Introduction to Variational and Projector Monte Carlo

Cyrus Umrigar

Physics Department, Cornell University, Ithaca.

Email: CyrusUmrigar@cornell.edu

Jul 29 - Aug 2, 2019, MOLSSI school on stochastic electronic structure

Lecture 1
Basics of MC

1. What is quantum Monte Carlo?
2. Essence of variational and projector Monte Carlo methods
3. Early history of MC and QMC.
4. Central limit theorem, Chebyshev inequality
5. Monte Carlo vs. deterministic integration
    1. Importance Sampling
6. Pseudo-random vs. quasi-random numbers
7. Sampling nonuniform probability densities
    1. Transformation method
    2. Rejection method
    3. Metropolis-Hastings algorithm (part of next lecture)
8. Unbiased estimators

Lecture 2

Variational Monte Carlo and Metropolis-Hastings Algorithm

1. Ingredients needed for accurate and efficient variational Monte Carlo
2. Metropolis-Hastings Algorithm
    1 Markov chains/matrices/kernels
    2 Stationarity condition and detailed balance condition
    3 Choice of acceptance matrix/kernel
    4 Choice of proposal matrix/kernel
    5 Various observations and connection to heat-bath/Gibbs sampler
    6 Optimizing the Markov matrix for finite spaces
3. Estimating errors of autocorrelated variables
4. Example of a variational wavefunction

Optimization of Many-body Wavefunctions (if there is time)

1. Measures of goodness of many-body wave functions
2. Pros and cons of optimizing energy versus variance of energy
3. Optimization methods
    1 Newton method
    2 linear method
    3 augmented hessian method
    4 perturbation theory
    5 stabilization
    6 optimization of linear combination of energy and variance

Solving the Many-Body Schrödinger Equation

Straightforward approach:

1. Expand the many-body wavefunction as a linear combination of (possibly nonorthogonal) basis states (determinants for Fermions).
2. Compute Hamiltonian and overlap matrices, H and S in this basis
3. Solve the generalized eigenvalue problem $Hc = ESc$

Problem:

The number of many-body states grows combinatorially in the number of single particle basis states and the number of particles, $\binom{N_{\rm orb}}{N_\uparrow} \times \binom{N_{\rm orb}}{N_\downarrow}$, e.g.

Molecules with 20 electrons in 200 orbitals: $\binom{200}{10}^2 = 5.0 \times 10^{32}$

(Partial) Solutions:

1. Selected CI: If only a small fraction, say 10^{12} of these states are important, then one can use smart methods for finding the most important say 10^9 states, diagonalizing and then include rest of 10^{12} states using perturbation theory.

2. Quantum Monte Carlo: Applicable to large finite Hilbert spaces as well as infinite Hilbert spaces!
3. DMRG: Very efficient for low-dimensional problems. Steve White, Garnet Chan

What is Quantum Monte Carlo?

Stochastic implementation of the power method for projecting out the dominant eigenvector of a matrix or integral kernel.

"Dominant state" means state with largest absolute eigenvalue.

If we repeatedly multiply an arbitrary vector, not orthogonal to the dominant state, by the matrix, we will eventually project out the dominant state.
Power method is an iterative method for eigenvalue problems (less efficient than Lanczos or Davidson). However, *stochastic* power method, QMC, is powerful.

QMC methods are used only when the number of states is so large ($> 10^{10}$) that it is not practical to store even a single vector in memory. Otherwise use exact diagonalization method, e.g., Lanczos or Davidson. At each MC generation, only a sample of the states are stored, and expectation values are accumulated.

QMC methods are used not only in a large discrete space but also in a continuously infinite space. Hence "matrix or integral kernel" above. In the interest of brevity I will use either discrete or continuous language (sums and matrices or integrals and integral kernels), but much of what is said will apply to both situations.

When to use Monte Carlo Methods

Monte Carlo methods: A class of computational algorithms that rely on repeated random sampling to compute results.

A few broad areas of applications are:

1. physics
2. chemistry
3. engineering
4. finance and risk analysis

When are MC methods likely to be the methods of choice?

1. When the problem is many-dimensional and approximations that factor the problem into products of lower dimensional problems are inaccurate.
2. A less important reason is that if one has a complicated geometry, a MC algorithm may be simpler than other choices.

Obvious drawback of MC methods: There is a statistical error.

Frequently there is a tradeoff between statistical error and systematic error (needed to overcome *sign problem*), so need to find the best compromise.

MC Simulations versus MC calculations

One can distinguish between two kinds of algorithms:

1. The system being studied is stochastic and the stochasticity of the algorithm mimics the stochasticity of the actual system. e.g. study of neutron transport and decay in nuclear reactor by following the trajectories of a large number of neutrons. Such problems are suitable for MC algorithms in a very obvious way.
2. Much more interesting are applications where the system being studied is not stochastic, but nevertheless a stochastic algorithm is the most efficient, or the most accurate, or the only feasible method for studying the system. e.g. the solution of a PDE in a large number of variables, e.g., the solution of the Schrödinger equation for an N-electron system, with say $N = 100$ or 1000. (Note: The fact that the wavefunction has a probabilistic interpretation has *nothing* to do with the stochasticity of the algorithm. The wavefunction itself is perfectly deterministic.)

I prefer to use the terminology that the former are MC simulations whereas the latter are MC calculations but not everyone abides by that terminology.

Early Recorded History of Monte Carlo

1777	Comte de Buffon: If a needle of length L is thrown at random onto a plane ruled with straight lines a distance $d (d > L)$ apart, then the probability P of the needle intersecting one of those lines is $P = \frac{2L}{\pi d}$. Laplace: This could be used to compute π (inefficiently).
1930s	First significant scientific application of MC: Enrico Fermi used it for neutron transport in fissile material. Segre: "Fermi took great delight in astonishing his Roman colleagues with his "too-good-to-believe" predictions of experimental results."
1940s	Monte Carlo named by Nicholas Metropolis and Stanislaw Ulam
1953	Algorithm for sampling any probability density Metropolis, Rosenbluth, Rosenbluth, Teller and Teller (generalized by Hastings in 1970)
1962,1974	First PMC calculations, Kalos, and, Kalos, Levesque, Verlet.
1965	First VMC calculations (of liquid He), Bill McMillan.

Compte de Buffon

I gave a series of lectures at the University of Paris.
After my first lecture, my host, Julien Toulouse, took me for a short walk to the Jardin de Plantes

Compte de Buffon

I gave a series of lectures at the University of Paris.
After my first lecture, my host, Julien Toulouse, took me for a short walk to the Jardin de Plantes to meet Buffon!

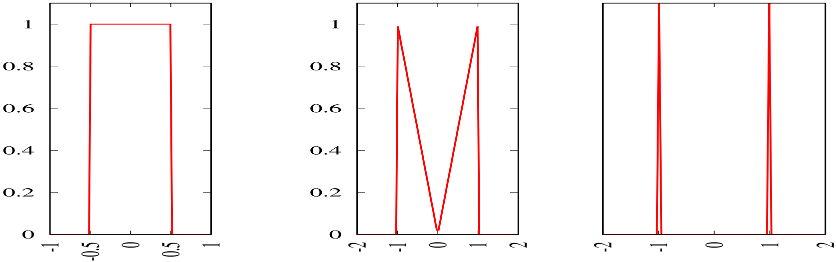
Here he is:

Among other things, he wrote a 36 volume set of books on the Natural History of the Earth!

Central Limit Theorem

de Moivre (1733), Laplace (1812), Lyapunov (1901), Pólya (1920)

Let $X_1, X_2, X_3, \cdots, X_N$ be a sequence of N independent random variables sampled from a probability density function with a finite expectation value, μ, and variance σ^2. The central limit theorem states that as the sample size N increases, the probability density of the sample average, $\bar{X}$, of these random variables approaches the normal distribution, $\sqrt{\frac{N}{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2/N)}$, with mean μ, and variance σ^2/N, irrespective of the original probability density function, e.g.:



The rate at which they converge will however depend on the original PDF.

(Weak) Law of Large Numbers

Cardano, Bernouli, Borel, Cantelli, Kolmogorov, Khinchin

Let $X_1, X_2, X_3, \cdots, X_N$ be a sequence of N independent random variables sampled from a probability density function with a finite expectation value, μ, but not necessarily a finite variance σ^2. Then for any $\epsilon > 0$,

$$\lim_{N\to\infty} P(|\bar{X} - \mu| \geq \epsilon) = 0$$

However, the rate at which it converges may be very slow.
So, employ distributions with a finite variance whenever possible.

Lorentzian

Does the Central Limit Theorem or the Law of Large Numbers apply to a Lorentzian (also known as Cauchy) probability density function

$$L(x) \quad = \quad \frac{1}{\pi}\frac{1}{1+x^2}?$$

Lorentzian

A Lorentzian (also known as Cauchy) probability density function

$$L(x) \quad = \quad \frac{1}{\pi}\frac{1}{1+x^2}$$

not only violates the conditions for the Central Limit Theorem but also the conditions for the Law of Large Numbers, since not only the variance but even the mean is undefined.

$$\begin{aligned}\int_{-\infty}^{\infty} xL(x)dx \quad &= \quad \left(\int_{-\infty}^{a} + \int_{a}^{\infty}\right) xL(x)dx \\ &= \quad -\infty + \infty\end{aligned}$$

Averages over a Lorentzian have the same spread of values as the original values!
So, although the Lorentzian looks much "nicer" than the other 3 functions we showed, it is a problem!

Chebychev Inequality

The Central Limit Theorem by itself does not tell you how quickly the averages converge to a Gaussian distribution.

If we have not averaged enough, for an arbitrary distribution with finite mean μ and finite variance σ^2, we have much weaker bounds given by Chebychev's inequality:
The probability of a variable lying between $\mu - n\sigma$ and $\mu + n\sigma$ is $> 1 - 1/n^2$, as compared to $\mathrm{erf}(n/\sqrt{2})$ for a Gaussian.

Prob. of being within 1σ of μ is $\geq$ 0% versus 68.3% for Gaussian
Prob. of being within 2σ of μ is $\geq$ 75% versus 95.4% for Gaussian
Prob. of being within 3σ of μ is $\geq$ 89% versus 99.7% for Gaussian
Prob. of being within 4σ of μ is $\geq$ 94% versus 99.994% for Gaussian

The worst case occurs for a distribution with probability $1 - 1/n^2$ at μ and probability $1/2n^2$ at $\mu - n\sigma$ and $\mu + n\sigma$.

