Introduction to the variational and diffusion Monte Carlo methods

Julien Toulouse, Roland Assaraf, C. J. Umrigar

Introduction to the variational and diffusion Monte Carlo methods

Julien Toulouse[1,2], Roland Assaraf[1,2], C. J. Umrigar[3]

[1]Sorbonne Universités, UPMC Univ Paris 06, UMR 7616, Laboratoire de Chimie Théorique, F-75005 Paris, France
[2]CNRS, UMR 7616, Laboratoire de Chimie Théorique, F-75005 Paris, France
[3]Laboratory of Atomic and Solid State Physics, Cornell University, Ithaca, New York 14853, USA

August 10, 2015

Contents

Abstract

We provide a pedagogical introduction to the two main variants of real-space quantum Monte Carlo methods for electronic-structure calculations: variational Monte Carlo (VMC) and diffusion Monte Carlo (DMC). Assuming no prior knowledge on the subject, we review in depth the Metropolis-Hastings algorithm used in VMC for sampling the square of an approximate wave function, discussing details important for applications to electronic systems. We also review in detail the more sophisticated DMC algorithm within the fixed-node approximation, introduced to avoid the infamous Fermionic sign problem, which allows one to sample a more accurate approximation to the ground-state wave function. Throughout this review, we discuss the statistical methods used for evaluating expectation values and statistical uncertainties. In particular, we show how to estimate nonlinear functions of expectation values and their statistical uncertainties.

Keywords: quantum Monte Carlo, electronic-structure calculations, Metropolis-Hastings algorithm, fixed-node approximation, statistical methods.

This chapter provides a pedagogical introduction to the two main variants of real-space quantum Monte Carlo (QMC) methods for electronic-structure calculations: variational Monte Carlo (VMC) and diffusion Monte Carlo (DMC). For more details of these methods, see, e.g., Refs. [1, 2, 3, 4, 5, 6]. For reviews on applications of QMC methods in chemistry and condensed-matter physics, see, e.g., Refs. [7, 8].

1 Variational Monte Carlo

1.1 Basic idea

The idea of the VMC method [9, 10] is simply to calculate the multidimensional integrals appearing in quantum mechanics using a Monte Carlo numerical integration technique[1]. The quantity of greatest interest is the variational energy associated with a Hamiltonian $\hat{H}$ and a wave function Ψ, which can be written as

$$E_{\rm v} = \frac{\langle\Psi|\hat{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle} = \frac{\int {\rm d}\mathbf{R}\,\Psi(\mathbf{R})^2 E_{\rm L}(\mathbf{R})}{\int {\rm d}\mathbf{R}\,\Psi(\mathbf{R})^2} = \int {\rm d}\mathbf{R}\,\rho(\mathbf{R}) E_{\rm L}(\mathbf{R}), \tag{1}$$

where $E_{\rm L}(\mathbf{R}) = (H\Psi(\mathbf{R}))/\Psi(\mathbf{R})$ is the *local energy* depending on the $3N$ coordinates $\mathbf{R}$ of the N electrons, and $\rho(\mathbf{R}) = \Psi(\mathbf{R})^2/\int {\rm d}\mathbf{R}\Psi(\mathbf{R})^2$ is the normalized probability density. For simplicity of notation we have assumed that $\Psi(\mathbf{R})$ is real valued; the extension to complex $\Psi(\mathbf{R})$ is straightforward. The variational energy can be estimated as the average value of $E_{\rm L}(\mathbf{R})$ on a sample of M points $\mathbf{R}_k$ sampled from the probability density $\rho(\mathbf{R})$,

$$E_{\rm v} \approx \overline{E}_{\rm L} = \frac{1}{M}\sum_{k=1}^{M} E_{\rm L}(\mathbf{R}_k). \tag{2}$$

In practice, the points $\mathbf{R}_k$ are sampled using the Metropolis-Hastings algorithm [12, 13].

The advantage of this approach is that it does not use an analytical integration involving the wave function, and thus does not impose severe constraints on the form of the wave function. The wave functions usually used in QMC are of the Jastrow-Slater form,

$$\Psi(\mathbf{R}) = J(\mathbf{R})\Phi(\mathbf{R}), \tag{3}$$

where $J(\mathbf{R})$ is a Jastrow factor and $\Phi(\mathbf{R})$ is a Slater determinant or a linear combination of Slater determinants[2]. The Jastrow factor is generally of the form $J(\mathbf{R}) = e^{f(\mathbf{R})}$. It depends explicitly on the interparticle distances r_{ij}, allowing for an efficient description of the so-called electron "dynamic" correlation.

In practice, the VMC method has two types of errors:

- a *systematic error*, due to the use of an approximate wave function (as in other wave-function methods),

[1] To the best of our knowledge, the first calculation of multidimensional integrals appearing in quantum mechanics by using Monte Carlo methods was done by Conroy [11].

[2] In QMC, it is convenient to use wave functions in which the values of the spin coordinates have been fixed, so Ψ is a function of the spatial coordinates $\mathbf{R}$ only.

- a *statistical uncertainty*, due to the sampling of finite size M (which is specific to Monte Carlo methods).

Of course, the variational energy is an upper bound of the exact ground-state energy, but the systematic error is generally unknown, since its determination requires knowing the exact solution. By contrast, the statistical uncertainty can be easily estimated by the usual statistical techniques. For this, let us examine more closely the meaning of Eq. (2). The average of the local energy $\overline{E}_{\mathrm{L}}$ on a finite sample is itself a random variable, taking different values on different samples. The central limit theorem establishes that, if $E_{\mathrm{L}}(\mathbf{R}_k)$ are random variables that are *independent* (i.e. not correlated) and *identically distributed*, with finite expected value $\mathrm{E}[E_{\mathrm{L}}]$ and finite variance, $\mathrm{V}[E_{\mathrm{L}}] = \mathrm{E}[(E_{\mathrm{L}} - E_{\mathrm{v}})^2]$, then in the large M limit the probability distribution of the random variable $\overline{E}_{\mathrm{L}}$ converges (in the mathematical sense of convergence in distribution) to a Gaussian (or normal) distribution of expected value $\mathrm{E}[E_{\mathrm{L}}]$ and variance $\mathrm{V}[E_{\mathrm{L}}]/M$,

$$\mathrm{E}\left[\overline{E}_{\mathrm{L}}\right] = \mathrm{E}[E_{\mathrm{L}}] = E_{\mathrm{v}}, \tag{4a}$$

$$\mathrm{V}\left[\overline{E}_{\mathrm{L}}\right] = \frac{\mathrm{V}[E_{\mathrm{L}}]}{M}. \tag{4b}$$

This means that $\overline{E}_{\mathrm{L}}$ is an *estimator* of E_{v} with a statistical uncertainty which can be defined by the standard deviation of its Gaussian distribution

$$\sigma\left[\overline{E}_{\mathrm{L}}\right] = \sqrt{\mathrm{V}\left[\overline{E}_{\mathrm{L}}\right]} = \sqrt{\frac{\mathrm{V}[E_{\mathrm{L}}]}{M}}. \tag{5}$$

The meaning of this standard deviation is that the desired expected value E_{v} has a probability of 68.3% of being in the interval $\left[\overline{E}_{\mathrm{L}} - \sigma, \overline{E}_{\mathrm{L}} + \sigma\right]$, a probability of 95.5% of being in the interval $\left[\overline{E}_{\mathrm{L}} - 2\sigma, \overline{E}_{\mathrm{L}} + 2\sigma\right]$, and a probability of 99.7% of being in the interval $\left[\overline{E}_{\mathrm{L}} - 3\sigma, \overline{E}_{\mathrm{L}} + 3\sigma\right]$. Note that, if the variance $\mathrm{V}[E_{\mathrm{L}}]$ is infinite but the expected value $\mathrm{E}[E_{\mathrm{L}}]$ is finite, then the law of large numbers guarantees the convergence of $\overline{E}_{\mathrm{L}}$ to $\mathrm{E}[E_{\mathrm{L}}]$ when $M \to \infty$ but with a statistical uncertainty which is more difficult to estimate and which decreases more slowly than $1/\sqrt{M}$.

It is important to note that the statistical uncertainty decreases as $1/\sqrt{M}$ *independently of the dimension of the problem*. This is in contrast to deterministic numerical integration methods for which the convergence of the integration error deteriorates with the spatial dimension d. For example, Simpson's integration rule converges as $1/M^{(4/d)}$ (provided the integrand has up to 4^{th}-order derivatives), so that for $d > 8$ Monte Carlo methods are more efficient for large M.

The statistical uncertainty is reduced if the variance of the local energy $\mathrm{V}[E_{\mathrm{L}}]$ is small. In the limit that Ψ is an exact eigenfunction of $\hat{H}$, the local energy E_{L} becomes exact, independent of $\mathbf{R}$, and thus its variance $\mathrm{V}[E_{\mathrm{L}}]$ and the statistical uncertainty of $\overline{E}_{\mathrm{L}}$ vanish. This is known as the *zero-variance* property. Since the systematic error (or bias) of the variational energy $\Delta E = E_{\mathrm{v}} - E_0$ (where E_0 is the exact energy) also vanishes in this limit, there is a zero-bias property as well. For these reasons, a great deal of effort has been expended on developing robust and efficient wave-function optimization methods.

1.2 Estimation of the statistical uncertainty

In practice, the probability density $\rho(\mathbf{R})$ is sampled with the Metropolis-Hastings algorithm which provides a sequence of points $\mathbf{R}_k$ correctly distributed according to $\rho(\mathbf{R})$ but *sequentially (or serially) correlated* (i.e. non independent). This is a consequence of each point being

sampled from a probability distribution conditional on the previous point. One can define an *autocorrelation time* (defined more precisely later) that is roughly speaking the average time for points to decorrelate. This sequential correlation must be taken into account when using the central limit theorem for evaluating the statistical uncertainty. This is done using the *blocking* technique, which is described next.

