallée, France*
[4)]*Yusuf Hamied Department of Chemistry, University of Cambridge, Cambridge CB2 1EW, United Kingdom*
[5)]*Physicochimie des électrolytes et Nanosystèmes Interfaciaux, Sorbonne Université, CNRS, F-75005 Paris, France*

(Dated: April 22, 2021)

Building upon recent developments of force-based estimators with a reduced variance for the computation of densities, radial distribution functions or local transport properties from molecular simulations, we show that the variance can be further reduced by considering optimal linear combinations of such estimators. This control variates approach, well known in Statistics and already used in other branches of computational Physics, has been comparatively much less exploited in molecular simulations. We illustrate this idea on the radial distribution function and the one-dimensional density of a bulk and confined Lennard-Jones fluid, where the optimal combination of estimators is determined for each distance or position, respectively. In addition to reducing the variance everywhere at virtually no additional cost, this approach cures an artefact of the initial force-based estimators, namely small but non-zero values of the quantities in regions where they should vanish. Beyond the examples considered here, the present work highlights more generally the underexplored potential of control variates to estimate observables from molecular simulations.

The purpose of particle-based atomistic or mesoscopic simulations is to sample phase space in order to compute observables as ensemble averages, with the ultimate goals of predicting physical properties and understanding their microscopic origin. In parallel to the development of more accurate models to describe real systems and of more efficient algorithms and of more powerful computers to generate more data, computational physicists, chemists and biologists always benefitted from advanced statistical tools to design their simulations and analyze their data. Beyond basic results of estimation theory to compute averages and uncertainties, more advanced approaches such as importance sampling have now become standards. While variance-reduction techniques have been around since the early days of molecular simulations, their full potential seems to have been under-explored in this field compared to others (*e.g.* finance) or to other branches of computational Physics such as Quantum Monte Carlo[1–3], or kinetic theory[4–8].

Recent promising examples of such endeavours include new strategies for the determination of generic observables[9–11] and transport properties[12–14], as well as force-based estimators to sample local properties such as number, charge and polarization densities, radial distribution functions (RDF) or local transport properties with a reduced variance[10,11,15–21]. The availability of different estimators for the same observable opens the possibility of exploiting another well known approach to further reduce the variance, namely the control variates method[22–30]. Here we show the potential of combining of estimators for the determination of RDF and one-dimensional density profiles of a bulk and confined fluid, de-

fined (for a one-component system of N particles in a volume V) respectively as the ensemble averages:

$$g(r) = \frac{V}{N^2} \frac{1}{4\pi r^2} \left\langle \sum_i \sum_{j \neq i} \delta(r_{ij} - r) \right\rangle \quad (1)$$

with r_{ij} the distance between two particles i and j, and

$$\rho(z) = \frac{1}{S} \left\langle \sum_i \delta(z_i - z) \right\rangle \quad (2)$$

with z_i the position of particle i along the z-axis and S the area of the system in the lateral direction. Before turning to the specific estimators for these observables, let us first introduce the generic idea of combining two estimators $E_0(x)$ and $E_1(x)$ of the same property, which in the present case explicitly depends on some parameter x (distance or position). For each value of x, one can consider linear combinations of the form:

$$E_\lambda(x) = (1 - \lambda)E_0(x) + \lambda E_1(x) = E_0(x) + \lambda \Delta(x) \quad (3)$$

with $\Delta(x) = E_1(x) - E_0(x)$. For any value of λ, the expectation value $\langle E_\lambda(x) \rangle$ is the same as both $E_0(x)$ and $E_1(x)$; $E_\lambda(x)$ is therefore another valid estimator of the same property. However, the variance $\text{var}(E_\lambda(x)) = \left\langle [E_\lambda(x) - \langle E_\lambda(x) \rangle]^2 \right\rangle$ is a quadratic function of λ. Importantly, one can find for each x the combination that minimizes the variance:

$$\lambda^*(x) = -\frac{\text{cov}(E_0(x), \Delta(x))}{\text{var}(\Delta(x))} = 1 - \frac{\text{cov}(E_1(x), \Delta(x))}{\text{var}(\Delta(x))}, \quad (4)$$

which involves covariances $\text{cov}(A, B) = \langle (A - \langle A \rangle)(B - \langle B \rangle) \rangle$. This provides for each x the optimal estimator $E_\lambda^*(x)$ in this