What if the population variance $\sigma^2 = \infty$ but we do not know that beforehand? The computed sample variance will ofcourse always be finite. The practical signature of an infinite variance estimator is that the estimated σ increases with sample size, N and tends to have upward jumps. So the estimated error of the sample mean, $\sigma_N = \sigma/\sqrt{N}$, goes down more slowly than $\frac{1}{\sqrt{N}}$, or even does not go down at all.

Monte Carlo versus Deterministic Integration methods

Deterministic Integration Methods:

Integration Error, ϵ, using $N_{\rm int}$ integration points:
1-dim Simpson rule: $\epsilon \leq cN_{\rm int}^{-4}$, (provided derivatives up to 4^{th} exist)
d-dim Simpson rule: $\epsilon \leq cN_{\rm int}^{-4/d}$, (provided derivatives up to 4^{th} exist)
This argument is correct for functions that are approximately separable.

Monte Carlo:

$\epsilon \sim \sigma(T_{\rm corr}/N_{\rm int})^{1/2}$, independent of dimension!, according to the central limit theorem since width of gaussian decreases as $(T_{\rm corr}/N_{\rm int})^{1/2}$ provided that the variance of the integrand is finite. ($T_{\rm corr}$ is the autocorrelation time.)

Very roughly, Monte Carlo becomes advantageous for $d > 8$.
For $d = 50$, even 2 grid points per dimensions gives $N_{\rm int} \approx 10^{15}$, so deterministic integration not possible.
For a many-body wavefunction $d = 3N_{\rm elec}$ and can be a few thousand!

Scaling with number of electrons

Simpson's rule integration

$$\epsilon \leq \frac{c}{N_{\text{int}}^{4/d}} = \frac{c}{N_{\text{int}}^{4/3N_{\text{elec}}}}$$

$$N_{\text{int}} \leq \left(\frac{c}{\epsilon}\right)^{\frac{3N_{\text{elec}}}{4}} \quad \text{exponential in } N_{\text{elec}}$$

Monte Carlo integration

$$\epsilon = \sigma\sqrt{\frac{N_{\text{elec}}}{N_{\text{MC}}}}$$

$$N_{\text{MC}} = \left(\frac{\sigma}{\epsilon}\right)^2 N_{\text{elec}} \quad \text{linear in } N_{\text{elec}}$$

(For both methods, computational cost is higher than this since the cost of evaluating the wavefunction increases with N_{elec}, e.g., as N_{elec}^3, (better if one uses "linear scaling"; worse if one increases N_{det} with N_{elec}.))

Monte Carlo Integration

$$I = \int_V f(x)dx = V\,\overline{f} \pm V\sqrt{\frac{\overline{f^2} - \overline{f}^2}{N-1}}$$

where $$\overline{f} = \frac{1}{N}\sum_i^N f(x_i), \quad \overline{f^2} = \frac{1}{N}\sum_i^N f^2(x_i)$$

and the points x_i are sampled uniformly in V. Many points may contribute very little.

Monte Carlo Integration

$$I = \int_V f(x)dx = V\,\overline{f} \pm V\sqrt{\frac{\overline{f^2} - \overline{f}^2}{N-1}}$$

where $\quad \overline{f} = \frac{1}{N}\sum_i^N f(x_i), \quad \overline{f^2} = \frac{1}{N}\sum_i^N f^2(x_i)$

and the points x_i are sampled uniformly in V. Many points may contribute very little.

Importance sampling

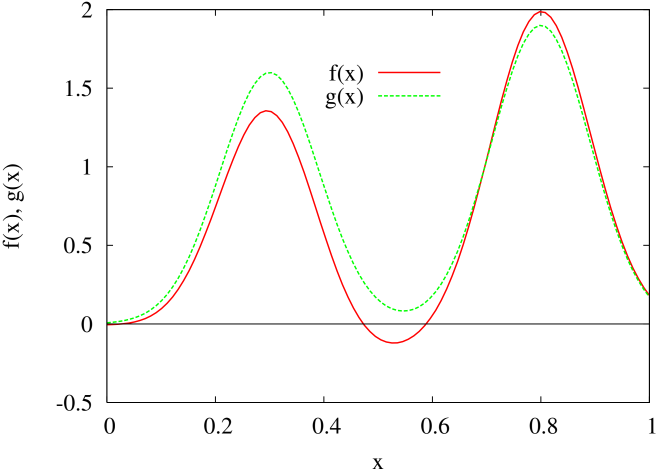
$$I = \int_V g(x)\frac{f(x)}{g(x)}dx = \overline{\left(\frac{f}{g}\right)} \pm \sqrt{\frac{\overline{\left(\frac{f}{g}\right)^2} - \overline{\left(\frac{f}{g}\right)}^2}{N-1}}$$

where the probability density function $g(x) \geq 0$ and $\int_V g(x)dx = 1$.
If $g(x) = 1/V$ in V then we recover original fluctuations but if $g(x)$ mimics $f(x)$ then the fluctuations are much reduced. Optimal g is $|f|$. Need: a) $g(x) \geq 0$, b) know integral of $g(x)$, and, c) be able to sample it.

Importance sampling can turn an $\infty-$variance estimator into a finite variance one!

Illustration of Importance Sampling

$f(x)$ is the function to be integrated. $g(x)$ is a function that is "similar" to $f(x)$ and has the required properties: a) $g(x) \geq 0$, b) $\int dx\, g(x) = 1$, and, c) we know how to sample it. $\int f(x)dx$ can be evaluated efficiently by sampling $g(x)$ and averaging $f(x)/g(x)$.



Infinite variance estimators

When variance σ^2 is finite, by the central limit theorem the average

$$F_N = \frac{\sum_{i=1}^{N} f(x_i)}{N}$$

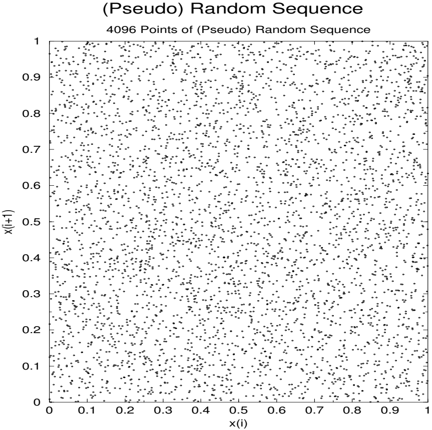
converges for increasing N to a gaussian of width $\sigma_N = \sigma/\sqrt{N}$.
Since we have a gaussian distribution the probability of F_N lying between $\mu - n\sigma_N$ and $\mu + n\sigma_N$ is $\mathrm{erf}(n/\sqrt{2})$

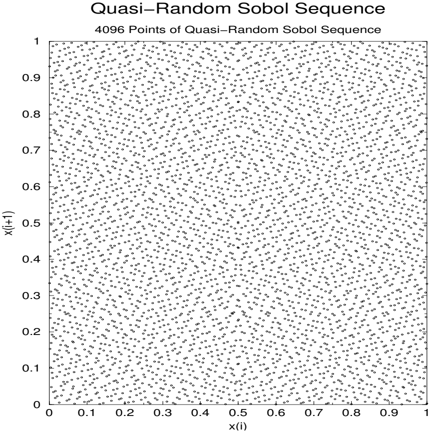
F_N being within $1\sigma_N$ of the true mean is 68.3%
F_N being within $2\sigma_N$ of the true mean is 95.4%
F_N being within $3\sigma_N$ of the true mean is 99.7%.

What if the population variance $\sigma^2 = \infty$ but we do not know that beforehand? The computed sample variance will ofcourse always be finite. The practical signature of an infinite variance estimator is that the estimated σ increases with sample size, N and tends to have upward jumps. So the estimated error of the sample mean, $\sigma_N = \sigma/\sqrt{N}$, goes down more slowly than $\frac{1}{\sqrt{N}}$, or even does not go down at all.

Pseudo-random vs quasi-random numbers

Terrible misnomers!





Reason why uniform grid is inefficient: Projection of $N = n^d$ points in d dimensions onto a line maps n^{d-1} points onto a single point.

Reason why quasi-MC is more efficient than pseudo-MC in intermediate # of dimensions (e.g. finance applications): Quasi-MC avoids clusters and voids.

Negatives for quasi-MC: Difficult to combine with importance sampling (needed for spiky functions), cannot choose # of MC points freely.

Random Number Generators

Conventional random number generators generate random numbers uniformly distributed on [0,1).

Of course no computer generated sequence of random numbers is truly random. For one, the random numbers must repeat after a finite (though hopefully very large) period. Also, if N bits are used to represent the random numbers, then the number of different numbers generated can by no larger than 2^N.

Note however, that the period can be (and typically is for the better generators) much larger than 2^N.

Many different algorithms exist for generating random numbers, e.g., linear congruential generators (with or without an additive constant), linear feedback shift register, lagged Fibonacci generator, XORshift algorithm etc. They are typically subjected to a battery of statistical tests, e.g., the Diehard tests of Marsaglia. Of course no random number generator can pass all the tests that one can invent, but hopefully the random number generator used does not have correlations that could significantly impact the system being studied.

Random Number Generators

For many MC calculations it is the short-ranged correlations that matter most, but one has to think for each application what is important. For example, if one were studying an Ising model with a power of two number of spins, it would be problematic to have random number generator that generated numbers with bits that repeat at an interval of 2^N.

In the old days, there were quite a few calculations that produced inaccurate results due to bad random number generators. For example, the standard generators that came with UNIX and with C were badly flawed. In the 1980s a special purpose computer was built at Santa Barbara to study the 3-D Ising model. However, at first it failed to reproduce the known exact results for the 2-D Ising model and that failure was traced back to a faulty random number generator. Fortunately, these days the standard random number generators are much more reliable.

Sampling random variables from nonuniform probability density functions

We say x is sampled from $f(x)$ if for any a and b in the domain,

$$\mathrm{Prob}[a \le x \le b] \quad = \quad \int_a^b dx' \, f(x')$$

1) Transformation method (For many simple functions)
2) Rejection method (For more complicated functions)
3) Metropolis-Hastings method (For any function)

1) Transformation method: Perform a transformation $x(\xi)$ on a uniform deviate ξ, to get x sampled from desired probability density $f(x)$.

$$|\mathrm{Prob}(\xi) d\xi| \quad = \quad |\mathrm{Prob}(x) dx| \quad \text{conservation of probability}$$

If we have sampled ξ from a uniform density $(\mathrm{Prob}(\xi) = 1)$ and we wish x to be sampled from the desired density, $f(x)$, then setting $\mathrm{Prob}(x) = f(x)$,

$$\left| \frac{d\xi}{dx} \right| \quad = \quad f(x)$$

Solve for $\xi(x)$ and invert to get $x(\xi)$, i.e., invert the cumulative distribution.