Let us consider a sequence of M realizations X_k (sequentially correlated) of a random variable X of expected value $\mathrm{E}[X]$ and of variance $\mathrm{V}[X]$. For example, X could be the local energy E_L. We divide this sequence into M_b successive blocks of M_s steps each. The *block average* $\overline{X}_\mathrm{b}$ is

$$\overline{X}_\mathrm{b} = \frac{1}{M_\mathrm{s}} \sum_{k=1}^{M_\mathrm{s}} X_k. \tag{6}$$

The expected value of $\overline{X}_\mathrm{b}$ is also the expected value of X, i.e. $E\left[\overline{X}_\mathrm{b}\right] = E[X]$, but its variance is not simply $V[X]/M_\mathrm{s}$ since the variables X_k are not independent. We can now define the *global average* $\overline{X}$ of the whole sample as the average over all the blocks of the block averages

$$\overline{X} = \frac{1}{M_\mathrm{b}} \sum_{b=1}^{M_\mathrm{b}} \overline{X}_b, \tag{7}$$

where $\overline{X}_b$ with a math subscript "b" indicates the block average for the b^{th} block (whereas $\overline{X}_\mathrm{b}$ with a Roman subscript "b" indicates the generic random variable). The global average $\overline{X}$ is another random variable with the same expected value as X, i.e. $\mathrm{E}\left[\overline{X}\right] = \mathrm{E}\left[\overline{X}_\mathrm{b}\right] = \mathrm{E}[X]$. If the length of the blocks is large compared to the autocorrelation time then the block averages $\overline{X}_b$ can be considered as being independent, and the variance of the global average is simply

$$\mathrm{V}\left[\overline{X}\right] = \frac{\mathrm{V}\left[\overline{X}_\mathrm{b}\right]}{M_\mathrm{b}}, \tag{8}$$

which leads to the statistical uncertainty of $\overline{X}$

$$\sigma\left[\overline{X}\right] = \sqrt{\mathrm{V}\left[\overline{X}\right]} = \sqrt{\frac{\mathrm{V}\left[\overline{X}_\mathrm{b}\right]}{M_\mathrm{b}}}. \tag{9}$$

In practice, the statistical uncertainty on a finite sample is calculated as

$$\sigma\left[\overline{X}\right] \approx \sqrt{\frac{1}{M_\mathrm{b}-1}\left(\frac{1}{M_\mathrm{b}}\sum_{b=1}^{M_\mathrm{b}} {\overline{X}_b}^2 - \left(\frac{1}{M_\mathrm{b}}\sum_{b=1}^{M_\mathrm{b}} \overline{X}_b\right)^2\right)}, \tag{10}$$

where the $M_\mathrm{b}-1$ term appearing instead of M_b is necessary to have an unbiased estimator of the standard deviation on the sample (see the appendix). It takes into account the fact that the computed variance is relative to the sample average rather than the true expected value.

Finally, let us examine the variance $\mathrm{V}\left[\overline{X}_\mathrm{b}\right]$. Since the variables X_k are not independent, the expansion of $\mathrm{V}\left[\overline{X}_\mathrm{b}\right]$ involves the covariances between the variables

$$\mathrm{V}\left[\overline{X}_\mathrm{b}\right] = \frac{1}{M_\mathrm{s}^2}\sum_{k,l}\mathrm{Cov}[X_k, X_l] = \frac{\mathrm{V}[X]}{M_\mathrm{s}} + \frac{2}{M_\mathrm{s}^2}\sum_{k<l}\mathrm{Cov}[X_k, X_l] = T_\mathrm{c}\frac{\mathrm{V}[X]}{M_\mathrm{s}}, \tag{11}$$

defining the autocorrelation time of X

$$T_\mathrm{c} = 1 + \frac{2}{\mathrm{V}[X] M_\mathrm{s}} \sum_{k<l} \mathrm{Cov}[X_k, X_l]. \tag{12}$$

The autocorrelation time is equal to 1 in the absence of correlation between the variables, i.e. $\mathrm{Cov}[X_k, X_l] = 0$ for $k \neq l$, but can be large in the presence of sequential correlation. It is instructive to express the statistical uncertainty as a function of T_c

$$\sigma\left[\overline{X}\right] = \sqrt{T_\mathrm{c} \frac{\mathrm{V}[X]}{M_\mathrm{s} M_\mathrm{b}}} = \sqrt{T_\mathrm{c} \frac{\mathrm{V}[X]}{M}}, \tag{13}$$

where $M = M_\mathrm{s} M_\mathrm{b}$ is the total size of the sample. The expression (13) allows one to interpret T_c as a factor giving the number of effectively independent points in the sample, $M_\mathrm{eff} = M/T_\mathrm{c}$. In practice, it is useful to calculate the autocorrelation time as $T_\mathrm{c} = M_\mathrm{s} \mathrm{V}\left[\overline{X}_\mathrm{b}\right] / \mathrm{V}[X]$ and check whether the length of the blocks is large enough for a correct estimation of the statistical uncertainty, e.g. $M_\mathrm{s} > 100\, T_\mathrm{c}$. If M_s is not much greater than T_c, then the statistical uncertainty $\sigma\left[\overline{X}\right]$ and the autocorrelation time T_c will be underestimated.

In the appendix, we further explain how to estimate the statistical uncertainty of nonlinear functions of expectation values, which often occur in practice.

1.3 Calculation cost

The calculation cost required to reach a given statistical uncertainty $\sigma\left[\overline{X}\right]$ is

$$t = t_\mathrm{s} M = t_\mathrm{s} \frac{T_\mathrm{c} \mathrm{V}[X]}{\sigma\left[\overline{X}\right]^2} \tag{14}$$

where t_s is the calculation time per iteration. The $1/\sigma\left[\overline{X}\right]^2$ dependence implies that decreasing the statistical uncertainty by a factor of 10 requires to increase the computational time by a factor of 100. This quadratic dependence directly stems from the central limit theorem and seems unavoidable[3]. However, one can play with the three other parameters:

- T_c depends on the sampling algorithm and on the random variable X. For efficient algorithms such as Umrigar's one [15, 3], the autocorrelation time of the local energy is close to 1 and little further improvement seems possible;
- t_s is usually dominated by the cost of evaluating X. For the local energy, the evaluation cost depends on the form of the wave function;
- $\mathrm{V}[X]$ depends on the choice of the random variable X with its associated probability distribution, the only constraint being that the expected value $\mathrm{E}[X]$ must equal the expectation value of the observable (otherwise, this is a biased estimator). The choice of a good probability distribution is usually called *importance sampling*. Even for a fixed probability distribution, it is possible to use various estimators for X, some of which have

[3] Quasi Monte Carlo methods [14] can in some cases achieve a convergence rate of $\mathcal{O}(\ln(M)/M)$ rather than $\mathcal{O}(1/\sqrt{M})$. However, they have not been used for QMC applications, in part because in QMC the sampled distributions, for systems with more than a few electrons, are very highly peaked.

smaller variance than others, since one has the freedom to add any quantity with zero expectation value. This has been exploited to construct improved estimators for diverse observables [16, 17, 18, 19, 20]. There is often a compromise to be found between a low computation time per iteration t_s and a low variance $V[X]$.

1.4 Sampling technique

The probability density, $\rho(\mathbf{R}) = \Psi(\mathbf{R})^2 / \int d\mathbf{R}\Psi(\mathbf{R})^2$, is generally complicated and cannot be sampled by direct methods such as the transformation method or the rejection method. Instead, the Metropolis-Hastings (or generalized Metropolis) algorithm, which can be used to sample any known probability density, is used. It employs a stochastic process, more specifically, a Markov chain.

Stochastic process

A *stochastic process* represents the evolution – say in "time" – of a random variable. It is described by a trajectory of successive points $\mathbf{R}_1$, $\mathbf{R}_2$, ..., $\mathbf{R}_M$ with an associated probability distribution $P(\mathbf{R}_M, ..., \mathbf{R}_2, \mathbf{R}_1)$. The idea of evolution in time can be made more explicit by decomposing the probability of the whole trajectory in products of the conditional probability of having a particular point knowing that all the previous points have already been realized. For example, for $M = 3$, the probability of the trajectory is

$$P(\mathbf{R}_3, \mathbf{R}_2, \mathbf{R}_1) = P(\mathbf{R}_3|\mathbf{R}_2, \mathbf{R}_1)P(\mathbf{R}_2|\mathbf{R}_1)P(\mathbf{R}_1). \tag{15}$$

Markov chain

A *Markov chain* is a stochastic process for which the conditional probability for the transition to a new point $\mathbf{R}_k$ depends only on the previous point $\mathbf{R}_{k-1}$

$$P(\mathbf{R}_k|\mathbf{R}_{k-1}, ..., \mathbf{R}_1) = P(\mathbf{R}_k|\mathbf{R}_{k-1}), \tag{16}$$

i.e. the process "forgets" the way it arrived at point $\mathbf{R}_{k-1}$. The probability of a trajectory can thus be simply written as, e.g. for $M = 3$,

$$P(\mathbf{R}_3, \mathbf{R}_2, \mathbf{R}_1) = P(\mathbf{R}_3|\mathbf{R}_2)P(\mathbf{R}_2|\mathbf{R}_1)P(\mathbf{R}_1), \tag{17}$$

and $P(\mathbf{R}_f|\mathbf{R}_i)$ is called the transition probability from point $\mathbf{R}_i$ to point $\mathbf{R}_f$. Note that, in general, the transition probability can depend on time (measured by the index k). We will consider here only the case of a stationary Markov chain for which the transition probability is time independent.

In the following, we will use notation corresponding to the case of states $\mathbf{R}_k$ in a continuous space ("integrals" instead of "sums"), but we will ignore the possibly subtle mathematical differences between the continuous and discrete cases, and we will often use the vocabulary of the discrete case (e.g., "matrix"). The transition probability matrix, P is a *stochastic matrix*, i.e., it has the following two properties:

$$P(\mathbf{R}_f|\mathbf{R}_i) \geq 0 \quad \text{(non negativity)}, \tag{18a}$$

$$\int \mathrm{d}\mathbf{R}_\mathrm{f}\, P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = 1 \quad \text{(column normalization)}. \tag{18b}$$

The second property expresses the fact that the probability that a point $\mathbf{R}_\mathrm{i}$ is somewhere at the next step must be 1. The eigenvalues of a stochastic matrix are between 0 and 1, and there is at least one eigenvalue equal to 1. The latter property is a consequence of the fact that, for a column-normalized matrix, the vector with all components equal to one is a left eigenvector with eigenvalue 1. The target probability distribution $\rho(\mathbf{R})$ is sampled by constructing a Markov chain converging to $\rho(\mathbf{R})$. A necessary condition is that the distribution $\rho(\mathbf{R})$ is a (right) eigenvector of $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ with the eigenvalue 1

$$\int \mathrm{d}\mathbf{R}_\mathrm{i}\, P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\rho(\mathbf{R}_\mathrm{i}) = \rho(\mathbf{R}_\mathrm{f}) = \int \mathrm{d}\mathbf{R}_\mathrm{i}\, P(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\rho(\mathbf{R}_\mathrm{f}) \qquad \forall\, \mathbf{R}_\mathrm{f}, \tag{19}$$

where the second equality simply comes from the normalization condition (18b). Eq. (19) is a *stationarity condition* for $\rho(\mathbf{R})$. It means that if we start from the target distribution $\rho(\mathbf{R})$ then we will continue to sample the same distribution by applying the Markov chain. However, we need more than that. We want that any initial distribution $\rho_\mathrm{ini}(\mathbf{R})$, e.g., a delta function at some initial point, evolves to the target stationary distribution $\rho(\mathbf{R})$ by repeated applications of the transition matrix

$$\lim_{M\to\infty} \int \mathrm{d}\mathbf{R}_1\, P^M(\mathbf{R}|\mathbf{R}_1)\rho_\mathrm{ini}(\mathbf{R}_1) =$$
$$\lim_{M\to\infty} \int \mathrm{d}\mathbf{R}_1 \mathrm{d}\mathbf{R}_2 ... \mathrm{d}\mathbf{R}_M\, P(\mathbf{R}|\mathbf{R}_M)P(\mathbf{R}_M|\mathbf{R}_{M-1})...P(\mathbf{R}_2|\mathbf{R}_1)\rho_\mathrm{ini}(\mathbf{R}_1) = \rho(\mathbf{R}), \tag{20}$$

i.e. $\rho(\mathbf{R})$ must be the dominant eigenvector of P (the unique eigenvector of largest eigenvalue). If the stationarity condition (19) is satisfied then this will always be the case except if P has several eigenvectors with eigenvalue 1. One can show that the matrix P has only one eigenvector of eigenvalue 1 if P is a primitive matrix, i.e. if there is an integer $n \geq 1$ such that all the elements of the matrix P^n are strictly positive, $P^n(\mathbf{R}_k|\mathbf{R}_l) > 0, \ \forall\, \mathbf{R}_k, \mathbf{R}_l$. This means that it must be possible to move between any pair of states $\mathbf{R}_k$ and $\mathbf{R}_l$ in n steps. This ensures that all states can be visited, and that the Markov chain converges to the unique stationary distribution $\rho(\mathbf{R})$. The Markov chain is then said to be *ergodic*.