————————
[a)]Electronic mail: benjamin.rotenberg@sorbonne-universite.fr

family, which has a reduced variance compared to both $E_0(x)$ and $E_1(x)$:

$$\text{var}(E_\lambda^*(x)) = \left[1 - \rho_{E_n,\Delta}^2(x)\right]\text{var}(E_n(x)) \quad (5)$$

with $\rho_{E_n,\Delta}(x) = \text{cov}(E_n,\Delta)/\sqrt{\text{var}(E_n)\,\text{var}(\Delta)}$ the correlation coefficient between $E_n(x)$ and $\Delta(x) = E_1(x) - E_0(x)$. In practice, given two estimators and a set of configurations, one simply computes for each x both expectation values $\langle E_n(x)\rangle$ as well as the (co-)variances entering in Eq. 4 to obtain the corresponding optimal estimator $E_\lambda^*(x)$ (see Eq. 3).

We now introduce pairs of estimators for the RDF (Eq. 1) and the one-dimensional density distribution (Eq. 2). In each case, both estimators use the force acting on the atoms in addition to their position and correspond to an "integral of the gradient", but they differ in the choice of the origin to perform this integration. For the RDF, we first consider the estimator introduced in Ref. 16, which uses the limit value of 1 for $r \to \infty$, $g_\infty(r) = \langle \hat{g}_{\lambda=0}(r)\rangle$ with

$$\hat{g}_{\lambda=0}(r) = 1 + \frac{V}{N^2}\frac{\beta}{4\pi}\sum_i\sum_{j\neq i}\frac{(\mathbf{f}_j - \mathbf{f}_i)}{2}\cdot\frac{\mathbf{r}_{ij}}{r_{ij}^3}H(r_{ij} - r)\,, \quad (6)$$

where $\mathbf{f}_i$ and $\mathbf{f}_j$ are the forces acting on atoms i and j, $\mathbf{r}_{ij}$ is the displacement between atoms i and j, $\beta = 1/k_B T$ with k_B the Boltzmann constant and T the temperature, and H is the Heaviside function. Compared to the standard histogram approach, obtained by discretizing the definition Eq. 1, in this force-based estimator all pairs separated by a distance larger than r contribute to the estimate of the RDF at r, which reduces the variance considerably. In addition, no bins are necessary and the RDF can be obtained with arbitrary resolution. However, Eq. 6 leads to a small yet spurious non-vanishing value (and non-zero variance) for $r \to 0$. An alternative expression was proposed in Ref. 11 using this condition inside the core and we therefore also consider $g_0(r) = \langle \hat{g}_{\lambda=1}(r)\rangle$, with

$$\hat{g}_{\lambda=1}(r) = \frac{V}{N^2}\frac{\beta}{4\pi}\sum_i\sum_{j\neq i}\frac{(\mathbf{f}_j - \mathbf{f}_i)}{2}\cdot\frac{\mathbf{r}_{ij}}{r_{ij}^3}H(r - r_{ij})\,, \quad (7)$$

where this time all pairs separated by a distance smaller than r contribute to the estimate of the RDF at r. Symmetrically to $g_\infty(r)$, $g_0(r)$ displays by construction a vanishing value with zero variance inside the core, but a larger variance at large distance. Note that the subscripts in g_∞ and g_0 refer to the distance from which the gradient is integrated, not to a value of the mixing parameter λ.

The gradient of the 1D-density is given by the force density $\langle\frac{1}{S}\sum_i\delta(z - z_i)\beta f_{z,i}\rangle$. For a fluid confined between walls, the density can therefore be obtained by integrating the force density from either side, starting from 0 inside the walls. This leads to two complementary estimators defined by:

$$\rho_0(z) = \langle\hat{\rho}_{\lambda=0}(z)\rangle = \left\langle\frac{1}{S}\sum_i H(z - z_i)\beta f_{z,i}\right\rangle\,, \quad (8)$$

and

$$\rho_L(z) = \langle\hat{\rho}_{\lambda=1}(z)\rangle = \left\langle\frac{1}{S}\sum_i H(z_i - z)\beta f_{z,i}\right\rangle\,. \quad (9)$$

Both estimators identically vanish (and so does the corresponding variance) on one side but have a spurious non-zero value (and corresponding variance) on the other.