Examples of Transformation Method

Example 1: $f(x) = ae^{-ax}$, $x \in [0, \infty)$

$$\left|\frac{d\xi}{dx}\right| = ae^{-ax}, \quad \text{or,} \quad \xi = e^{-ax}, \quad \text{i.e.,} \quad \boxed{x = \frac{-\ln(\xi)}{a}}$$

Examples of Transformation Method

Example 1: $f(x) = ae^{-ax}$, $x \in [0, \infty)$

$$\left|\frac{d\xi}{dx}\right| = ae^{-ax}, \quad \text{or,} \quad \xi = e^{-ax}, \quad \text{i.e.,} \quad \boxed{x = \frac{-\ln(\xi)}{a}}$$

Example 2: $f(x) = \frac{x^{-1/2}}{2}$, $x \in [0, 1]$

$$\left|\frac{d\xi}{dx}\right| = \frac{x^{-1/2}}{2}, \quad \text{or} \quad \xi = x^{1/2}, \quad \text{i.e.,} \quad \boxed{x = \xi^2}$$

Note that in this case we are sampling a probability density that is infinite at 0, but that is OK!

Examples of Transformation Method

Example 1: $f(x) = ae^{-ax}$, $x \in [0, \infty)$

$$\left|\frac{d\xi}{dx}\right| = ae^{-ax}, \quad \text{or,} \quad \xi = e^{-ax}, \quad \text{i.e.,} \quad \boxed{x = \frac{-\ln(\xi)}{a}}$$

Example 2: $f(x) = \frac{x^{-1/2}}{2}$, $x \in [0, 1]$

$$\left|\frac{d\xi}{dx}\right| = \frac{x^{-1/2}}{2}, \quad \text{or} \quad \xi = x^{1/2}, \quad \text{i.e.,} \quad \boxed{x = \xi^2}$$

Note that in this case we are sampling a probability density that is infinite at 0, but that is OK!

Example 3: $f(x) = xe^{-x^2/2}$, $x \in [0, \infty)$

$$\left|\frac{d\xi}{dx}\right| = xe^{-x^2/2}, \quad \text{or,} \quad \xi = e^{-x^2/2}, \quad \text{i.e.,} \quad \boxed{x = \sqrt{-2\ln(\xi)}}$$

Examples of Transformation Method

Example 4a: $f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$, $x \in (-\infty, \infty)$ (using Box-Müller method)

$$\frac{1}{2\pi} e^{-(\frac{x_1^2}{2} + \frac{x_2^2}{2})}\, dx_1\, dx_2 \quad = \quad \left(r\, e^{-\frac{r^2}{2}}\, dr\right) \left(\frac{d\phi}{2\pi}\right)$$

$$r = \sqrt{-2\log(\xi_1)}, \qquad \phi = 2\pi\xi_2$$

$$\boxed{x_1 = \sqrt{-2\log(\xi_1)}\cos(2\pi\xi_2)}, \qquad \boxed{x_2 = \sqrt{-2\log(\xi_1)}\sin(2\pi\xi_2)}$$

(x_1 and x_2 are uncorrelated)

Examples of Transformation Method

Example 4a: $f(x) = \frac{e^{-x^2/2}}{\sqrt{2\pi}}$, $x \in (-\infty, \infty)$ (using Box-Müller method)

$$\frac{1}{2\pi} e^{-(\frac{x_1^2}{2} + \frac{x_2^2}{2})} \, dx_1 \, dx_2 \quad = \quad \left(r \, e^{-\frac{r^2}{2}} \, dr \right) \left(\frac{d\phi}{2\pi} \right)$$

$$r = \sqrt{-2\log(\xi_1)}, \qquad \phi = 2\pi\xi_2$$

$$\boxed{x_1 = \sqrt{-2\log(\xi_1)}\cos(2\pi\xi_2)}, \qquad \boxed{x_2 = \sqrt{-2\log(\xi_1)}\sin(2\pi\xi_2)}$$

(x_1 and x_2 are uncorrelated)

Example 4b: $f(x) \approx \frac{e^{-x^2/2}}{\sqrt{2\pi}}$, $x \in (-\infty, \infty)$ (using central-limit theorem)

$\xi - 0.5$ is in $[-1/2, 1/2]$. Since σ^2 for uniform distribution about 0 is $\int_{-1/2}^{1/2} dx \, x^2 = \frac{1}{12}$

$$x = \lim_{N \to \infty} \sqrt{\frac{12}{N}} \left(\sum_{i=1}^{N} \xi_i - \frac{N}{2} \right) \approx \boxed{\sum_{i=1}^{12} \xi_i - 6}$$

(avoids log, sqrt, cos, sin, but, misses tiny tails beyond ± 6)

Rejection Method

We wish to sample $f(x)$.
Find a function $g(x)$ that can be sampled by another method (say transformation) and that preferably mimics the behaviour of $f(x)$.
Let $C \geq \max(f(x)/g(x))$.
Then $f(x)$ is sampled by sampling $g(x)$ and keep the sampled points with probability

$$P = \frac{f(x)}{Cg(x)}$$

The efficiency of the method is the fraction of the sampled points that are kept.

$$\begin{aligned} Eff &= \int dx \frac{f(x)}{Cg(x)} g(x) \\ &= \frac{1}{C} \end{aligned}$$

Drawback: It is often hard to know C and a "safe" upperbound choice for C may lead to low efficiency. An alternative is to associate weights with the sampled points.

Sampling from Discrete Distributions

Suppose we need to repeatedly sample from N discrete events with probabilities $p_1, p_2, \cdots p_N$, where N is large.

What is the best possible scaling of the time per sample?

Is it $\mathcal{O}(N), \quad \mathcal{O}(\log_2(N)), \quad \mathcal{O}(1)$?

Sampling from Discrete Distributions

Suppose we need to repeatedly sample from N discrete events with probabilities $p_1, p_2, \cdots p_N$, where N is large.

What is the best possible scaling of the time per sample?

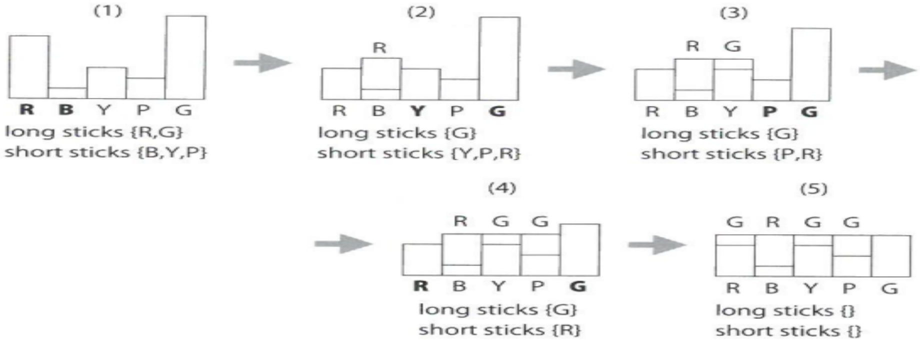
Is it $\mathcal{O}(N)$, $\quad \mathcal{O}(\log_2(N))$, $\quad \mathcal{O}(1)$?

Straightforward $\mathcal{O}(\log_2(N))$ method with binary search:

1. Before starting sampling, construct array of cumulative probabilities.
2. Draw a random number, ξ, in $[0, 1]$.
3. Do a binary search to find the interval it falls in.

Can we do it in $\mathcal{O}(1)$?

Sampling from Discrete Distributions: $\mathcal{O}(1)$ Alias Method



1. Before starting sampling, construct an integer array, $\{A_i\}$, that contains the aliases and a real array, $\{p_i\}$ that contains the probabilities of staying at i.
2. Draw a random number in $[0, 1]$.
3. Go to the $i = \lceil N\xi \rceil$ bin.
4. With probability p_i sample i and with probability $(1 - p_i)$ sample A_i.

Figure taken from book by Gubernatis, Kawashima and Werner

Expectation Values

Interested in calculating expection values (ensemble averages) of a variable X with respect to a probability density ρ, (discrete or continuous).

$$\langle X \rangle_\rho = \frac{\sum_{\mathbf{R}} X(\mathbf{R})\rho(\mathbf{R})}{\sum_{\mathbf{R}} \rho(\mathbf{R})} \approx \frac{1}{T}\sum_{i=1}^{T} X(\mathbf{R}_i)$$

with configurations $\mathbf{R}_i$ sampled from $\rho(\mathbf{R})/\sum_{\mathbf{R}} \rho(\mathbf{R})$.
Ensemble average approximated by time average.
Equality when $T \to \infty$.
We need a means to sample ρ.

Importance Sampling for computing integrals efficiently

Now that we know how to sample simple probability density functions, we study how to use *importance sampling* to compute integrals more efficiently.

Example of Importance Sampling to Calculate Integrals More Efficiently

Suppose we wish to compute

$$\int_0^1 dx\, f(x) \;=\; \int_0^1 dx\, \frac{1}{x^p + x} \left(= \frac{\log\left(\frac{x+x^p}{x^p}\right)}{1-p}\Bigg|_0^1 = \frac{\log(2)}{1-p}, \quad \text{but pretend not known} \right.$$

Note that

$$\int_0^1 dx(f(x))^2 \;=\; \infty, \quad (\text{for } p \geq 0.5)$$

so if we estimate the integral by sampling points uniformly in $[0,1]$ then this would be an infinite variance estimator and the error of the estimate will go down more slowly than $N^{-1/2}$. However, we can instead sample points from the density

$$g(x) \;=\; \frac{1-p}{x^p}$$

Now the variance of $f(x)/g(x)$ is finite and the error decreases as $N^{-1/2}$, and, with a small prefactor. (Still would not use this in 1D.)