In practice, instead of imposing the stationarity condition (19), the Markov matrix is constructed by imposing the more stringent *detailed balance* condition,

$$P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\rho(\mathbf{R}_\mathrm{i}) = P(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\rho(\mathbf{R}_\mathrm{f}), \tag{21}$$

which forces the probability flux between the two states $\mathbf{R}_\mathrm{i}$ and $\mathbf{R}_\mathrm{f}$ to be the same in both directions. This is a sufficient (but not necessary) condition for $\rho(\mathbf{R})$ to be the stationary distribution. A Markov chain satisfying condition (21) is said to be reversible.

In practice, a Markov chain is realized by a *random walk*. Starting from an initial point $\mathbf{R}_1$ (or walker) – i.e. a delta-function distribution $\delta(\mathbf{R} - \mathbf{R}_1)$ – sample the second point $\mathbf{R}_2$ by drawing from the probability distribution $P(\mathbf{R}_2|\mathbf{R}_1)$, then a third point $\mathbf{R}_3$ by drawing from $P(\mathbf{R}_3|\mathbf{R}_2)$, and so on. After disregarding a certain number of iterations M_eq corresponding to a transient phase called *equilibration*, the random walk samples the stationary distribution $\rho(\mathbf{R})$ in the sense that $\rho(\mathbf{R}) = \mathrm{E}[\delta(\mathbf{R} - \mathbf{R}_k)] \approx (1/M)\sum_{k=1}^{M} \delta(\mathbf{R} - \mathbf{R}_k)$ and the averages of the estimators of the observables of interest are calculated. The rate of convergence to the stationary distribution $\rho(\mathbf{R})$ and the autocorrelation times of the observables are determined by the second

largest eigenvalue of the matrix P (see, e.g., Ref. [21]). The random walk must be sufficiently long so as to obtain a representative sample of the states making a non negligible contribution to the expected values. If the transitions between states belonging to two contributing regions of the space of states are too improbable, as may happen for example for dissociated atoms, then there is a risk that the random walk remains stuck in a region of space and seems converged, even though the true stationary distribution is not yet reached. To avoid this problem, smart choices for the transition matrix can be crucial in various applications of Monte Carlo methods [22, 23].

Metropolis-Hastings algorithm

In the Metropolis-Hastings algorithm [12, 13], one realizes a Markov chain with the following random walk. Starting from a point $\mathbf{R}_\mathrm{i}$, a new point $\mathbf{R}_\mathrm{f}$ is determined in two steps:

1. a temporary point $\mathbf{R}'_\mathrm{f}$ is proposed with the probability $P_\mathrm{prop}(\mathbf{R}'_\mathrm{f}|\mathbf{R}_\mathrm{i})$,
2. the point $\mathbf{R}'_\mathrm{f}$ is accepted (i.e. $\mathbf{R}_\mathrm{f} = \mathbf{R}'_\mathrm{f}$) with probability $P_\mathrm{acc}(\mathbf{R}'_\mathrm{f}|\mathbf{R}_\mathrm{i})$, or rejected (i.e. $\mathbf{R}_\mathrm{f} = \mathbf{R}_\mathrm{i}$) with probability $P_\mathrm{rej}(\mathbf{R}'_\mathrm{f}|\mathbf{R}_\mathrm{i}) = 1 - P_\mathrm{acc}(\mathbf{R}'_\mathrm{f}|\mathbf{R}_\mathrm{i})$

The corresponding transition probability can be written as

$$P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \begin{cases} P_\mathrm{acc}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) & \text{if } \mathbf{R}_\mathrm{f} \neq \mathbf{R}_\mathrm{i} \\ 1 - \int \mathrm{d}\mathbf{R}'_\mathrm{f}\, P_\mathrm{acc}(\mathbf{R}'_\mathrm{f}|\mathbf{R}_\mathrm{i}) P_\mathrm{prop}(\mathbf{R}'_\mathrm{f}|\mathbf{R}_\mathrm{i}) & \text{if } \mathbf{R}_\mathrm{f} = \mathbf{R}_\mathrm{i} \end{cases} \tag{22}$$

or, in a single expression,

$$P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = P_\mathrm{acc}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) + \left[1 - \int \mathrm{d}\mathbf{R}'_\mathrm{f}\, P_\mathrm{acc}(\mathbf{R}'_\mathrm{f}|\mathbf{R}_\mathrm{i}) P_\mathrm{prop}(\mathbf{R}'_\mathrm{f}|\mathbf{R}_\mathrm{i})\right] \delta(\mathbf{R}_\mathrm{i} - \mathbf{R}_\mathrm{f}). \tag{23}$$

The proposal probability $P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ is a stochastic matrix, i.e. $P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \geq 0$ and $\int \mathrm{d}\mathbf{R}_\mathrm{f} P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = 1$, ensuring that $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ fulfils the non-negativity condition (18a). The second term in Eq. (23) with the delta function ensures that $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ fulfils the normalization condition (18b). The acceptance probability is chosen so as to fulfil the detailed balance condition (21), for $\mathbf{R}_\mathrm{f} \neq \mathbf{R}_\mathrm{i}$,

$$\frac{P_\mathrm{acc}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{P_\mathrm{acc}(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})} = \frac{P_\mathrm{prop}(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f}) \rho(\mathbf{R}_\mathrm{f})}{P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \rho(\mathbf{R}_\mathrm{i})}. \tag{24}$$

Several choices are possibles. The choice of Metropolis *et al.* [12] maximizes the acceptance probability

$$P_\mathrm{acc}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \min\left\{1, \frac{P_\mathrm{prop}(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f}) \rho(\mathbf{R}_\mathrm{f})}{P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \rho(\mathbf{R}_\mathrm{i})}\right\}. \tag{25}$$

The acceptance probability is not a stochastic matrix, even though both the proposal and the total Markov matrices are stochastic. Since only the ratio $\rho(\mathbf{R}_\mathrm{f})/\rho(\mathbf{R}_\mathrm{i})$ is involved in Eq. (25), it is not necessary to calculate the normalization constant of the probability density $\rho(\mathbf{R})$. It is clear that the acceptance probability of Eq. (25) is optimal, but there is considerable scope for ingenuity in choosing a proposal probability $P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ that leads to a small autocorrelation time.

Choice of the proposal probability

The original paper of Metropolis *et al.* [12] employed a symmetric proposal matrix, in which case the proposal matrix drops out of the formula for the acceptance. The advantage of having a nonsymmetric proposal matrix was pointed out by Hastings [13]. One has a lot of freedom in the choice of the proposal probability $P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})$. The only constraints are that $P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})$ must be a stochastic matrix leading to an ergodic Markov chain and that it must be possible to efficiently sample $P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})$ with a direct sampling method. The proposal probability determines the average size of the proposed moves $\mathbf{R}_{\rm i} \rightarrow \mathbf{R}_{\rm f}$ and the average acceptance rate of these moves. In order to reduce sequential correlation, one has to make moves as large as possible but with a high acceptance rate. In practice, for a given form of the proposal matrix, there is a compromise to be found between the average size of the proposed moves and the average acceptance rate.

The simplest choice for $P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})$ is a distribution that is uniform inside a small cube $\Omega(\mathbf{R}_{\rm i})$ centered in $\mathbf{R}_{\rm i}$ and of side length Δ and zero outside

$$P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i}) = \begin{cases} \frac{1}{\Delta^{3N}} & \text{if } \mathbf{R}_{\rm f} \in \Omega(\mathbf{R}_{\rm i}) \\ 0 & \text{elsewhere .} \end{cases} \tag{26}$$

In practice, a move according to Eq. (26) is proposed,

$$\mathbf{R}_{\rm f} = \mathbf{R}_{\rm i} + \frac{\Delta}{2}\,\boldsymbol{\chi}, \tag{27}$$

where $\boldsymbol{\chi}$ is a vector of $3N$ random numbers drawn from the uniform distribution between -1 and 1. The size of the cube Δ can be adjusted so as to minimize the autocorrelation time of the local energy, but the latter remains large and the sampling is inefficient.

Clever choices use information from the distribution $\rho(\mathbf{R})$, in particular its local gradient, to guide the sampling. A choice for $P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})$ which would lead to large moves with an acceptance probability equal to 1 would be $P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i}) = \rho(\mathbf{R}_{\rm f})$, independently from $\mathbf{R}_{\rm i}$, but we would then be back to the initial problem of sampling a complicated distribution $\rho(\mathbf{R})$. A good choice for $P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})$ is the Green function of the Fokker-Planck equation in the short-time approximation

$$P_{\rm prop}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i}) = \frac{1}{(2\pi\tau)^{3N/2}} e^{-\frac{(\mathbf{R}_{\rm f}-\mathbf{R}_{\rm i}-\mathbf{v}(\mathbf{R}_{\rm i})\tau)^2}{2\tau}}, \tag{28}$$

where $\mathbf{v}(\mathbf{R}) = \boldsymbol{\nabla}\Psi(\mathbf{R})/\Psi(\mathbf{R})$ is called the *drift velocity* of the wave function and τ is the time step which can be adjusted so as to minimize the autocorrelation time of the local energy. In practice, a move according to Eq. (28) is proposed

$$\mathbf{R}_{\rm f} = \mathbf{R}_{\rm i} + \mathbf{v}(\mathbf{R}_{\rm i})\tau + \boldsymbol{\eta}, \tag{29}$$

where $\boldsymbol{\eta}$ is a vector of $3N$ random numbers drawn from the Gaussian distribution of average 0 and standard deviation $\sqrt{\tau}$. The term $\boldsymbol{\eta}$ describes an isotropic Gaussian diffusion process (or Wiener process). The term $\mathbf{v}(\mathbf{R}_{\rm i})\tau$ is a drift term which moves the random walk in the direction of increasing $|\Psi(\mathbf{R})|$.

The optimal size of the move is smaller in regions where $\mathbf{v}(\mathbf{R})$ is changing rapidly. For example, $\mathbf{v}(\mathbf{R})$ has a discontinuity at the nuclear positions. Hence, it is more efficient to make smaller moves for electrons in the core than for electrons in the valence regions. In doing this,

care must be taken to ensure the detailed balance condition. An elegant solution is provided in the VMC algorithm of Refs. [15, 3] where the electron moves are made in spherical coordinates centered on the nearest nucleus and the size of radial moves is proportional to the distance to the nearest nucleus. In addition, the size of the angular moves gets larger as one approaches a nucleus. This algorithm allows one to achieve, in many cases, an autocorrelation time of the local energy close to 1.