We illustrate the advantage of combining estimators for a simple Lennard-Jones (LJ) fluid at a reduced density $\rho^* = \rho\sigma^3 = 0.8$ and reduced temperature $T^* = k_B T/\varepsilon = 1.35$, with σ and ε the LJ diameter and energy, respectively. For the RDF we consider a bulk system with $N = 864$ particles, while for the 1D-density we simulate a fluid with $N = 1152$ particles confined between two walls, consisting each of 72 LJ particles identical to the fluid on a centered square lattice (with lattice spacing $\sqrt{2}\sigma$) and separated by a distance 22σ. We use periodic boundary conditions in 3 and 2 dimensions for the bulk and confined systems, respectively. In both cases we use a time step of $10^{-3}\,t^*$ with $t^* = \sqrt{m\sigma^2/\varepsilon}$ the LJ time unit and m the mass of the particles, and a Nosé-Hoover thermostat with a time constant of $0.1\,t^*$. In order to estimate the relevant averages, (co-)variances and resulting optimal combination parameter λ^*, we use 10^3 configurations separated by t^* in both the bulk and confined cases. Even though the force-based estimators do not require bins and can be evaluated at arbitrary positions, the results are shown for evenly spaced distances and positions with $\Delta r = 0.005\sigma$ and $\Delta z = 0.005\sigma$ for the RDF and 1D-density, respectively.

Results for the RDF and the 1D-density profile are shown in Figures 1 and 2, respectively. In both cases all estimators provide the same results for the observables (panels 1a and 2a), which are typical of bulk and interfacial fluids. However the variances (panels 1b and 2b) differ. The variance of the optimal combination is lower than that of both initial estimators, as expected. It is also instructive to consider the optimal weight λ^*, shown in panels 1c and 2c. In both cases, the optimal combination coincides with the zero-variance estimator in the corresponding regions: inside the core for g_0 and at large distance for g_∞, for $z \to 0$ and $z \to L$ for ρ_0 and ρ_L, respectively. Beyond these limits, the contributions of the two estimators are equal in the central region for the 1D-density, but their evolution is more complex for the RDF, with oscillations following that of the RDF itself and even a negative region for λ^* near the first peak of $g(r)$.

Beyond the illustration on the RDF and 1D-density, this work highlights the potential of the control variates method to combine estimators of the same quantity in order to obtain new estimators with reduced variance. This approach is particularly beneficial when the regions in which the initial estimators perform well in complementary regions, as it mitigates their respective limitations. The additional cost is limited, since it only requires computing (co-)variances of the initial estimators. The recent development of alternative estimators for molecular simulations, *e.g.* force-based or within the mapped-averaging framework, can directly benefit from the present control variates approach, for other observables such as angular distributions or two-particle densities. One could also consider other combinations, such as direct estimates of response functions (*e.g.* heat capacity or capacitance) with their fluctuation counterparts (*e.g.* energy or charge fluctuations).
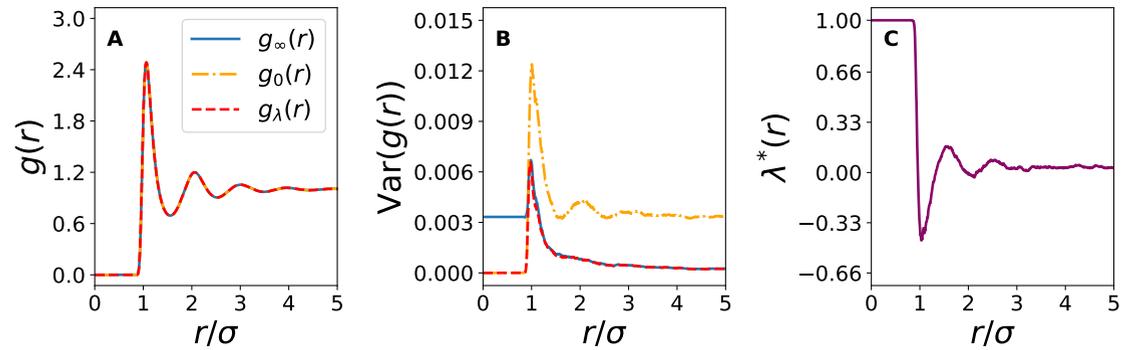
Figure 1. (a) Radial distribution function for a bulk Lennard-Jones fluid with the two estimators g_∞ (see Eq. 6, solid blue line) and g_0 (see Eq. 7, dashed-dotted yellow line), corresponding to $\lambda = 0$ and $\lambda = 1$, respectively, and with the optimal linear combination for each distance (g_λ, see Eqs. 3 and 4, dashed red line). (b) Variance with each estimator. (c) Optimal mixing parameter λ^* as a function of distance. In all panels the distance is normalized by the LJ diameter σ.
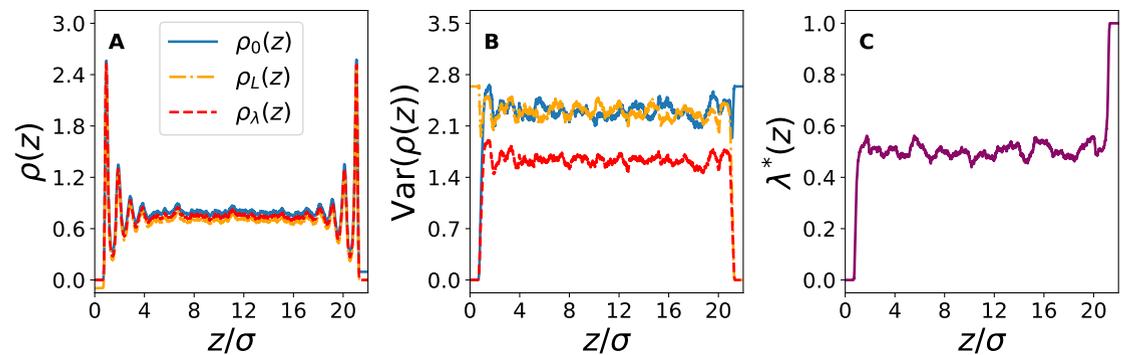