Homework Problem 1

Compute

$$I = \int_0^1 dx\, f(x) = \int_0^1 dx\, \frac{1}{x^p + x} \quad \left(= \frac{\log(2)}{1-p}, \text{ but pretend not known}\right) \approx \frac{1}{N_{\rm MC}} \sum_{k=1}^{N_{\rm MC}} \frac{1}{\xi_k^p + \xi_k}$$

with/without importance sampling, using for the importance sampling function

$$g(x) = \frac{(1-p)}{x^p}$$

To sample $g(x)$: $$\left|\frac{d\xi}{dx}\right| = (1-p)x^{-p}, \quad \text{i.e.,} \quad \xi = x^{1-p}, \quad \text{i.e.,} \quad \boxed{x = \xi^{\frac{1}{1-p}}}$$

$$\begin{aligned}\int_0^1 dx\, f(x) &= \int_0^1 dx\, g(x) \frac{f(x)}{g(x)} = \int_0^1 dx\, \frac{1-p}{x^p} \frac{1}{(1-p)(1+x^{1-p})} \\ &\approx \frac{1}{N_{\rm MC}(1-p)} \sum_{k=1}^{N_{\rm MC}} \frac{1}{(1+x_k^{1-p})} = \frac{1}{N_{\rm MC}(1-p)} \sum_{k=1}^{N_{\rm MC}} \frac{1}{(1+\xi_k)}\end{aligned}$$

Do this for $p = 0.25, 0.5, 0.75, 0.95$ and $N_{\rm MC} = 10^3, 10^4, 10^5, 10^6, 10^7, 10^8, 10^9$.
Plot 2 graphs, each having 8 curves (4 values of p, and, with/without importance sampling):

1. Log of estimated 1-standard deviation statistical error versus $\log(N_{\rm MC})$.
2. Actual error in I, with estimated 1-std. dev. statistical error as an error bar versus $\log(N_{\rm MC})$.

Unbiased Estimators

Population mean: $\langle f \rangle$

Sample (of size N) mean: $\bar{f}$

$\tilde{F}(\bar{f})$ is an unbiased estimator if $\left\langle \tilde{F}(\bar{f}) \right\rangle = F(\langle f \rangle)$
or more generally
$\tilde{F}(\bar{f}_1, \bar{f}_2, \cdots)$ is an unbiased estimator if $\left\langle \tilde{F}(\bar{f}_1, \bar{f}_2, \cdots) \right\rangle = F(\langle f_1 \rangle, \langle f_2 \rangle, \cdots)$

1) Is $\left\langle \bar{f} - \bar{g} \right\rangle = \langle f \rangle - \langle g \rangle$?

2) Is $\left\langle \bar{f}\bar{g} \right\rangle = \langle f \rangle \langle g \rangle$?

3) Is $\left\langle \bar{f}/\bar{g} \right\rangle = \langle f \rangle / \langle g \rangle$?

4) Is $\left\langle \bar{f^2} - \bar{f}^2 \right\rangle = \left\langle f^2 \right\rangle - \langle f \rangle^2$?

Unbiased Estimators

Population mean: $\langle f \rangle$

Sample (of size N) mean: $\bar{f}$

$\tilde{F}(\bar{f})$ is an unbiased estimator if $\left\langle \tilde{F}(\bar{f}) \right\rangle = F(\langle f \rangle)$
or more generally
$\tilde{F}(\bar{f}_1, \bar{f}_2, \cdots)$ is an unbiased estimator if $\left\langle \tilde{F}(\bar{f}_1, \bar{f}_2, \cdots) \right\rangle = F(\langle f_1 \rangle, \langle f_2 \rangle, \cdots)$

1) Is $\left\langle \bar{f} - \bar{g} \right\rangle = \langle f \rangle - \langle g \rangle$? yes

2) Is $\left\langle \bar{f}\bar{g} \right\rangle = \langle f \rangle \langle g \rangle$? no

3) Is $\left\langle \bar{f}/\bar{g} \right\rangle = \langle f \rangle / \langle g \rangle$? no

4) Is $\left\langle \overline{f^2} - \bar{f}^2 \right\rangle = \left\langle f^2 \right\rangle - \langle f \rangle^2$? no. Correct: $\frac{N}{N-1} \left\langle \overline{f^2} - \bar{f}^2 \right\rangle = \left\langle f^2 \right\rangle - \langle f \rangle^2$

Unbiased Estimators

In this section, denote population means by brackets and sample means by bars.
Let $f(x)$ be a random variable with probability density $\rho(x)$.

$$\text{Population mean:} \quad \mu \quad = \quad \int dx\, f(x)\, \rho(x) \;\equiv\; \langle f \rangle_\rho$$

$$\text{Population variance:} \quad \sigma^2 \quad = \quad \int dx\, (f(x) - \langle f \rangle_\rho)^2\, \rho(x) = \langle f^2 \rangle_\rho - \langle f \rangle_\rho^2$$

Estimator is unbiased if averaging over an infinite number of samples of size N gives the same result as that from a single infinite sample. Can estimate mean and variance from independent finite samples of size N, but the "obvious" estimators often have $\mathcal{O}(1/N)$ errors, so we provide here estimators that are correct at least to $\mathcal{O}(1/N)$.

$$\text{Unbiased estimator for} \quad \langle f \rangle_\rho \quad : \quad \frac{1}{N} \sum_{i=1}^{N} f(x_i) \;\equiv\; \overline{f_\rho}$$

$$\text{Unbiased estimators for} \quad \sigma^2 \quad : \quad \frac{1}{N} \sum_{i=1}^{N} (f(x_i) - \langle f \rangle_\rho)^2 \quad \text{(but } \langle f \rangle_\rho \text{ not known)}$$

$$\text{and} \qquad \frac{1}{N-1} \sum_{i=1}^{N} (f(x_i) - \overline{f_\rho})^2 = \frac{N}{N-1} \left(\overline{f_\rho^2} - \overline{f_\rho}^2 \right)$$

Estimating Unbiased Variance from Uncorrelated Samples

Let $\langle f(x)\rangle$ denote the population mean and $\bar{f}(x)$ denote the sample mean.
Then $\overline{f^2} - (\bar{f})^2 =$

$$\left\langle \frac{\sum_i f^2(x_i)}{N} - \left[\frac{\sum_i f(x_i)}{N}\right]^2 \right\rangle = \langle f^2\rangle - \left\langle \frac{\sum_i f^2(x_i) + \sum_{i,j\neq i}\sum_j f(x_i)f(x_j)}{N^2} \right\rangle$$

Since $f(x_i)$ and $f(x_j)$ are independent

$$RHS = \left(1 - \frac{1}{N}\right)\langle f^2\rangle - \frac{N(N-1)}{N^2}\langle f\rangle^2 \; = \; \frac{N-1}{N}(\langle f^2\rangle - \langle f\rangle^2) \; = \; \frac{N-1}{N}\sigma^2$$

So, the unbiased estimate for σ^2 is

$$\boxed{\sigma^2 \approx \frac{N}{N-1}\left(\overline{f^2} - (\bar{f})^2\right)}$$

Loss of one degree of freedom because sample variance is computed relative to sample mean rather than the true mean.

Examples of Unbiased and Biased Estimators

$$E_T = \frac{\int d\mathbf{R}\, \psi_{\rm T}(\mathbf{R})\, \mathcal{H}\, \psi_{\rm T}(\mathbf{R})}{\int d\mathbf{R}\, \psi^2_{\rm T}(\mathbf{R})} = \int d\mathbf{R}\; \frac{\psi^2_{\rm T}(\mathbf{R})}{\int d\mathbf{R}\, \psi^2_{\rm T}(\mathbf{R})}\; \frac{\mathcal{H}\psi_{\rm T}(\mathbf{R})}{\psi_{\rm T}(\mathbf{R})}$$

$$= \frac{1}{N}\sum_{i=1}^{N} \frac{\mathcal{H}\Psi_{\rm T}(\mathbf{R}_i)}{\Psi_{\rm T}(\mathbf{R}_i)} = \frac{1}{N}\sum_{i=1}^{N} E_L(\mathbf{R}_i) \qquad \text{unbiased}$$

$$E_T = \frac{\int d\mathbf{R}\, \psi_{\rm T}(\mathbf{R})\, \mathcal{H}\, \psi_{\rm T}(\mathbf{R})}{\int d\mathbf{R}\, \psi^2_{\rm T}(\mathbf{R})} = \frac{\int d\mathbf{R}\; \frac{|\psi_{\rm T}(\mathbf{R})|}{\int d\mathbf{R}\, |\psi_{\rm T}(\mathbf{R})|}\; \mathrm{sgn}(\psi_{\rm T}(\mathbf{R}))\, \mathcal{H}\psi_{\rm T}(\mathbf{R})}{\int d\mathbf{R}\; \frac{|\psi_{\rm T}(\mathbf{R})|}{\int d\mathbf{R}\, |\psi_{\rm T}(\mathbf{R})|}\; |\psi_{\rm T}(\mathbf{R})|}$$

$$= \frac{\sum_{i=1}^{N}\; \mathrm{sgn}(\psi_{\rm T}(\mathbf{R}))\, \mathcal{H}\Psi_{\rm T}(\mathbf{R}_i)}{\sum_{i=1}^{N} |\psi_{\rm T}(\mathbf{R})|} \qquad \mathcal{O}\left(\frac{1}{N}\right) \text{ bias}$$

Can do better by calculating covariances.

Unbiased Estimators to $\mathcal{O}(1/N)$ of functions of expectation values and their variance

$\langle x \rangle$ $\equiv$ population averages of x, i.e., true expectation value
$\bar{x}$ $\equiv$ average of x over sample of size N

Let F be a function of expectation values, $\{\langle f_i \rangle\}$.
$F\left(\{\bar{f}_i\}\right)$ is unbiased estimator for $F(\{\langle f_i \rangle\})$ iff F is linear function of $\{\langle f_i \rangle\}$.