Expectation values

The expectation value of an operator $\hat{O}$ can be computed by averaging the corresponding local value $O(\mathbf{R}_\mathrm{f}) = \langle \mathbf{R}_\mathrm{f}|\hat{O}|\Psi\rangle/\Psi(\mathbf{R}_\mathrm{f})$ at the Monte Carlo points $\mathbf{R}_\mathrm{f}$ after the accept/reject step. A somewhat smaller statistical error can be achieved by instead averaging

$$P_\mathrm{acc}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\, O(\mathbf{R}_\mathrm{f}) + (1 - P_\mathrm{acc}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}))\, O(\mathbf{R}_\mathrm{i}), \tag{30}$$

regardless of whether the proposed move is accepted or rejected.

Moving the electrons all at once or one by one?

So far we have assumed that, for a many-electron system, all the electrons are moved and then this move is accepted or rejected in a single step. In fact, it is also possible to move the electrons one by one, i.e. move the first electron, accept or reject this move, then move the second electron, accept or reject this move, and so on. In this case, the transition probability for N electrons can be formally decomposed as

$$\begin{aligned} P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) &= P(\mathbf{r}_{1,\mathrm{f}}\mathbf{r}_{2,\mathrm{f}}...\mathbf{r}_{N,\mathrm{f}}|\mathbf{r}_{1,\mathrm{f}}\mathbf{r}_{2,\mathrm{f}}...\mathbf{r}_{N,\mathrm{i}}) \times \\ &\quad ... \times P(\mathbf{r}_{1,\mathrm{f}}\mathbf{r}_{2,\mathrm{f}}...\mathbf{r}_{N,\mathrm{i}}|\mathbf{r}_{1,\mathrm{f}}\mathbf{r}_{2,\mathrm{i}}...\mathbf{r}_{N,\mathrm{i}}) \times P(\mathbf{r}_{1,\mathrm{f}}\mathbf{r}_{2,\mathrm{i}}...\mathbf{r}_{N,\mathrm{i}}|\mathbf{r}_{1,\mathrm{i}}\mathbf{r}_{2,\mathrm{i}}...\mathbf{r}_{N,\mathrm{i}}), \end{aligned} \tag{31}$$

where each one-electron transition probability (knowing that the other electrons are fixed) is made of a proposal probability and an acceptance probability just as before. If each one-electron transition probability satisfies the stationary condition (19), then the global transition probability satisfies it as well.

Moving the N electrons one by one requires more calculation time than moving the electrons all at once, since the wave function must be recalculated after each move to calculate the acceptance probability. The calculation time does not increase by a factor of N as one may naively think but typically by a factor of 2 if the value of the wave function is recalculated in a clever way after an one-electron move. For example, for Slater determinants, one can use the matrix determinant lemma in conjunction with the Sherman-Morrison formula (see, e.g., Ref. [24]) to efficiently recalculate the values of the determinants when only one row or column has been changed. In spite of the increase in the calculation time, it has been repeatedly shown in the literature (see, e.g., Refs. [10, 15, 25, 26]) that, for systems with many electrons, moving the electrons one by one leads to a more efficient algorithm: larger moves can be made for the same average acceptance, so the points $\mathbf{R}_k$ are less sequentially correlated and the autocorrelation time of the local energy is smaller (by a factor larger than the one necessary for compensating the increase of the calculation time per iteration).

2 Diffusion Monte Carlo

2.1 Basic idea

While the VMC method is limited by the use of an approximate wave function Ψ, the idea of the DMC method [27, 28, 29, 5, 30] is to sample from the exact wave function Ψ_0 of the ground state of the system. If we have this exact wave function Ψ_0, then the associated exact energy E_0 can be obtained from the mixed expectation value using the trial wave function Ψ,

$$E_0 = \frac{\langle\Psi_0|\hat{H}|\Psi\rangle}{\langle\Psi_0|\Psi\rangle} = \frac{\int \mathrm{d}\mathbf{R}\,\Psi_0(\mathbf{R})\Psi(\mathbf{R})E_\mathrm{L}(\mathbf{R})}{\int \mathrm{d}\mathbf{R}\,\Psi_0(\mathbf{R})\Psi(\mathbf{R})}, \tag{32}$$

since Ψ_0 is an eigenfunction of the Hamiltonian $\hat{H}$. The advantage of the mixed expectation value (32) is that it does not require calculating the action of $\hat{H}$ on Ψ_0. The integral in Eq. (32) involves the local energy of the trial wave function, $E_\mathrm{L}(\mathbf{R}) = (H\Psi(\mathbf{R}))/\Psi(\mathbf{R})$, and can be estimated in a similar way as in VMC by calculating the average of $E_\mathrm{L}(\mathbf{R})$ on a sample of points $\mathbf{R}_k$ representing the mixed distribution $\Psi_0(\mathbf{R})\Psi(\mathbf{R})/\int \mathrm{d}\mathbf{R}\,\Psi_0(\mathbf{R})\Psi(\mathbf{R})$.

But how to access to the exact wave function Ψ_0? Let us consider the action of the *imaginary-time evolution operator* ($t \to -it$) on an arbitrary wave function such as the trial wave function Ψ

$$|\Psi(t)\rangle = e^{-(\hat{H}-E_\mathrm{T})t}|\Psi\rangle, \tag{33}$$

where E_T is for now an undetermined trial energy. Using the spectral decomposition of the evolution operator (written with the eigenstates Ψ_i and the eigenenergies E_i of $\hat{H}$), we see that the limit of an infinite propagation time is dominated by the state Ψ_0 with the lowest energy having a nonzero overlap with Ψ

$$\lim_{t\to\infty}|\Psi(t)\rangle = \lim_{t\to\infty}\sum_i e^{-(E_i-E_\mathrm{T})t}|\Psi_i\rangle\langle\Psi_i|\Psi\rangle = \lim_{t\to\infty} e^{-(E_0-E_\mathrm{T})t}|\Psi_0\rangle\langle\Psi_0|\Psi\rangle, \tag{34}$$

since all the other states of energies $E_i > E_0$ decay exponentially faster. The exponential $e^{-(E_0-E_\mathrm{T})t}$ can be eliminated by adjusting E_T to E_0, and we then obtain that $\Psi(t)$ becomes proportional to Ψ_0

$$\lim_{t\to\infty}|\Psi(t)\rangle \propto |\Psi_0\rangle. \tag{35}$$

In position representation, Eq. (33) is written as

$$\Psi(\mathbf{R}_\mathrm{f}, t) = \int \mathrm{d}\mathbf{R}_\mathrm{i}\, G(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}; t)\Psi(\mathbf{R}_\mathrm{i}), \tag{36}$$

where $G(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}; t) = \langle\mathbf{R}_\mathrm{f}|e^{-(\hat{H}-E_\mathrm{T})t}|\mathbf{R}_\mathrm{i}\rangle$ is called the *Green function* (or the imaginary-time propagator from $\mathbf{R}_\mathrm{i}$ to $\mathbf{R}_\mathrm{f}$). Multiplying and dividing by $\Psi(\mathbf{R}_\mathrm{f})$ and $\Psi(\mathbf{R}_\mathrm{i})$, we obtain the evolution equation of the mixed distribution $f(\mathbf{R}, t) = \Psi(\mathbf{R}, t)\Psi(\mathbf{R})$

$$f(\mathbf{R}_\mathrm{f}, t) = \int \mathrm{d}\mathbf{R}_\mathrm{i}\, \tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}; t)\, \Psi(\mathbf{R}_\mathrm{i})^2, \tag{37}$$

where $\tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}; t)$ is the *importance-sampling* Green function,

$$\tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}; t) = \Psi(\mathbf{R}_\mathrm{f})\, G(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}; t)\, \frac{1}{\Psi(\mathbf{R}_\mathrm{i})}, \tag{38}$$

i.e. $\tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i};t)$ is $G(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i};t)$ similarity transformed by the diagonal matrix that has the values of Ψ along the diagonal. In the limit of infinite time, the mixed distribution becomes proportional to the target stationary distribution: $f(\mathbf{R}) = \lim_{t\to\infty} f(\mathbf{R},t) \propto \Psi_0(\mathbf{R})\Psi(\mathbf{R})$.

In practice, an analytical expression of the Green function is known only in the limit of a short propagation time, $\tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i};\tau)$, where τ is a small time step, and one must thus iterate to obtain the stationary distribution

$$f(\mathbf{R}) = \lim_{M\to\infty} \int \mathrm{d}\mathbf{R}_1 \mathrm{d}\mathbf{R}_2 ... \mathrm{d}\mathbf{R}_M \, \tilde{G}(\mathbf{R}|\mathbf{R}_M;\tau)\tilde{G}(\mathbf{R}_M|\mathbf{R}_{M-1};\tau)...\tilde{G}(\mathbf{R}_2|\mathbf{R}_1;\tau)\,\Psi(\mathbf{R}_1)^2. \quad (39)$$

A short-time approximation to the Green function is obtained by applying the Trotter-Suzuki formula, $e^{-(\hat{T}+\hat{V})\tau} = e^{-\hat{V}\tau/2}e^{-\hat{T}\tau}e^{-\hat{V}\tau/2} + O(\tau^3)$, where $\hat{T}$ and $\hat{V}$ are the kinetic and potential energy operators. In position representation, this approximation leads to the following expression

$$G(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i};\tau) \approx \frac{1}{(2\pi\tau)^{3N/2}} e^{-\frac{(\mathbf{R}_\mathrm{f}-\mathbf{R}_\mathrm{i})^2}{2\tau}} e^{-\left(\frac{V(\mathbf{R}_\mathrm{f})+V(\mathbf{R}_\mathrm{i})}{2}-E_\mathrm{T}\right)\tau}, \quad (40)$$

where $V(\mathbf{R})$ is the potential energy. Similarly, assuming for now that the trial wave function is of the same sign in $\mathbf{R}_\mathrm{i}$ and $\mathbf{R}_\mathrm{f}$, i.e. $\Psi(\mathbf{R}_\mathrm{f})/\Psi(\mathbf{R}_\mathrm{i}) > 0$, a short-time approximation to the importance-sampling Green function is [5, 31]

$$\tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i};\tau) \approx \frac{1}{(2\pi\tau)^{3N/2}} e^{-\frac{(\mathbf{R}_\mathrm{f}-\mathbf{R}_\mathrm{i}-\mathbf{v}(\mathbf{R}_\mathrm{i})\tau)^2}{2\tau}} e^{-\left(\frac{E_\mathrm{L}(\mathbf{R}_\mathrm{f})+E_\mathrm{L}(\mathbf{R}_\mathrm{i})}{2}-E_\mathrm{T}\right)\tau}, \quad (41)$$

where the drift velocity $\mathbf{v}(\mathbf{R}) = \boldsymbol{\nabla}\Psi(\mathbf{R})/\Psi(\mathbf{R})$ and the local energy $E_\mathrm{L}(\mathbf{R})$ were assumed constant between $\mathbf{R}_\mathrm{i}$ and $\mathbf{R}_\mathrm{f}$. This short-time approximation implies a *finite time-step error* in the calculation of all observables, which should in principle be eliminated by extrapolating the results to $\tau = 0$ (see Refs. [32, 33, 34] for proofs that the time-step error vanishes in the $\tau \to 0$ limit).