Figure 2. (a) One-dimensional density profile for a confined Lennard-Jones fluid, estimated with the two estimators ρ_0 (see Eq. 8, solid blue line) and ρ_L (see Eq. 9, dashed-dotted yellow line), corresponding to $\lambda = 0$ and $\lambda = 1$, respectively, and with the optimal linear combination for each distance (ρ_λ, see Eqs. 3 and 4, dashed red line). (b) Variance divided by the density with each estimator. (c) Optimal mixing parameter λ^* as a function of position. In all panels the position is normalized by the LJ diameter σ.

DATA AVAILABILITY

The data that support the findings of this study are available from the corresponding author upon reasonable request.

REFERENCES

[1]R. Assaraf and M. Caffarel, "Zero-variance principle for monte carlo algorithms," Physical review letters **83**, 4682 (1999).

[2]R. Assaraf, M. Caffarel, and A. Scemama, "Improved Monte Carlo estimators for the one-body density," Physical Review E **75** (2007).

[3]J. Toulouse, R. Assaraf, and C. J. Umrigar, "Zero-variance zero-bias quantum Monte Carlo estimators of the spherically and system-averaged pair density," The Journal of Chemical Physics **126**, 244112 (2007).

[4]G. A. Radtke and N. G. Hadjiconstantinou, "Variance-reduced particle simulation of the Boltzmann transport equation in the relaxation-time approximation," Physical Review E **79**, 056711 (2009).

[5]G. A. Radtke, J.-P. M. Péraud, and N. G. Hadjiconstantinou, "On efficient simulations of multiscale kinetic transport," Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences **371**, 20120182 (2013).

[6]J.-P. M. Péraud and N. G. Hadjiconstantinou, "An alternative approach to efficient simulation of micro/nanoscale phonon transport," Applied Physics Letters **101**, 153114 (2012).

[7]G. A. Radtke, N. G. Hadjiconstantinou, and W. Wagner, "Low-noise Monte Carlo simulation of the variable hard sphere gas," Physics of Fluids **23**, 030606 (2011).

[8]J.-P. M. Péraud and N. G. Hadjiconstantinou, "Efficient simulation of multidimensional phonon transport using energy-based variance-reduced Monte Carlo formulations," Physical Review B **84**, 205331 (2011).

[9] A. J. Schultz, S. G. Moustafa, W. Lin, S. J. Weinstein, and D. A. Kofke, "Reformulation of Ensemble Averages via Coordinate Mapping," Journal of Chemical Theory and Computation **12**, 1491–1498 (2016).