In general

$$
\begin{aligned}
F(\{\langle f_i \rangle\}) &= \left\langle F\left(\{\bar{f}_i\}\right)\right\rangle - \frac{1}{2}\sum_{i,j} \frac{\partial^2 F}{\partial f_i \partial f_j} \, \frac{\mathrm{cov}(f_i, f_j)}{N} \, + \, \mathcal{O}\left(\frac{1}{N^2}\right) \\
\mathrm{var}\big(F(\{\langle f_i \rangle\})\big) &= \sum_{i,j} \frac{\partial F}{\partial f_i}\frac{\partial F}{\partial f_j} \, \mathrm{cov}(f_i, f_j) \, + \, \mathcal{O}\left(\frac{1}{N}\right)
\end{aligned}
$$

Unbiased Estimators to $\mathcal{O}(1/N)$ or better (cont)

$$
\begin{aligned}
\text{Estim. for mean} \quad \langle f\rangle_\rho &= \overline{f_\rho} \\
\text{Estim. for variance} \quad \langle f^2\rangle_\rho - \langle f\rangle_\rho^2 &= \frac{N}{N-1}\left(\overline{f_\rho^2} - \overline{f_\rho}^2\right) \\
\text{Estim. for error of sample mean} \quad &= \sqrt{\frac{1}{N-1}\left(\overline{f_\rho^2} - \overline{f_\rho}^2\right)} \\
\text{Estim. for covariance} \quad \mathrm{cov}(f,g) \equiv \langle fg\rangle_\rho - \langle f\rangle_\rho\langle g\rangle_\rho &= \frac{N}{N-1}\left(\overline{fg_\rho} - \overline{f_\rho}\,\overline{g_\rho}\right) \\
\text{HW-2 Estim. for product of expec. values} \quad \langle f\rangle_\rho\langle g\rangle_\rho &= \frac{\overline{f_\rho}\,\overline{g_\rho}}{\left[1 + \frac{1}{N}\frac{\mathrm{cov}(f,g)}{\langle f\rangle_\rho\langle g\rangle_\rho}\right]} \\
\text{HW-2 Estim. for ratio of expec. values} \quad \frac{\langle f\rangle_\rho}{\langle g\rangle_\rho} &\approx \frac{\overline{f_\rho}/\overline{g_\rho}}{\left[1 + \frac{1}{N}\left(\frac{\sigma_g^2}{\langle g\rangle_\rho^2} - \frac{\mathrm{cov}(f,g)}{\langle f\rangle_\rho\langle g\rangle_\rho}\right)\right]} \\
\mathrm{Var}\left(\overline{f_\rho}\;\overline{g_\rho}\right) &= \frac{1}{N}\langle f\rangle_\rho^2\langle g\rangle_\rho^2\left(\frac{\sigma_f^2}{\langle f\rangle_\rho^2} + \frac{\sigma_g^2}{\langle g\rangle_\rho^2} + 2\frac{\mathrm{cov}(f,g)}{\langle f\rangle_\rho\langle g\rangle_\rho}\right) \\
\mathrm{Var}\left(\frac{\overline{f_\rho}}{\overline{g_\rho}}\right) &= \frac{1}{N}\frac{\langle f\rangle_\rho^2}{\langle g\rangle_\rho^2}\left[\frac{\sigma_f^2}{\langle f\rangle_\rho^2} + \frac{\sigma_g^2}{\langle g\rangle_\rho^2} - 2\frac{\mathrm{cov}(f,g)}{\langle f\rangle_\rho\langle g\rangle_\rho}\right].
\end{aligned}
$$

Note that the product, $\overline{f_\rho}\,\overline{g_\rho}$ is unbiased if $\mathrm{cov}(f,g)=0$, but the ratio $\frac{\overline{f_\rho}}{\overline{g_\rho}}$ has $\mathcal{O}(1/N)$ bias even if $\mathrm{cov}(f,g)=0$. The ratio has no bias (and no fluctuations) when f and g are perfectly correlated.
In practice replace population means by sample means on RHS.

Unbiased Estimators to $\mathcal{O}(1/N)$ or better (cont)

$$
\begin{aligned}
\text{Estim. of mean} \quad \langle f\rangle_\rho &= \overline{f_\rho} \\
\text{Estim. of variance} \quad \langle f^2\rangle_\rho - \langle f\rangle_\rho^2 &= \frac{N}{N-1}\left(\overline{f_\rho^2} - \overline{f_\rho}^2\right) \\
\text{Estim. of error of sample mean} \quad &= \sqrt{\frac{1}{N-1}\left(\overline{f_\rho^2} - \overline{f_\rho}^2\right)} \\
\text{Estim. of covar.} \quad \mathrm{cov}(f,g) \equiv \langle fg\rangle_\rho - \langle f\rangle_\rho\langle g\rangle_\rho &= \frac{N}{N-1}\left(\overline{fg_\rho} - \overline{f_\rho}\,\overline{g_\rho}\right) \\
\text{Estim. of product of expec. values} \quad \langle f\rangle_\rho\langle g\rangle_\rho &= \overline{f_\rho}\,\overline{g_\rho} - \frac{1}{N}\mathrm{cov}(f,g) \\
\text{Estim. of ratio of expec. values} \quad \frac{\langle f\rangle_\rho}{\langle g\rangle_\rho} &\approx \frac{\overline{f_\rho}}{\overline{g_\rho}}\left[1 - \frac{1}{N}\left(\frac{\sigma_g^2}{\langle g\rangle_\rho^2} - \frac{\mathrm{cov}(f,g)}{\langle f\rangle_\rho\langle g\rangle_\rho}\right)\right] \\
\mathrm{Var}\left(\overline{f_\rho}\ \overline{g_\rho}\right) &= \frac{1}{N}\langle f\rangle_\rho^2\langle g\rangle_\rho^2\left(\frac{\sigma_f^2}{\langle f\rangle_\rho^2} + \frac{\sigma_g^2}{\langle g\rangle_\rho^2} + 2\frac{\mathrm{cov}(f,g)}{\langle f\rangle_\rho\langle g\rangle_\rho}\right) \\
\mathrm{Var}\left(\frac{\overline{f_\rho}}{\overline{g_\rho}}\right) &= \frac{1}{N}\frac{\langle f\rangle_\rho^2}{\langle g\rangle_\rho^2}\left[\frac{\sigma_f^2}{\langle f\rangle_\rho^2} + \frac{\sigma_g^2}{\langle g\rangle_\rho^2} - 2\frac{\mathrm{cov}(f,g)}{\langle f\rangle_\rho\langle g\rangle_\rho}\right].
\end{aligned}
$$

Note that the product, $\overline{f_\rho}\,\overline{g_\rho}$ is unbiased if $\mathrm{cov}(f,g)=0$, but the ratio $\frac{\overline{f_\rho}}{\overline{g_\rho}}$ has $\mathcal{O}(1/N)$ bias even if $\mathrm{cov}(f,g)=0$. The ratio has no bias (and no fluctuations) when f and g are perfectly correlated. In practice replace population means by sample means on RHS.

Unbiased Estimators of autocorrelated variables

Independent samples:

Estim. for error of sample mean $\quad \overline{\Delta_f} = \sqrt{\frac{1}{N-1}\left(\overline{f^2_\rho} - \overline{f_\rho}^2\right)}$

Autocorrelated samples (e.g. from Metropolis-Hastings):

Estim. for error of sample mean $\quad \overline{\Delta_f} = \sqrt{\frac{1}{N_{\rm eff}-1}\left(\overline{f^2_\rho} - \overline{f_\rho}^2\right)}$

where

$$N_{\rm eff} = \frac{N}{(1+2\tau_f)} \equiv \frac{N}{T_{\rm corr}}$$

$$\tau_f = \frac{\sum_{t=1}^{\infty}\left[\left\langle f_1 f_{1+t}\right\rangle_\rho - \left\langle f\right\rangle_\rho^2\right]}{\sigma_f^2}$$

If samples are indep., $\left\langle f_1 f_{1+t}\right\rangle_\rho = \left\langle f\right\rangle_\rho^2$ and integrated autocorrelation time $\tau_f = 0$.
Since the relevant quantity for MC calculations is $(1+2\tau_f) \equiv T_{\rm corr}$ we will refer to it as the autocorrelation time of f, though this is not standard usage.

Quantum Monte Carlo in a Nutshell

Definitions

Given a complete or incomplete basis: $\{|\phi_i\rangle\}$, either discrete or continuous

$$\begin{array}{llcll} \text{Exact} & |\Psi_0\rangle & = & \sum_i e_i|\phi_i\rangle, & \text{where,} \quad e_i = \langle\phi_i|\Psi_0\rangle \\ \text{Trial} & |\Psi_T\rangle & = & \sum_i t_i|\phi_i\rangle, & \text{where,} \quad t_i = \langle\phi_i|\Psi_T\rangle \\ \text{Guiding} & |\Psi_G\rangle & = & \sum_i g_i|\phi_i\rangle, & \text{where,} \quad g_i = \langle\phi_i|\Psi_G\rangle \end{array}$$

(If basis incomplete then "exact" means "exact in that basis".)

Ψ_T used to calculate variational and mixed estimators of operators $\hat{A}$, i.e., $\langle\Psi_{\rm T}|\hat{A}|\Psi_{\rm T}\rangle / \langle\Psi_{\rm T}|\Psi_{\rm T}\rangle$, $\langle\Psi_{\rm T}|\hat{A}|\Psi_0\rangle / \langle\Psi_{\rm T}|\Psi_0\rangle$

Ψ_G used to alter the probability density sampled, i.e., $\Psi_{\rm G}^2$ in VMC, $\Psi_{\rm G}\Psi_0$ in PMC.

$\Psi_{\rm G}$ must be such that $g_i \neq 0$ if $e_i \neq 0$. If $\Psi_{\rm T}$ also satisfies this condition then $\Psi_{\rm G}$ can be chosen to be $\Psi_{\rm T}$. Reasons to have $\Psi_{\rm G} \neq \Psi_{\rm T}$ are: a) rapid evaluation of "local energy", b) have finite-variance estimators. To simplify expressions, we sometimes use $\Psi_{\rm G} = \Psi_{\rm T}$ or $\Psi_{\rm G} = 1$ in what follows.