2.2 Stochastic realization

The stochastic realization of Eq. (39) is less trivial than for VMC. The Green function $\tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i};\tau)$ is generally not a stochastic matrix, since it does not conserve the normalization of the probability density: $\int \mathrm{d}\mathbf{R}_\mathrm{f}\, \tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i};\tau) \neq 1$. We can nevertheless write the elements of $\tilde{G}$ as the product of the corresponding elements of a stochastic matrix P and a weight matrix W,

$$\tilde{G}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i};\tau) = P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})W(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}), \quad (42)$$

where, in the short-time approximation, $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = (2\pi\tau)^{-3N/2}\, e^{-(\mathbf{R}_\mathrm{f}-\mathbf{R}_\mathrm{i}-\mathbf{v}(\mathbf{R}_\mathrm{i})\tau)^2/2\tau}$ and $W(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = e^{-((E_\mathrm{L}(\mathbf{R}_\mathrm{f})+E_\mathrm{L}(\mathbf{R}_\mathrm{i}))/2-E_\mathrm{T})\tau}$. Note that $\tilde{G}$ reduces to a stochastic matrix in the limit $\Psi \to \Psi_0$. The stochastic realization is then a weighted random walk. Start from a walker at an initial position $\mathbf{R}_1$ with a weight $w_1 = 1$, i.e., a distribution $w_1\delta(\mathbf{R}-\mathbf{R}_1)$. Sample the position $\mathbf{R}_2$ of the walker at the next iteration from the probability distribution $P(\mathbf{R}_2|\mathbf{R}_1)$ [according to Eq. (29)] and give it weight $w_2 = W(\mathbf{R}_2|\mathbf{R}_1) \times w_1$, sample the third position $\mathbf{R}_3$ from the probability distribution $P(\mathbf{R}_3|\mathbf{R}_2)$ and give it weight $w_3 = W(\mathbf{R}_3|\mathbf{R}_2) \times w_2$, and so on. After an equilibration phase, the random walk should sample the stationary distribution $f(\mathbf{R}) \propto \mathrm{E}[w_k\delta(\mathbf{R}-\mathbf{R}_k)] \approx (1/M)\sum_{k=1}^{M} w_k\delta(\mathbf{R}-\mathbf{R}_k)$. In reality, this procedure is terribly inefficient. Because the weights w_k are products of a large number of random variables, they

can become very large at some iterations and very small at other iterations. Consequently, the averages are dominated by a few points with large weights, even though the calculation of any point of the Markov chain takes the same computational time regardless of its weight. This problem can be alleviated by keeping the product of the weights for only a finite number n of consecutive iterations [35]

$$w_k = \prod_{l=k-n+1}^{k} W(\mathbf{R}_l|\mathbf{R}_{l-1}). \tag{43}$$

However, using a finite n introduces a bias in the sampled stationary distribution. In practice, for an n large enough to have a reasonably small bias, this procedure still remains inefficient.

The solution is to use at each iteration k a population of M_k walkers, with positions $\mathbf{R}_{k,\alpha}$ and weights $w_{k,\alpha}$ (where $\alpha = 1, 2, ..., M_k$), performing random walks with a *branching or birth-death process* designed to make the weights $w_{k,\alpha}$ vary in only a small range from walker to walker in a given iteration, and from iteration to iteration, while still sampling the correct distribution $f(\mathbf{R}) \propto \mathrm{E}[\sum_{\alpha=1}^{M_k} w_{k,\alpha}\delta(\mathbf{R} - \mathbf{R}_{k,\alpha})] \approx (1/M)\sum_{k=1}^{M}\sum_{\alpha=1}^{M_k} w_{k,\alpha}\delta(\mathbf{R} - \mathbf{R}_{k,\alpha})$. Various unbiased variants are possible, characterized by a population size M_k that either varies or is constant from iteration to iteration, and by weights $w_{k,\alpha}$ that can either be equal or different for each walker.

The simplest variant uses a varying population size M_k and weights all equal to one, $w_{k,\alpha} = 1$. At each iteration k, each walker α is replaced by $m_{k,\alpha}$ unit-weight copies of itself, where $m_{k,\alpha}$ is an integer equal on average to what should be the current weight $W_{k,\alpha} = W(\mathbf{R}_{k,\alpha}|\mathbf{R}_{k-1,\alpha})$. For example, if the walker α should have the weight $W_{k,\alpha} = 2.7$ at iteration k, this walker is replaced by $m_{k,\alpha} = 3$ copies of itself with a probability 0.7 or replaced by $m_{k,\alpha} = 2$ copies of itself with a probability 0.3. More generally, $m_{k,\alpha} = \lfloor W_{k,\alpha} \rfloor + 1$ with probability $W_{k,\alpha} - \lfloor W_{k,\alpha} \rfloor$ and $m_{k,\alpha} = \lfloor W_{k,\alpha} \rfloor$ otherwise, where $\lfloor W_{k,\alpha} \rfloor$ is the nearest integer smaller than $W_{k,\alpha}$. If $m_{k,\alpha} = 0$ the walker is terminated. This procedure does not change the sampled stationary distribution [4]. This variant has the disadvantage that the integerization of the weights results in unnecessary duplications of walkers, leading to more correlated walkers and thus to a smaller number of statistically independent points in the sample. Another disadvantage is that it leads to unnecessary fluctuations in the sum of the weights, a quantity that is relevant for computing the growth estimator of the energy.

A better solution is the split-join algorithm [6] which limits the duplication of walkers by keeping residual noninteger weights $w_{k,\alpha}$. At each iteration k, after updating the weights according to $w_{k,\alpha} = W(\mathbf{R}_{k,\alpha}|\mathbf{R}_{k-1,\alpha}) \times w_{k-1,\alpha}$, each walker α with a weight $w_{k,\alpha} > 2$ is split into $\lfloor w_{k,\alpha} \rfloor$ walkers, each being attributed the weight $w_{k,\alpha}/\lfloor w_{k,\alpha} \rfloor$. If walkers α and β each have weight $< 1/2$, keep walker α with probability $w_{k,\alpha}/(w_{k,\alpha} + w_{\beta,k})$ and walker β otherwise. In either case, the surviving walker gets weight, $w_{k,\alpha} + w_{\beta,k}$. This algorithm has the advantage that it conserves the total weight of the population of walkers $W_k = \sum_{\alpha=1}^{M_k} w_{k,\alpha}$ for a given iteration. Yet another possibility is the stochastic reconfiguration algorithm [36, 37], which uses a fixed population size M_k, and walkers of equal noninteger weights within each iteration, though the weights of the walkers fluctuate from one iteration to the next.

To avoid the explosion or extinction of the population of walkers (or their weights if M_k is kept fixed), the value of E_T can be adjusted during the iterations. For example, a choice

[4] One can write: $\mathrm{E}\left[\sum_{\alpha=1}^{M_k} W_{k,\alpha}\delta(\mathbf{R} - \mathbf{R}_{k,\alpha})\right] = \mathrm{E}\left[\sum_{\alpha=1}^{M_k} m_{k,\alpha}\delta(\mathbf{R} - \mathbf{R}_{k,\alpha})\right] = \mathrm{E}\left[\sum_{\alpha=1}^{M_{k+1}} \delta(\mathbf{R} - \mathbf{R}_{k+1,\alpha})\right]$, where $\mathbf{R}_{k+1,\alpha}$ are the positions of the $M_{k+1} = \sum_{\alpha=1}^{M_k} m_{k,\alpha}$ walkers used for the next iteration $k+1$ obtained after making $m_{k,\alpha}$ copies of the α^{th} walker.

for E_T at iteration $k+1$ is $E_\mathrm{T}(k+1) = E_0^\mathrm{est}(k) - C\log(W_k/W_0)$ where $E_0^\mathrm{est}(k)$ is an estimate of E_0 at iteration k, C is a constant, W_k is the total weight of the population of walkers and W_0 is the target total weight. Because of fluctuations, E_T thus slightly varies with respect to E_0 during the iterations, which introduces a systematic bias on the weights and thus on the stationary distribution $f(\mathbf{R})$. The adjustment of E_T makes $f(\mathbf{R})$ too small in regions where $E_\mathrm{L}(\mathbf{R}) < E_0$ and too large in regions where $E_\mathrm{L}(\mathbf{R}) > E_0$. Both of these have the effect of raising the energy. This is called *population-control error*. This error is generally small and decreases with increasing number of walkers as $1/M_k$ [6]. Besides, it is possible to eliminate almost completely this error by undoing the modification of weights introduced by the variation of E_T for the last several iterations [38, 6].

In the limit of an infinitesimal time step, the transition matrix $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ has a stationary distribution $\Psi(\mathbf{R})^2$, and the weight term $W(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ converts this distribution into the mixed distribution $\Psi_0(\mathbf{R})\Psi(\mathbf{R})$. One can get rid of the finite time-step error in the transition matrix $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ by introducing an accept/reject step as in the Metropolis-Hastings algorithm [5]. For this, the transition matrix is redefined as $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = P_\mathrm{acc}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$, for $\mathbf{R}_\mathrm{i} \neq \mathbf{R}_\mathrm{f}$, with the proposal probability

$$P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \frac{1}{(2\pi\tau)^{3N/2}} e^{-\frac{(\mathbf{R}_\mathrm{f}-\mathbf{R}_\mathrm{i}-\mathbf{v}(\mathbf{R}_\mathrm{i})\tau)^2}{2\tau}}, \tag{44}$$

and the acceptance probability

$$P_\mathrm{acc}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \min\left\{1, \frac{P_\mathrm{prop}(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\Psi(\mathbf{R}_\mathrm{f})^2}{P_\mathrm{prop}(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\Psi(\mathbf{R}_\mathrm{i})^2}\right\}. \tag{45}$$

With this modification, $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ has the stationary distribution $\Psi(\mathbf{R})^2$ even for a finite time step. Of course, the finite time-step error persists in the term $W(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$. Since certain moves are rejected, $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ corresponds now to a process of diffusion with drift with an effective time step $\tau_\mathrm{eff} < \tau$. This effective time step can be estimated during the calculation from the average acceptance rate and it is consistent to use it in the term $W(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ in place of τ. In practice, just as in VMC, it is also more efficient in DMC to move the electrons one by one, i.e. to decompose $P(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ according to Eq. (31). We then arrive at a DMC algorithm very similar to the VMC algorithm, with weights in addition. Note, however, that since a relatively small time step must be used in DMC, the average moves are smaller than in VMC and the autocorrelation time of the local energy is larger than in VMC.

The energy is calculated as the average of the local energy over the distribution $f(\mathbf{R})/\int \mathrm{d}\mathbf{R} f(\mathbf{R})$. For M iterations (after the equilibration phase) and M_k walkers, we have

$$E_0 \approx \overline{E}_\mathrm{L} = \frac{\sum_{k=1}^{M}\sum_{\alpha=1}^{M_k} w_{k,\alpha} E_\mathrm{L}(\mathbf{R}_{k,\alpha})}{\sum_{k=1}^{M}\sum_{\alpha=1}^{M_k} w_{k,\alpha}}. \tag{46}$$

Just as in VMC, there is a zero-variance principle on the energy in DMC. In the limit that the trial wave function Ψ is an exact eigenfunction of the Hamiltonian, E_L is independent of $\mathbf{R}$, the weights reduce to 1, and the variance on $\overline{E}_\mathrm{L}$ vanishes.