[10] A. J. Schultz and D. A. Kofke, "Alternatives to conventional ensemble averages for thermodynamic properties," Current Opinion in Chemical Engineering Frontiers of Chemical Engineering: Molecular Modeling, **23**, 70–76 (2019).

[11] A. Purohit, A. J. Schultz, and D. A. Kofke, "Force-sampling methods for density distributions as instances of mapped averaging," Molecular Physics **117**, 2822–2829 (2019).

[12] L. Ercole, A. Marcolongo, and S. Baroni, "Accurate thermal conductivities from optimally short molecular dynamics simulations," Scientific Reports **7**, 1–11 (2017).

[13] S. Baroni, R. Bertossa, L. Ercole, F. Grasselli, and A. Marcolongo, "Heat Transport in Insulators from Ab Initio Green-Kubo Theory," in *Handbook of Materials Modeling: Applications: Current and Emerging Materials*, edited by W. Andreoni and S. Yip (Springer International Publishing, Cham, 2018) pp. 1–36.

[14] R. Bertossa, F. Grasselli, L. Ercole, and S. Baroni, "Theory and Numerical Simulation of Heat Transport in Multicomponent Systems," Physical Review Letters **122**, 255901 (2019).

[15] A. B. Adib and C. Jarzynski, "Unbiased estimators for spatial distribution functions of classical fluids," The Journal of Chemical Physics **122**, 014114 (2005).

[16] D. Borgis, R. Assaraf, B. Rotenberg, and R. Vuillemier, "Computation of pair distribution functions and three-dimensional densities with a reduced variance principle," Molecular Physics **111**, 3486–3492 (2013).

[17] D. de las Heras and M. Schmidt, "Better Than Counting: Density Profiles from Force Sampling," Phys. Rev. Lett. **120**, 218001 (2018).

[18] A. Trokhymchuk, A. J. Schultz, and D. A. Kofke, "Alternative ensemble averages in molecular dynamics simulation of hard spheres," Molecular Physics , 1–20 (2019).

[19] S. W. Coles, D. Borgis, R. Vuilleumier, and B. Rotenberg, "Computing three-dimensional densities from force densities improves statistical efficiency," J. Chem. Phys. **151**, 064124 (2019).

[20] E. Mangaud and B. Rotenberg, "Sampling mobility profiles of confined fluids with equilibrium molecular dynamics simulations," The Journal of Chemical Physics **153**, 044125 (2020).

[21] B. Rotenberg, "Use the force! Reduced variance estimators for densities, radial distribution functions, and local mobilities in molecular simulations," The Journal of Chemical Physics **153**, 150902 (2020).

[22] H. Hammer and H. Tjelmeland, "Control Variates for the Metropolis–Hastings Algorithm," Scandinavian Journal of Statistics **35**, 400–414 (2008).

[23] S. Boyaval and T. Lelièvre, "A variance reduction method for parametrized stochastic differential equations using the reduced basis paradigm," Communications in Mathematical Sciences **8**, 735–762 (2010).

[24] P. Dellaportas and I. Kontoyiannis, "Control variates for estimation based on reversible Markov chain Monte Carlo samplers," Journal of the Royal Statistical Society. Series B (Statistical Methodology) **74**, 133–161 (2012).

[25] A. Mira, R. Solgi, and D. Imparato, "Zero variance Markov chain Monte Carlo for Bayesian estimators," Statistics and Computing **23**, 653–662 (2013).

[26] T. Lelièvre and G. Stoltz, "Partial differential equations and stochastic methods in molecular dynamics," Acta Numerica **25**, 681–880 (2016).

[27] C. J. Oates, M. Girolami, and N. Chopin, "Control functionals for Monte Carlo integration," Journal of the Royal Statistical Society: Series B (Statistical Methodology) **79**, 695–718 (2017).

[28] J. Roussel and G. Stoltz, "A Perturbative Approach to Control Variates in Molecular Dynamics," Multiscale Modeling & Simulation **17**, 552–591 (2019).

[29] T. H. Nguyen and D. D. L. Minh, "Implicit ligand theory for relative binding free energies: II. An estimator based on control variates," Journal of Physics Communications **4**, 115010 (2020).

[30] D. Belomestny, L. Iosipoi, E. Moulines, A. Naumov, and S. Samsonov, "Variance reduction for Markov chains with application to MCMC," Statistics and Computing **30**, 973–997 (2020).