Variational MC

$$E_V = \frac{\langle \Psi_\mathrm{T}|\hat{H}|\Psi_\mathrm{T}\rangle}{\langle \Psi_\mathrm{T}|\Psi_\mathrm{T}\rangle} = \frac{\sum_{ij}^{N_\mathrm{st}} \langle \Psi_\mathrm{T}|\phi_i\rangle\ \langle \phi_i|\hat{H}|\phi_j\rangle\ \langle \phi_j|\Psi_\mathrm{T}\rangle}{\sum_{i}^{N_\mathrm{st}} \langle \Psi_\mathrm{T}|\phi_k\rangle\ \langle \phi_k|\Psi_\mathrm{T}\rangle}$$

$$= \frac{\sum_{ij}^{N_\mathrm{st}} t_i H_{ij} t_j}{\sum_{k}^{N_\mathrm{st}} t_k^2} = \sum_{i}^{N_\mathrm{st}} \frac{t_i^2}{\sum_{k}^{N_\mathrm{st}} t_k^2}\ \frac{\sum_{j}^{N_\mathrm{st}} H_{ij} t_j}{t_i}$$

$$= \sum_{i}^{N_\mathrm{st}} \frac{t_i^2}{\sum_{k}^{N_\mathrm{st}} t_k^2}\ E_\mathrm{L}(i) \ \approx\ \frac{\left[\sum_{i}^{N_\mathrm{MC}} E_\mathrm{L}(i)\right]_{\Psi_\mathrm{T}^2}}{N_\mathrm{MC}} \rightarrow_{\Psi_G \neq \Psi_T} \frac{\left[\sum_{i}^{N_\mathrm{MC}} \left(\frac{t_i}{g_i}\right)^2 E_\mathrm{L}(i)\right]_{\Psi_\mathrm{G}^2}}{\left[\sum_{k}^{N_\mathrm{MC}} \left(\frac{t_k}{g_k}\right)^2\right]_{\Psi_\mathrm{G}^2}}$$

Sample probability density function $\frac{g_i^2}{\sum_k^{N_\mathrm{st}} g_k^2}$ using Metropolis-Hastings, if Ψ_G complicated.

Value depends only on Ψ_T. Statistical error depend on Ψ_T and Ψ_G.

Energy bias and statistical error vanish as $\Psi_\mathrm{T} \rightarrow \Psi_0$.

For fixed Ψ_T, $\Psi_G = \Psi_T$ does not minimize statistical fluctuations!

In fact $\Psi_G \neq \Psi_T$ needed when optim. to get finite variance.

$\Psi_G = \Psi_T$ allows simple unbiased estimator. Ratio of expec. val. $\neq$ expec. val. of ratios.

Projector MC

Pure and Mixed estimators for energy are equal: $E_0 = \frac{\langle\Psi_0|\hat{H}|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle} = \frac{\langle\Psi_0|\hat{H}|\Psi_{\rm T}\rangle}{\langle\Psi_0|\Psi_{\rm T}\rangle}$

Projector: $|\Psi_0\rangle = \hat{P}(\infty)\,|\Psi_{\rm T}\rangle = \lim_{n\to\infty} \hat{P}^n(\tau)\,|\Psi_{\rm T}\rangle$

$$
\begin{aligned}
E_0 &= \frac{\langle\Psi_0|\hat{H}|\Psi_{\rm T}\rangle}{\langle\Psi_0|\Psi_{\rm T}\rangle} = \frac{\sum_{ij}^{N_{\rm st}} \langle\Psi_0|\phi_i\rangle\ \langle\phi_i|\hat{H}|\phi_j\rangle\ \langle\phi_j|\Psi_{\rm T}\rangle}{\sum_k^{N_{\rm st}} \langle\Psi_0|\phi_k\rangle\ \langle\phi_k|\Psi_{\rm T}\rangle} \\
&= \frac{\sum_{ij}^{N_{\rm st}} e_i H_{ij} t_j}{\sum_k^{N_{\rm st}} e_k t_k} = \sum_i^{N_{\rm st}} \frac{e_i t_i}{\sum_k^{N_{\rm st}} e_k t_k}\ \frac{\sum_j^{N_{\rm st}} H_{ij} t_j}{t_i} \\
&= \sum_i^{N_{\rm st}} \frac{e_i t_i}{\sum_k^{N_{\rm st}} e_k t_k}\, E_{\rm L}(i) \ \approx\ \frac{\left[\sum_i^{N_{\rm MC}} E_{\rm L}(i)\right]_{\Psi_{\rm T}\Psi_0}}{N_{\rm MC}} \xrightarrow{}_{\Psi_G\neq\Psi_T} \frac{\left[\sum_i^{N_{\rm MC}} \left(\frac{t_i}{g_i}\right) E_{\rm L}(i)\right]_{\Psi_{\rm G}\Psi_0}}{\left[\sum_k^{N_{\rm MC}} \left(\frac{t_k}{g_k}\right)\right]_{\Psi_{\rm G}\Psi_0}}
\end{aligned}
$$

Sample $e_i g_i / \sum_k^{N_{\rm st}} e_k g_k$ using *importance-sampled* projector.

For exact PMC, value indep. of $\Psi_{\rm T}$, $\Psi_{\rm G}$, statistical error depends on $\Psi_{\rm T}$, $\Psi_{\rm G}$.

Statistical error vanishes as $\Psi_{\rm T} \to \Psi_0$.

For fixed Ψ_T, $\Psi_G = \Psi_T$ does not minimize statistical fluctuations!

Variational and Projector MC

$$E_V = \frac{\left[\sum_i^{N_{\rm MC}} \left(\frac{t_i}{g_i}\right)^2 E_{\rm L}(i)\right]_{\Psi_{\rm G}^2}}{\left[\sum_k^{N_{\rm MC}} \left(\frac{t_k}{g_k}\right)^2\right]_{\Psi_{\rm G}^2}} \quad \text{(Value depends on } \Psi_{\rm T}\text{, error } \Psi_{\rm T}, \Psi_{\rm G}\text{)}$$

$$E_0 = \frac{\left[\sum_i^{N_{\rm MC}} \left(\frac{t_i}{g_i}\right) E_{\rm L}(i)\right]_{\Psi_{\rm G}\Psi_0}}{\left[\sum_k^{N_{\rm MC}} \left(\frac{t_k}{g_k}\right)\right]_{\Psi_{\rm G}\Psi_0}} \quad \text{(Value exact}^{\dagger}\text{. Error depends on } \Psi_{\rm T}, \Psi_{\rm G}\text{.)}$$

$$E_{\rm L}(i) = \frac{\sum_j^{N_{\rm st}} H_{ij} t_j}{t_i}$$

In both VMC and PMC weighted average of the *configuration value of $\hat{H}$* aka *local energy, $E_{\rm L}(i)$,* but from points sampled from different distributions.

This is practical for systems that are large enough to be interesting if

1. $t_i = \langle \phi_i | \Psi_{\rm T} \rangle$, $g_i = \langle \phi_i | \Psi_{\rm G} \rangle$ can be evaluated in polynomial time, say N^3
2. the sum in $E_{\rm L}(i)$ can be done quickly, i.e., $\hat{H}$ is sparse (if space discrete) or semi-diagonal, i.e. $V(\mathbf{R})$ is local (if space continuous).

† In practice, usually necessary to make approximation (e.g. FN) and value depends on $\Psi_{\rm G}$.

Projector MC

Projector: $|\Psi_0\rangle = \hat{P}(\infty)\,|\Psi_{\mathrm{T}}\rangle = \lim_{n\to\infty} \hat{P}^n(\tau)\,|\Psi_{\mathrm{T}}\rangle$

Projector is any function of the Hamiltonian that maps the ground state eigenvalue of $\hat{H}$ to 1, and the higher eigenvalues of $\hat{H}$ to absolute values that are < 1 (preferably close to 0).

Exponential projector: $\hat{P} = e^{\tau(E_T \hat{\mathbf{1}} - \hat{H})}$ (usually has time-step error)

Linear projector: $\hat{P} = \hat{\mathbf{1}} + \tau(E_T \hat{\mathbf{1}} - \hat{H})$ $\left(\tau < \frac{2}{E_{\mathrm{max}} - E_0}\right)$

Green's function projector: $\hat{P} = \dfrac{1}{\hat{\mathbf{1}} - \tau(E_T \hat{\mathbf{1}} - \hat{H})}$

Importance Sampling in Projector Monte Carlo

We want to sample from $g_i e_i = \langle\phi_i|\Psi_{\rm G}\rangle \langle\phi_i|\Psi_0\rangle$ rather than $e_i = \langle\phi_i|\Psi_0\rangle$.

If

$$\sum_j P_{ij} e_j \quad = \quad e_i$$

the similarity transformed matrix with elements $\tilde{P}_{ij} = \frac{g_i P_{ij}}{g_j}$ has eigenstate with elements $g_i e_i$:

$$\sum_j \tilde{P}_{ij}(g_j e_j) \quad = \quad \sum_j \left(\frac{g_i P_{ij}}{g_j}\right)(g_j e_j) \; = \; g_i e_i$$

$\tilde{P}_{ij}$ is called the *importance sampled* projector.

Variational Monte Carlo in Real Space

Now we make things more concrete by considering the example of variational MC in real space, i.e. the MC walk is in the space of position eigenstates.

Variational Monte Carlo in Real Space

W. L. McMillan, Phys. Rev. **138**, A442 (1965)

Monte Carlo is used to perform the many-dimensional integrals needed to calculate quantum mechanical expectation values. e.g.

$$E_T = \frac{\int d\mathbf{R}\,\Psi_{\mathrm{T}}^*(\mathbf{R})\,\mathcal{H}\,\psi_{\mathrm{T}}(\mathbf{R})}{\int d\mathbf{R}\,\psi_{\mathrm{T}}^2(\mathbf{R})}$$

$$= \int d\mathbf{R}\;\frac{\psi_{\mathrm{T}}^2(\mathbf{R})}{\int d\mathbf{R}\,\psi_{\mathrm{T}}^2(\mathbf{R})}\;\frac{\mathcal{H}\psi_{\mathrm{T}}(\mathbf{R})}{\psi_{\mathrm{T}}(\mathbf{R})}$$

$$= \frac{1}{N}\sum_i \frac{\mathcal{H}\Psi_{\mathrm{T}}(\mathbf{R}_i)}{\Psi_{\mathrm{T}}(\mathbf{R}_i)} = \frac{1}{N}\sum_i E_L(\mathbf{R}_i)$$

Energy is obtained as an arithmetic sum of the *local energies* $E_L(\mathbf{R}_i)$ evaluated for configurations sampled from $\psi_{\mathrm{T}}^2(\mathbf{R})$ using a generalization of the Metropolis method. If ψ_{T} is an eigenfunction the $E_L(\mathbf{R}_i)$ do not fluctuate. Accuracy of VMC depends crucially on the quality of $\psi_{\mathrm{T}}(\mathbf{R})$. Diffusion MC does better by projecting onto ground state.

Three ingredients for accurate Variational Monte Carlo

1. A method for sampling an arbitrary wave function Metropolis-Hastings.
2. A functional form for the wave function that is capable of describing the correct physics/chemistry.
3. An efficient method for optimizing the parameters in the wave functions.

Metropolis-Hastings Monte Carlo

Metropolis, Rosenbluth^2, Teller^2, JCP, **21** 1087 (1953)

W.K. Hastings, Biometrika, **57** (1970)

Metropolis method originally used to sample the Boltzmann distribution. This is still one of its more common uses.