Note that for an observable $\hat{O}$ that does not commute with the Hamiltonian, the average $\overline{O}_\mathrm{L}$ over the mixed DMC distribution is an estimator of $\langle\Psi_0|\hat{O}|\Psi\rangle/\langle\Psi_0|\Psi\rangle$ which is only an approximation to the exact expectation value $\langle\Psi_0|\hat{O}|\Psi_0\rangle/\langle\Psi_0|\Psi_0\rangle$ with an $\mathcal{O}(||\Psi-\Psi_0||)$ error. Since the average $\overline{O}_\mathrm{L}$ over the VMC distribution also has an error that is linear in $||\Psi-\Psi_0||$

but with a prefactor that is twice as large, an $\mathcal{O}(||\Psi - \Psi_0||^2)$ approximation is provided by twice the average of $O_{\rm L}$ over the mixed DMC distribution minus the average of $O_{\rm L}$ over the VMC distribution [39]. For a recent survey of exact methods for sampling the pure distribution Ψ_0^2, see Ref. [40], and for a discussion of the techniques used for calculating pure expectation values of various classes of operators see Ref. [2].

2.3 Fermionic sign problem

In Eq. (41), we have assumed that the trial wave function $\Psi(\mathbf{R})$ is always of the same sign, i.e. that it does not have any nodes (points $\mathbf{R}$ so that $\Psi(\mathbf{R}) = 0$). This is the case for the ground-state wave function of a Bosonic system, and for a few simple electronic systems (two electrons in a spin-singlet state, such as the ground state of the He atom or of the H_2 molecule). In this case, the algorithm presented above allows one to obtain the exact energy of the system, after elimination of the finite time-step error and the population-control error. If the wave function of the Fermionic ground state has nodes, then there is always at least one Bosonic state of lower energy, and the true ground state of the Hamiltonian is a Bosonic state for which the wave function $\Psi_{\rm B}(\mathbf{R})$ can be chosen strictly positive. If one applied the Green function exactly, starting from the distribution $\Psi(\mathbf{R})^2$ the distribution would correctly converge to $\Psi_0(\mathbf{R})\Psi(\mathbf{R})$ since the trial wave function is antisymmetric (with respect to the exchange of two electrons) and has a zero overlap with all the Bosonic states which are symmetric. However, in reality one applies the Green function using a finite sampling in position space which does not allow one to impose the antisymmetry. Starting from an antisymmetric wave function Ψ, a small component of $\Psi_{\rm B}$ can thus appear, and it grows and eventually dominates. The distribution tends to $\Psi_{\rm B}(\mathbf{R})\Psi(\mathbf{R})$ and the energy formula in Eq. (46) only gives 0/0 (the positive and negative weights cancel out) with statistical noise. Even if one imposed antisymmetry and eliminated the Bosonic states, e.g. by considering all electron permutations in each walker, the problem persists because different paths between the same points in this antisymmetrized space can contribute with opposite sign. Since Ψ_0 and $-\Psi_0$ are equally good solutions of the Schrödinger equation, the algorithm would sample each with approximately equal probability, leading again to the cancellation of positive and negative weight contributions. These are different manifestations of the infamous *Fermionic sign problem*.

2.4 Fixed-node approximation

To avoid the sign problem in DMC, the *fixed-node approximation* (FN) [28, 41, 29] is introduced. The idea is to force the convergence to a wave function approximating the Fermionic ground state by fixing its nodes to be the same as those of the trial wave function $\Psi(\mathbf{R})$. Formally, one can define the FN Hamiltonian, $\hat{H}_{\rm FN}$, by adding to the true Hamiltonian $\hat{H}$ infinite potential barriers at the location of the nodes of $\Psi(\mathbf{R})$ [42]. The ground-state wave function of this Hamiltonian is called the FN wave function $\Psi_{\rm FN}$ and its energy is the FN energy $E_{\rm FN}$,

$$\hat{H}_{\rm FN}|\Psi_{\rm FN}\rangle = E_{\rm FN}|\Psi_{\rm FN}\rangle. \tag{47}$$

In the $3N$-dimensional space of positions $\mathbf{R}$, the nodes of $\Psi(\mathbf{R})$ define hypersurfaces of dimension $3N-1$. The position space is then partitioned in nodal pockets of $\Psi(\mathbf{R})$, delimited by nodal surfaces, in which the wave function has a fixed sign. In each nodal pocket, the FN wave function is the solution to the Schrödinger equation satisfying vanishing boundary conditions on the nodal

surface. The FN Green function corresponding to the Hamiltonian $\hat{H}_{\rm FN}$ is

$$G_{\rm FN}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i};t) = \langle \mathbf{R}_{\rm f}|e^{-(\hat{H}_{\rm FN}-E_{\rm T})t}|\mathbf{R}_{\rm i}\rangle, \tag{48}$$

and only permits moves $\mathbf{R}_{\rm i} \to \mathbf{R}_{\rm f}$ inside a nodal pocket. The importance-sampling FN Green function,

$$\tilde{G}_{\rm FN}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i};t) = \Psi(\mathbf{R}_{\rm f})\, G_{\rm FN}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i};t)\, \frac{1}{\Psi(\mathbf{R}_{\rm i})}, \tag{49}$$

also confines the moves inside a nodal pocket, and it is thus always greater or equal to zero. A short-time approximation to $\tilde{G}_{\rm FN}(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i};t)$ is then again given by Eq. (41). The stochastic algorithm previously described can thus be applied directly. Thanks to the FN approximation, the weights always remain positive, and the stationary mixed distribution $f(\mathbf{R})$ is proportional to $\Psi_{\rm FN}(\mathbf{R})\Psi(\mathbf{R})$.

The largest contributions to the finite time-step error come from singularities of the drift velocity $\mathbf{v}(\mathbf{R}) = \boldsymbol{\nabla}\Psi(\mathbf{R})/\Psi(\mathbf{R})$ and of the local energy $E_{\rm L}(\mathbf{R})$ in the Green function of Eq. (41). Since the gradient of the trial wave function $\boldsymbol{\nabla}\Psi(\mathbf{R})$ (and of the exact wave function) does not generally vanish at the location of the nodes, the drift velocity $\mathbf{v}(\mathbf{R})$ diverges at the nodes, which leads to too large moves near the nodes for finite time steps. The drift velocity has discontinuities also at particle coalescences (both electron-nucleus and electron-electron). Similarly, for an approximate trial wave function $\Psi(\mathbf{R})$, the local energy $E_{\rm L}(\mathbf{R})$ also diverges at the nodes and at particle coalescences (unless the Kato cusp conditions [43, 44] are imposed). The finite time-step error can be greatly reduced by replacing $\mathbf{v}(\mathbf{R})$ and $E_{\rm L}(\mathbf{R})$ in the Green function by approximate integrals of these quantities over the time step τ [6].

If importance sampling is not used, it is necessary to kill walkers that cross the nodes of Ψ to impose the FN boundary condition. In practice importance sampling is almost always used. In that case, it is better to reject the moves of walkers crossing the nodes, which is consistent with the FN approximation, but even this is not necessary since the number of walkers that cross the node per unit time goes to zero as $\tau \to 0$ [6][5]. For a finite time step, there are node crossing events, but these are just part of the finite time-step error and in practice essentially the same time-step error is obtained whether the walkers are allowed to cross nodes or not.

We may wonder whether the walkers have to sample all the nodal pockets. The tiling theorem [45] establishes that all the nodal pockets of the ground-state wave function of a many-electron Hamiltonian with a reasonable local potential are equivalent, i.e., the permutations of any nodal pocket are sufficient to cover the entire space. This means that, for ground-state calculations, the distribution of the walkers over the nodal pockets is irrelevant.

By applying the variational principle, it is easy to show that the FN energy is an upper bound to the exact energy

$$E_{\rm FN} = \frac{\langle \Psi_{\rm FN}|\hat{H}_{\rm FN}|\Psi_{\rm FN}\rangle}{\langle \Psi_{\rm FN}|\Psi_{\rm FN}\rangle} = \frac{\langle \Psi_{\rm FN}|\hat{H}|\Psi_{\rm FN}\rangle}{\langle \Psi_{\rm FN}|\Psi_{\rm FN}\rangle} \geq E_0, \tag{50}$$

the second equality coming from the fact that the infinite potential barriers in $\hat{H}_{\rm FN}$ do not contribute to the expectation value since $\Psi_{\rm FN}$ is zero on the nodal surface. Since the wave

[5] The drift velocity moves electrons away from the nodal surface, but for small τ the diffusion term dominates and can cause walkers to cross nodes. The density of walkers goes quadratically to zero near nodes and walkers that are roughly within a distance $\sqrt{\tau}$ can cross. Hence the number that cross per Monte Carlo step goes as $\int_0^{\sqrt{\tau}} x^2 {\rm d}x \sim \tau^{3/2}$, and so the number that cross per unit time goes to zero as $\sqrt{\tau}$.

function $\Psi_{\rm FN}$ is an eigenfunction of $\hat{H}_{\rm FN}$, the FN energy can also be expressed using the mixed expectation value

$$E_{\rm FN} = \frac{\langle \Psi_{\rm FN} | \hat{H}_{\rm FN} | \Psi \rangle}{\langle \Psi_{\rm FN} | \Psi \rangle} = \frac{\langle \Psi_{\rm FN} | \hat{H} | \Psi \rangle}{\langle \Psi_{\rm FN} | \Psi \rangle}, \tag{51}$$

where the Hamiltonian $\hat{H}_{\rm FN}$ has been replaced by $\hat{H}$ for essentially the same reason as before, viz., both Ψ and $\Psi_{\rm FN}$ are zero where $\hat{H}_{\rm FN}$ is infinite. In practice, the FN energy is thus obtained by the same energy formula (46).

The accuracy of the DMC results with the FN approximation thus depends on the quality of the nodal surface of the trial wave function. For a trial wave function with a single Slater determinant, the error due to the FN approximation can often be large, even for energy differences. For example, for the C_2 molecule, the FN error for a single-determinant trial wave function is 1.6 eV for the total energy and 0.8 eV for the dissociation energy [46]. In order to reduce this error, one can use several Slater determinants and optimize the parameters of the wave function (Jastrow parameters, coefficients of determinants, coefficients that express the orbitals in terms of the basis functions, and exponents of the basis functions) in VMC (see Refs. [47, 48, 49, 50, 51, 52, 53, 46]). This allows one to reach near chemical accuracy ($\sim$ 1 kcal/mol) in DMC for calculations of energy differences such as molecular atomization energies [54].

Appendix: Statistical estimator of nonlinear functions of expectation values

We often need to estimate nonlinear functions of expectation values. The simplest example is the variance,

$$\mathrm{V}[X] = \mathrm{E}[X^2] - \mathrm{E}[X]^2, \tag{52}$$

which is a quadratic function of the expectation values of two random variables X^2 and X. Another example is the calculation of the energy in DMC using weights [see Eq. (46)], with simplified notation,

$$E_0 = \frac{\mathrm{E}[wE_\mathrm{L}]}{\mathrm{E}[w]}, \tag{53}$$

involving a ratio of two expectation values.

Consider a nonlinear function, $f(\mathrm{E}[X], \mathrm{E}[Y])$, of two expectation values, $\mathrm{E}[X]$ and $\mathrm{E}[Y]$. The usual simple estimator of $f(\mathrm{E}[X], \mathrm{E}[Y])$ is $f(\overline{X}, \overline{Y})$ where

$$\overline{X} = \frac{1}{M_\mathrm{b}} \sum_{b=1}^{M_\mathrm{b}} \overline{X}_b, \tag{54}$$

and

$$\overline{Y} = \frac{1}{M_\mathrm{b}} \sum_{b=1}^{M_\mathrm{b}} \overline{Y}_b, \tag{55}$$

are averages over a finite number of blocks M_b, and $\overline{X}_b$ and $\overline{Y}_b$ are the block averages of X and Y, respectively [see Eq. (6)]. As discussed before, each block average is itself an average over a sufficiently large number of steps, M_s, so that the block averages can be assumed to be independent of each other. The simple estimator can be justified as follows. (i) When the law of large numbers holds, $\overline{X}$ and $\overline{Y}$ converge, with increasing M_b, almost surely to $\mathrm{E}[X]$ and $\mathrm{E}[Y]$, respectively. (ii) Hence, $f(\overline{X}, \overline{Y})$ converges to $f(\mathrm{E}[X], \mathrm{E}[Y])$ provided that f is continuous at the point $(\mathrm{E}[X], \mathrm{E}[Y])$. However, because f is nonlinear, $f(\overline{X}, \overline{Y})$ has a systematic error, i.e. $E[f(\overline{X}, \overline{Y})] \neq f(\mathrm{E}[X], \mathrm{E}[Y])$, that vanishes only in the limit of infinite sample size, $M_\mathrm{b} \to \infty$. Though not necessary, in the following, for the sake of simplicity, we assume that $f(\overline{X}, \overline{Y})$ has a finite expectation value and a finite variance[6].

Systematic error

Let us first consider the case of a nonlinear function $f(x)$ of a single variable. By definition, the systematic error of the estimator $f(\overline{X})$ is $\mathrm{E}[f(\overline{X})] - f(\mathrm{E}[X])$. The systematic error can be evaluated using a second-order Taylor expansion of the function $f(\overline{X})$ around $\mathrm{E}[X]$ (assuming that f is at least a C^2 function in the neighborhood of $\mathrm{E}[X]$)

$$f(\overline{X}) = f(\mathrm{E}[X]) + \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right) (\overline{X} - \mathrm{E}[X]) + \frac{1}{2}\left(\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right) (\overline{X} - \mathrm{E}[X])^2 + \cdots, \tag{56}$$

[6] $E[f(\overline{X}, \overline{Y})]$ can be undefined when f has a point at which it diverges, e.g., $f(x, y) = x/y$. In this case, this definition of the systematic error does not have a strict meaning. In practice, this is not a problem for this f provided that the absolute value of the expectation value of Y over a block is larger than a few times the square root of its variance, say, $|E[\overline{Y_\mathrm{b}}]| > 5\sqrt{V[\overline{Y_\mathrm{b}}]}$.

where the derivatives of f are evaluated at $\mathrm{E}[X]$. If we take the expectation value of this expression, the linear term vanishes

$$\mathrm{E}[f(\overline{X})] = f(\mathrm{E}[X]) + \frac{1}{2}\left(\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right)\,\mathrm{E}\big[(\overline{X} - \mathrm{E}[X])^2\big] + \cdots . \tag{57}$$

Assuming the random variable X has a finite variance and that the higher-order terms can be neglected, the systematic error is thus

$$\mathrm{E}[f(\overline{X})] - f(\mathrm{E}[X]) \;=\; \frac{1}{2}\left(\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right)\,\mathrm{V}[\overline{X}] + \cdots \;=\; \frac{1}{2}\left(\frac{\mathrm{d}^2 f}{\mathrm{d}x^2}\right)\,\frac{\mathrm{V}[\overline{X}_\mathrm{b}]}{M_\mathrm{b}} + \cdots . \tag{58}$$

Hence, the estimator $f(\overline{X})$ has a systematic error with a leading term proportional to $1/M_\mathrm{b}$. Note that if the hypotheses (especially the finite variance) are not satisfied, the systematic error can decrease more slowly than $1/M_\mathrm{b}$. Equation (58) can easily be generalized to a function of several variables. For example, for two variables, the systematic error is

$$\begin{aligned}\mathrm{E}[f(\overline{X},\overline{Y})] - f(\mathrm{E}[X],\mathrm{E}[Y]) \;=\;& \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}\right)\,\frac{\mathrm{V}[\overline{X}_\mathrm{b}]}{M_\mathrm{b}} + \frac{1}{2}\left(\frac{\partial^2 f}{\partial y^2}\right)\,\frac{\mathrm{V}[\overline{Y}_\mathrm{b}]}{M_\mathrm{b}} \\ &+ \left(\frac{\partial^2 f}{\partial x \partial y}\right)\,\frac{\mathrm{Cov}[\overline{X}_\mathrm{b},\overline{Y}_\mathrm{b}]}{M_\mathrm{b}} + \cdots ,\end{aligned} \tag{59}$$

where the second-order derivatives are evaluated at $(\mathrm{E}[X],\mathrm{E}[Y])$. The leading neglected term is $O(1/M_\mathrm{b}^2)$ if the third moments of X and Y are finite. The second-order derivatives in Eq. (59) can in practice be evaluated at $(\overline{X},\overline{Y})$ without changing the order of the approximation. Hence, an estimator for $f(\mathrm{E}[X],\mathrm{E}[Y])$ with an $O(1/M_\mathrm{b}^2)$ error is

$$\begin{aligned}f(\mathrm{E}[X],\mathrm{E}[Y]) \;\approx\;& f(\overline{X},\overline{Y}) - \frac{1}{2}\left(\frac{\partial^2 f}{\partial x^2}\right)\,\frac{\mathrm{V}[\overline{X}_\mathrm{b}]}{M_\mathrm{b}} - \frac{1}{2}\left(\frac{\partial^2 f}{\partial y^2}\right)\,\frac{\mathrm{V}[\overline{Y}_\mathrm{b}]}{M_\mathrm{b}} \\ &- \left(\frac{\partial^2 f}{\partial x \partial y}\right)\,\frac{\mathrm{Cov}[\overline{X}_\mathrm{b},\overline{Y}_\mathrm{b}]}{M_\mathrm{b}} + \cdots ,\end{aligned} \tag{60}$$

where the second-order derivatives are evaluated at $(\overline{X},\overline{Y})$.

This approach is general and can be used to recover some well-known unbiased estimators. For example, let us consider the covariance of two random variables

$$\mathrm{Cov}[X,Y] \;=\; \mathrm{E}[XY] - \mathrm{E}[X]\mathrm{E}[Y] \;=\; f(\mathrm{E}[XY],\mathrm{E}[X],\mathrm{E}[Y]), \tag{61}$$

for which $f(x,y,z) = x - yz$. In this case, the generalization of Eq. (59) to three variables with $\overline{X} = (1/M)\sum_{i=1}^{M} X_i$ and $\overline{Y} = (1/M)\sum_{i=1}^{M} Y_i$ where X_i and Y_i are M uncorrelated realizations of X and Y, respectively, gives

$$\mathrm{E}[\overline{XY} - \overline{X}\,\overline{Y}] - \mathrm{Cov}[X,Y] \;=\; -\frac{\mathrm{Cov}[X,Y]}{M}, \tag{62}$$

which leads to the usual unbiased estimator for the covariance

$$\mathrm{Cov}[X,Y] \;\approx\; \frac{M}{M-1}\left(\overline{XY} - \overline{X}\,\overline{Y}\right) \;=\; \frac{1}{M-1}\sum_{i=1}^{M}\left(X_i - \overline{X}\right)\left(Y_i - \overline{Y}\right). \tag{63}$$

Statistical uncertainty

First consider a function of a single variable. The statistical uncertainty of $f(\overline{X})$ is given by $\sigma[f(\overline{X})] = \sqrt{\mathrm{V}[f(\overline{X})]}$ where the variance of $f(\overline{X})$ is $\mathrm{V}[f(\overline{X})] = \mathrm{E}\left[\left(f(\overline{X}) - \mathrm{E}[f(\overline{X})]\right)^2\right]$. Subtracting Eq. (57) from Eq. (56) gives

$$f(\overline{X}) - \mathrm{E}[f(\overline{X})] = \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right) \, (\overline{X} - \mathrm{E}[X]) + \cdots . \tag{64}$$

Taking the square of this equation and the expectation value leads to the leading term in the variance of $f(\overline{X})$

$$\mathrm{V}[f(\overline{X})] = \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right)^2 \, \mathrm{V}[\overline{X}] + \cdots . \tag{65}$$

This equation can be generalized to a function of several variables. For example, for two variables, the variance of $f(\overline{X}, \overline{Y})$ is

$$\mathrm{V}[f(\overline{X}, \overline{Y})] = \left(\frac{\partial f}{\partial x}\right)^2 \, \mathrm{V}[\overline{X}] + \left(\frac{\partial f}{\partial y}\right)^2 \, \mathrm{V}[\overline{Y}] + 2\left(\frac{\partial f}{\partial x}\right)\left(\frac{\partial f}{\partial y}\right) \mathrm{Cov}[\overline{X}, \overline{Y}] + \cdots . \tag{66}$$

Equation (66) can be used for estimating the variance of $f(\overline{X}, \overline{Y})$ at the cost of evaluating the variances $\mathrm{V}[\overline{X}]$ and $\mathrm{V}[\overline{Y}]$ and the covariance $\mathrm{Cov}[\overline{X}, \overline{Y}]$. Note however, that it can give a severe underestimate of the error if $\partial f/\partial x$ and $\partial f/\partial y$ are small and M_b is not sufficiently large.

There is a simple alternative for estimating the variance of f that does not suffer from this limitation and does not require calculating covariances. Consider again the case of a single variable. Instead of defining the block average of f in the obvious way, i.e. $\overline{f}_b = f(\overline{X}_b)$, we define the block average of f as

$$\begin{aligned} \overline{f}_1 &= f(\overline{X}_1) \quad \text{for the first block } b = 1 \\ \overline{f}_b &= bf(\overline{X}(b)) - (b-1)f(\overline{X}(b-1)) \quad \text{for any block } b \geq 2, \end{aligned} \tag{67}$$

where $\overline{X}(b)$ is the running global average up to block b

$$\overline{X}(b) = \frac{1}{b}\sum_{b'=1}^{b} \overline{X}_{b'}. \tag{68}$$

With this definition of the block average, it is easy to check that

$$f(\overline{X}) = \frac{1}{M_\mathrm{b}}\sum_{b=1}^{M_\mathrm{b}} \overline{f}_b, \tag{69}$$

i.e. we have written $f(\overline{X})$ as an average of random variables $\overline{f}_b$. Provided that the variance of X is finite, the block average $\overline{f_b}$ introduced in Eq. (67) can be expanded as

$$\overline{f}_b = f(\mathrm{E}[X]) + \left(\frac{\mathrm{d}f}{\mathrm{d}x}\right) \, (\overline{X}_b - \mathrm{E}[X]) + \cdots . \tag{70}$$

Assuming that f has a second-order Taylor expansion, the neglected term converges to zero in probability for large b, at least as $1/(bM_\mathrm{s})$. Therefore, according to Eq. (70), for large b, the random variables $\overline{f}_b$ converge to independent and equidistributed random variables (since the

random variables $\overline{X}_b$ are)[7]. Consequently, the variance of $f(\overline{X})$ can be estimated with the usual formula

$$\mathrm{V}[f(\overline{X})] \approx \frac{\mathrm{V}[\overline{f}_b]}{M_\mathrm{b}} \approx \frac{1}{M_\mathrm{b}-1}\left(\frac{1}{M_\mathrm{b}}\sum_{b=1}^{M_\mathrm{b}}\overline{f}_b^2 - f(\overline{X})^2\right). \tag{71}$$

This formula applies similarly for functions of several variables. The advantage of Eq. (71) over Eq. (66) for estimating the variance is that it is much simpler to implement and compute, especially for functions of many variables. The estimation of the variance can be simply updated at each block, just as for the expectation value.