General method for sampling **any known** discrete or continuous density. (Other quantum Monte Carlo methods, e.g., diffusion MC, enable one to sample densities that are not explicitly known but are the eigenstates of known matrices or integral kernels.)

Metropolis-Hastings has serial correlations. Hence, direct sampling methods preferable, but rarely possible for complicated densities in many dimensions.

Metropolis-Hastings Monte Carlo (cont)

A *Markov chain* is specified by two ingredients:
1) an initial state
2) a transition matrix $M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ (probability of transition $\mathbf{R}_\mathrm{i} \rightarrow \mathbf{R}_\mathrm{f}$.)

$$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \geq 0, \quad \sum_{\mathbf{R}_\mathrm{f}} M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = 1. \qquad \textit{Column-stochastic matrix}$$

To sample $\rho(\mathbf{R})$, start from an arbitrary $\mathbf{R}_\mathrm{i}$ and evolve the system by repeated application of M that satisfies the *stationarity condition* (flux into state $\mathbf{R}_\mathrm{i}$ equals flux out of $\mathbf{R}_\mathrm{i}$):

$$\sum_{\mathbf{R}_\mathrm{f}} M(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\, \rho(\mathbf{R}_\mathrm{f}) = \sum_{\mathbf{R}_\mathrm{f}} M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\, \rho(\mathbf{R}_\mathrm{i}) = \rho(\mathbf{R}_\mathrm{i}) \quad \forall\, \mathbf{R}_\mathrm{i}$$

i.e., $\rho(\mathbf{R})$ is a right eigenvector of M with eigenvalue 1.
Stationarity $\Rightarrow$ if we start with ρ, will continue to sample ρ.
Want more than that: *any* initial density should evolve to ρ.

$$\lim_{n\rightarrow\infty} M^n(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\, \delta(\mathbf{R}_\mathrm{i}) = \rho(\mathbf{R}_\mathrm{f}), \quad \forall\, \mathbf{R}_\mathrm{i}.$$

i.e., ρ should be the *dominant* right eigenvector.

Metropolis-Hastings Monte Carlo (cont)

Want that any initial density should evolve to ρ.

$$\lim_{n\to\infty} M^n(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\delta(\mathbf{R}_\mathrm{i}) = \rho(\mathbf{R}_\mathrm{f}), \quad \forall\, \mathbf{R}_\mathrm{i}.$$

ρ should be the *dominant* right eigenvector. Additional conditions needed to guarantee this.

A nonnegative matrix M is said to be *primitive* if $\exists\ n$ such that M^n has all elements positive. (Can go from any state to any other in finite number of steps.)

(Special case of) Perron-Frobenius Theorem: A column-stochastic primitive matrix has a unique dominant eigenvalue of 1, with a positive right eigenvector and a left eigenvector with all components equal to 1 (by definition of column-stochastic matrix).

In practice, length of Monte Carlo should be long enough that there be a significant probability of the system making several transitions between the neighborhoods of any pair of representative states that make a significant contribution to the average. This ensures that states are visited with the correct probability with only small statistical fluctuations.
For example in a double-well system many transitions between the 2 wells should occur, but we can choose our proposal matrix to achieve this even if barrier between wells is high.

Metropolis-Hastings Monte Carlo (cont)

Construction of M

Need a prescription to construct M, such that ρ is its stationary state. Impose *detailed balance* condition

$$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i}) = M(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})$$

Detailed balance more stringent than stationarity condition (removed the sums).
Detailed balance is not necessary but provides way to construct M.
Write elements of M as product of elements of a proposal matrix T and an acceptance Matrix A,

$$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$$

$M(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ and $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ are stochastic matrices, but $A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ is not.
Detailed balance is now:

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i}) = A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})$$

$$\text{or} \qquad \frac{A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})} = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}\ .$$

Metropolis-Hastings Monte Carlo (cont)

Choice of Acceptance Matrix A

$$\frac{A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{A(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})} = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}\ .$$

Infinity of choices for A. Any function

$$F\left(\frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}\right) = A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$$

for which $F(x)/F(1/x) = x$ and $0 \le F(x) \le 1$ will do.
Choice of Metropolis *et al.* $F(x) = \min\{1, x\}$, maximizes the acceptance:

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \min\left\{1, \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}\right\}.$$

Other less good choices for $A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ have been made, *e.g.* $F(x) = \frac{x}{1+x}$

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f}) + T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}.$$

Metropolis: $T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f}) = T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$, Hastings: $T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f}) \neq T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$

Metropolis-Hastings Monte Carlo (cont)

Choice of Proposal Matrix T

So, the optimal choice for the acceptance matrix $A(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})$ is simple and known.

However, there is considerable scope for using one's ingenuity to come up with good proposal matrices, $T(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})$, that allow one to make large moves with large acceptances, in order to make the autocorrelation time small.

Choice of Proposal Matrix T in Metropolis-Hastings (cont)

CJU, PRL **71**, 408 (1993)

$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \min\left\{1, \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}\right\}$$

Use freedom in T to make $\quad \dfrac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})\ \rho(\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})} \approx 1.$

$T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \propto \rho(\mathbf{R}_\mathrm{f})$ optimal if $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ can be sampled over all space – usually not the case. And if it is, then one would not use Metropolis-Hastings in the first place.

Otherwise, let $\quad T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \dfrac{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{\int d\mathbf{R}_\mathrm{f}\ S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})} \approx \dfrac{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{i})\Omega(\mathbf{R}_\mathrm{i})}$

$S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ is non-zero only in domain $D(\mathbf{R}_\mathrm{i})$ of volume $\Omega(\mathbf{R}_\mathrm{i})$ around $\mathbf{R}_\mathrm{i}$).

$$\frac{A(\mathbf{R}_\mathrm{f},\mathbf{R}_\mathrm{i})}{A(\mathbf{R}_\mathrm{i},\mathbf{R}_\mathrm{f})} = \frac{T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})}{T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}\frac{\rho(\mathbf{R}_\mathrm{f})}{\rho(\mathbf{R}_\mathrm{i})} \approx \frac{\Omega(\mathbf{R}_\mathrm{i})}{\Omega(\mathbf{R}_\mathrm{f})}\frac{S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{i})}{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{f})}\frac{S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})}{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}\frac{\rho(\mathbf{R}_\mathrm{f})}{\rho(\mathbf{R}_\mathrm{i})}$$

from which it is apparent that the choice

$S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \sqrt{\rho(\mathbf{R}_\mathrm{f})/\Omega(\mathbf{R}_\mathrm{f})} \quad$ yields $\quad A(\mathbf{R}_\mathrm{f},\mathbf{R}_\mathrm{i})/A(\mathbf{R}_\mathrm{i},\mathbf{R}_\mathrm{f}) \approx 1.$

Choice of Proposal Matrix T in Metropolis-Hastings (cont)

To be more precise, if the log-derivatives of $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ equal those of $\sqrt{\rho(\mathbf{R}_\mathrm{f})/\Omega(\mathbf{R}_\mathrm{f})}$ at $\mathbf{R}_\mathrm{f}=\mathbf{R}_\mathrm{i}$, the acceptance goes as $1-\mathcal{O}((\mathbf{R}^{'}-\mathbf{R})^3)$, i.e., the average acceptance goes as $1-\mathcal{O}(\Delta^4)$, where Δ is the linear dimension of $D(\mathbf{R}_\mathrm{i})$.

Considerable improvement compared to using a symmetric $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ or choosing $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \rho(\mathbf{R}_\mathrm{f})$ for either of which we have $1-\mathcal{O}((\mathbf{R}^{'}-\mathbf{R})^1)$ and $1-\mathcal{O}(\Delta^2)$.

Another possible choice, motivated by (DMC) is

$$T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \frac{1}{(2\pi\tau)^{3/2}} \exp\left[\frac{-(\mathbf{R}_\mathrm{f}-\mathbf{R}_\mathrm{i}-\mathbf{V}(\mathbf{R}_\mathrm{i})\tau)^2}{2\tau}\right], \quad \mathbf{V}(\mathbf{R}_\mathrm{i}) = \frac{\nabla\Psi(\mathbf{R}_\mathrm{i})}{\Psi(\mathbf{R}_\mathrm{i})}$$

Advantage: allows Metropolis Monte Carlo and diffusion Monte Carlo programs to share almost all the code.
Such an algorithm is more efficient than one with a symmetric $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ or one for which $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \rho(\mathbf{R}_\mathrm{f})$, but less efficient than one for which $S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \overset{\propto}{\sim} \sqrt{\rho(\mathbf{R}_\mathrm{f})/\Omega(\mathbf{R}_\mathrm{f})}$.

These arguments are rigorous only in the small-step limit and are applicable only to functions with sufficiently many derivatives within $D(\mathbf{R}_\mathrm{i})$. In practice these ideas yield large reduction in the autocorrelation time provided that we employ a coordinate system such that ρ has continuous derivatives within $D(\mathbf{R}_\mathrm{i})$.