[7]The naive definition of the block average as $\overline{f}_b = f(\overline{X}_b)$ would also lead to Eq. (70) but the neglected term would not converge to zero for large b.

References

[1] B. L. Hammond, Jr. W. A. Lester, and P. J. Reynolds. *Monte Carlo Methods in Ab Initio Quantum Chemistry*. World Scientific, Singapore, 1994.

[2] M. P. Nightingale and C. J. Umrigar. In D. M. Ferguson, J. I. Siepmann, and D. G. Truhlar, editors, *Monte Carlo Methods in Chemistry*, Advances in Chemical Physics Vol. 105, page Chapter 4. Wiley, NY, 1998.

[3] C. J. Umrigar. In M. P. Nightingale and C. J. Umrigar, editors, *Quantum Monte Carlo Methods in Physics and Chemistry*, NATO ASI Ser. C 525, page 129. Kluwer, Dordrecht, 1999.

[4] W. M. C. Foulkes, L. Mitas, R. J. Needs, and G. Rajagopal. Quantum Monte Carlo simulations of solids. *Rev. Mod. Phys.*, 73:33, 2001.

[5] P. J. Reynolds, D. M. Ceperley, B. J. Alder, and W. A. Lester. Fixed-node quantum Monte Carlo for molecules. *J. Chem. Phys.*, 77:5593, 1982.

[6] C. J. Umrigar, M. P. Nightingale, and K. J. Runge. A diffusion Monte Carlo algorithm with very small time-step errors. *J. Chem. Phys.*, 99:2865, 1993.

[7] B. M. Austin, D. Yu. Zubarev, and W. A. Lester. Quantum Monte Carlo and related approaches. *Chem. Rev.*, 112:263, 2012.

[8] J. Kolorenč and L. Mitas. Applications of quantum Monte Carlo methods in condensed systems. *Rep. Prog. Phys.*, 74:026502, 2011.

[9] W. L. McMillan. Ground state of liquid He^4. *Phys. Rev.*, 138:A442, 1965.

[10] D. Ceperley, G. V. Chester, and M. H. Kalos. Monte Carlo simulation of a many-fermion study. *Phys. Rev. B*, 16:3081, 1977.

[11] H. Conroy. Molecular Schrödinger equation. II. Monte Carlo evaluation of integrals. *J. Chem. Phys.*, 41:1331, 1964.

[12] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *J. Chem. Phys.*, 21:1087, 1953.

[13] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97, 1970.

[14] R. E. Caflisch. Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 1998:1, 1998.

[15] C. J. Umrigar. Accelerated Metropolis method. *Phys. Rev. Lett.*, 71:408, 1993.

[16] R. Assaraf and M. Caffarel. Zero-variance principle for Monte Carlo algorithms. *Phys. Rev. Lett.*, 83:4682, 1999.

[17] R. Assaraf and M. Caffarel. Computing forces with quantum Monte Carlo. *J. Chem. Phys.*, 113:4028, 2000.

[18] R. Assaraf and M. Caffarel. Zero-variance zero-bias principle for observables in quantum Monte Carlo: Application to forces. *J. Chem. Phys.*, 119:10536, 2003.

[19] R. Assaraf, M. Caffarel, and A. Scemama. Improved Monte Carlo estimators for the one-body density. *Phys. Rev. E*, 75:035701(R), 2007.

[20] J. Toulouse, R. Assaraf, and C. J. Umrigar. Zero-variance zero-bias quantum Monte Carlo estimators of the spherically and system-averaged pair density. *J. Chem. Phys.*, 126:244112, 2007.

[21] W. R. Gilks, S. Richardson, and D. J. Spiegelhalter. *Markov Chain Monte Carlo in Practice*. Chapman & Hall, London, 1996.

[22] U. Wolff. Collective Monte Carlo updating for spin systems. *Phys. Rev. Lett.*, 62:361–364, 1989.

[23] R. G. Melko and A. W. Sandvik. Stochastic series expansion algorithm for the $S = 12$ XY model with four-site ring exchange. *Phys. Rev. E*, 72:026702, 2005.

[24] W. Press, S. Teukolsky, W. Vetterling, and B. Flannery. *Numerical Recipes*. Cambridge University Press, Cambridge, 1992.

[25] P. López Rios, A. Ma, N. D. Drummond, M. D. Towler, and R. J. Needs. Inhomogeneous backflow transformations in quantum Monte Carlo calculations. *Phys. Rev. E*, 74:066701, 2006.

[26] R. M. Lee, G. J. Conduit, N. Nemec, P. López Rios, and N. D. Drummond. Strategies for improving the efficiency of quantum Monte Carlo calculations. *Phys. Rev. E*, 83:066706, 2011.

[27] R. Grimm and R. G. Storer. Monte-Carlo solution of Schrödinger's equation. *J. Comput. Phys.*, 7:134, 1971.

[28] J. B. Anderson. A random-walk simulation of the Schrödinger equation: H_3^+. *J. Chem. Phys.*, 63:1499, 1975.

[29] J. B. Anderson. Quantum chemistry by random walk. H ^2P, H_3^+ D_{3h} $^1A_1'$, H_2 $^3\Sigma_u^+$, H_4 $^1\Sigma_g^+$, Be ^1S. *J. Chem. Phys.*, 65:4121, 1976.

[30] J. W. Moskowitz, K. E. Schmidt, M. A. Lee, and M. H. Kalos. A new look at correlation energy in atomic and molecular systems. II. The application of the Green's function Monte Carlo method to LiH. *J. Chem. Phys.*, 77:349, 1982.

[31] K. E. Schmidt. Variational and Green's function Monte Carlo calculations of few-body systems. In L. S. Ferreira, A. C. Fonseca, and L. Streit, editors, *Models and Methods in Few-Body Physics, Lecture Notes in Physics Vol. 273*, pages 363–407. Springer, Berlin Heidelberg, 1987.

[32] S. M. Rothstein and J. Vrbik. A Green's function used in diffusion Monte Carlo. *J. Chem. Phys.*, 87:1902, 1987.

[33] J. B. Anderson and D. R. Garmer. Validity of random walk methods in the limit of small time steps. *J. Chem. Phys.*, 87:1903, 1987.

[34] P. J. Reynolds, R. K. Owen, and W. A. Lester. Is there a zeroth order timestep error in diffusion quantum Monte Carlo? *J. Chem. Phys.*, 87:1905, 1987.

[35] M. Caffarel and P. Claverie. Development of a pure diffusion quantum Monte Carlo method using a full generalized Feynman–Kac formula. II. Applications to simple systems. *J. Chem. Phys.*, 88(2):1100–1109, 1988.

[36] M. Calandra Buonaura and S. Sorella. Numerical study of the two-dimensional Heisenberg model using a Green function Monte Carlo technique with a fixed number of walkers. *Phys. Rev. B*, 57:11446, 1998.

[37] R. Assaraf, M. Caffarel, and A. Khelif. Diffusion Monte Carlo methods with a fixed number of walkers. *Phys. Rev. E*, 61:4566, 2000.

[38] M. P. Nightingale and H. W. J. Blöte. Gap of the linear spin-1 Heisenberg antiferromagnet: A Monte Carlo calculation. *Phys. Rev. B*, 33:659, 1986.

[39] D. M. Ceperley and M. H. Kalos. Quantum many-body problem. In K. Binder, editor, *Monte Carlo Methods in Statistical Physics*, pages 145–194. Springer, Berlin, 1979.

[40] S. M. Rothstein. A survey on pure sampling in quantum Monte Carlo methods. *Can. J. Chem.*, 91:505, 2013.

[41] D. J. Klein and H. M. Pickett. Nodal hypersurfaces and Anderson's random-walk simulation of the Schrödinger equation. *J. Chem. Phys.*, 64:4811, 1976.

[42] A. Badinski, P. D. Haynes, and R. J. Needs. Nodal Pulay terms for accurate diffusion quantum Monte Carlo forces. *Phys. Rev. B*, 77:085111, 2008.

[43] T. Kato. On the eigenfunctions of many-particle systems in quantum mechanics. *Comm. Pure Appl. Math.*, 10:151, 1957.

[44] R. T. Pack and W. Byers-Brown. Cusp conditions for molecular wavefunctions. *J. Chem. Phys.*, 45:556, 1966.

[45] D. M. Ceperley. Fermion nodes. *J. Stat. Phys.*, 63:1237, 1991.

[46] J. Toulouse and C. J. Umrigar. Full optimization of Jastrow-Slater wave functions with application to the first-row atoms and homonuclear diatomic molecules. *J. Chem. Phys.*, 128:174101, 2008.

[47] C. J. Umrigar, K. G. Wilson, and J. W. Wilkins. Optimized trial wave functions for quantum Monte Carlo calculations. *Phys. Rev. Lett.*, 60:1719, 1988.

[48] M. P. Nightingale and V. Melik-Alaverdian. Optimization of ground- and excited-state wave functions and van der Waals clusters. *Phys. Rev. Lett.*, 87:043401, 2001.

[49] F. Schautz and C. Filippi. Optimized Jastrow-Slater wave functions for ground and excited states: Application to the lowest states of ethene. *J. Chem. Phys.*, 120:10931, 2004.

[50] C. J. Umrigar and C. Filippi. Energy and variance optimization of many-body wave functions. *Phys. Rev. Lett.*, 94:150201, 2005.

[51] S. Sorella. Wave function optimization in the variational Monte Carlo method. *Phys. Rev. B*, 71:241103, 2005.

[52] C. J. Umrigar, J. Toulouse, C. Filippi, S. Sorella, and R. G. Hennig. Alleviation of the fermion-sign problem by optimization of many-body wave functions. *Phys. Rev. Lett.*, 98:110201, 2007.

[53] J. Toulouse and C. J. Umrigar. Optimization of quantum Monte Carlo wave functions by energy minimization. *J. Chem. Phys.*, 126:084102, 2007.

[54] F. R. Petruzielo, J. Toulouse, and C. J. Umrigar. Approaching chemical accuracy with quantum Monte Carlo. *J. Chem. Phys.*, 136:124116, 2012.