Some examples

We want to sample from $|\Psi(\mathbf{R})|^2$.
We propose moves with probability density

$$T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \frac{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{\int d\mathbf{R}_\mathrm{f}\, S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})} \approx \frac{S(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}{S(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{i})\Omega(\mathbf{R}_\mathrm{i})}$$

and since the acceptance is

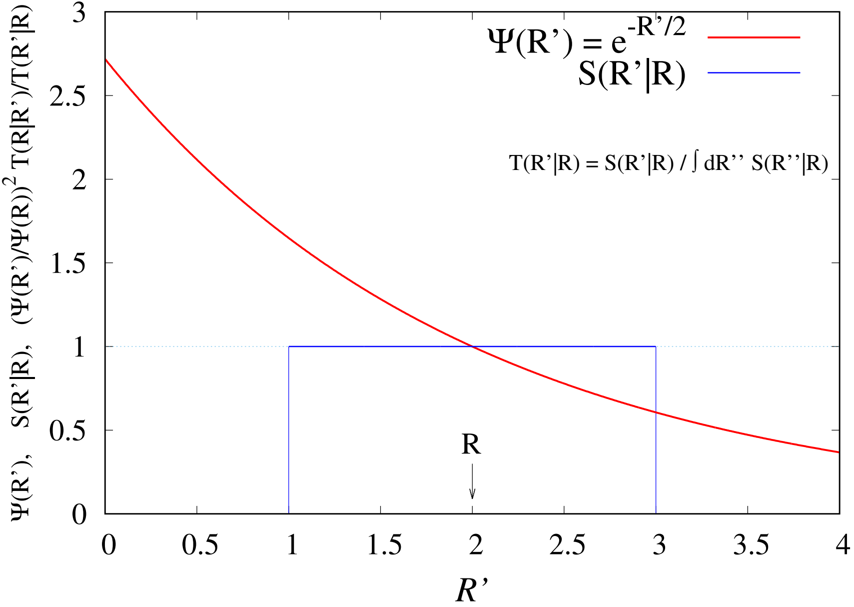
$$A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \min\left\{1, \frac{|\Psi(\mathbf{R}_\mathrm{f})|^2\ T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})}{|\Psi(\mathbf{R}_\mathrm{i})|^2\ T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}\right\}$$

we want

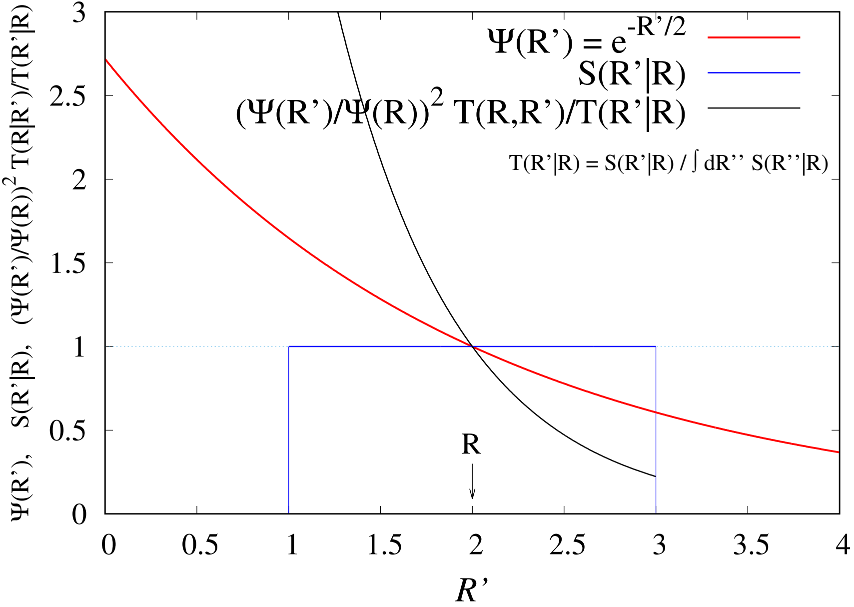
$$\frac{|\Psi(\mathbf{R}_\mathrm{f})|^2\ T(\mathbf{R}_\mathrm{i}|\mathbf{R}_\mathrm{f})}{|\Psi(\mathbf{R}_\mathrm{i})|^2\ T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})}$$

to be as close to 1 as possible. Let's see how it changes with $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$.

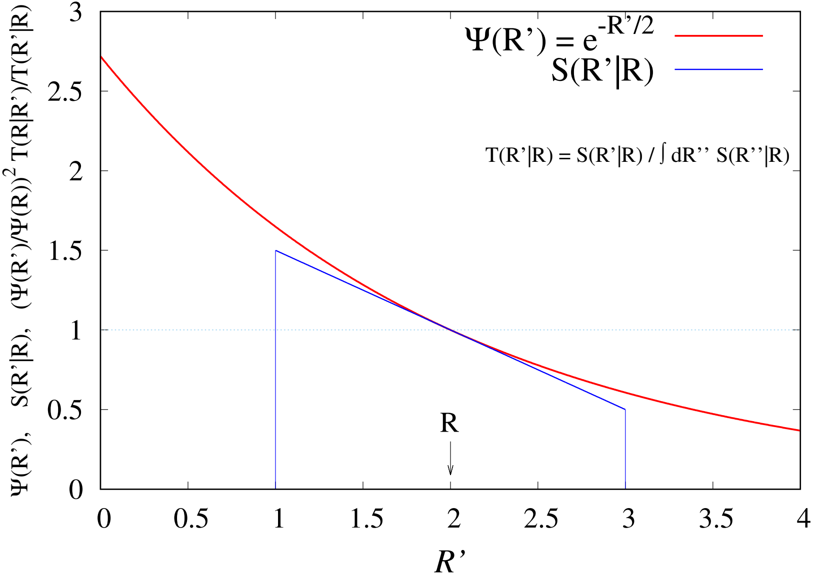
Symmetrical T in Metropolis



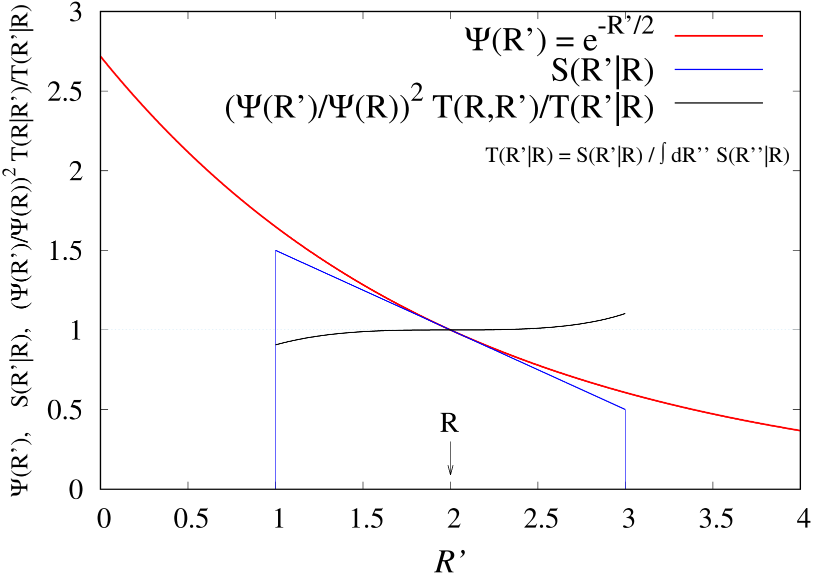
Symmetrical T in Metropolis



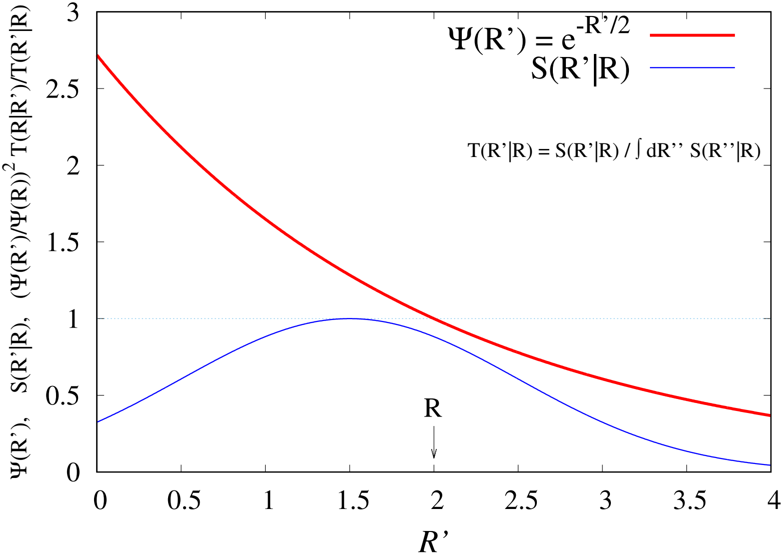
Non-symmetrical linear T in Metropolis-Hastings



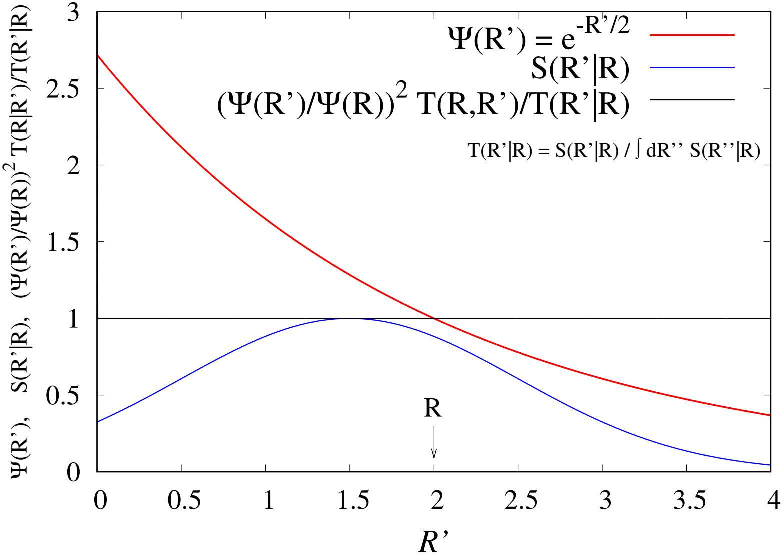
Non-symmetrical linear T in Metropolis-Hastings



Non-symmetrical drifted Gaussian T in Metropolis-Hastings

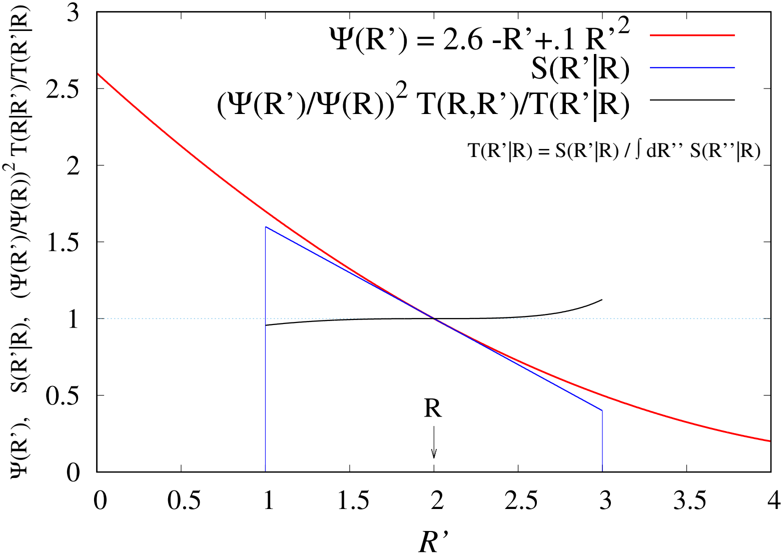


Non-symmetrical drifted Gaussian T in Metropolis-Hastings



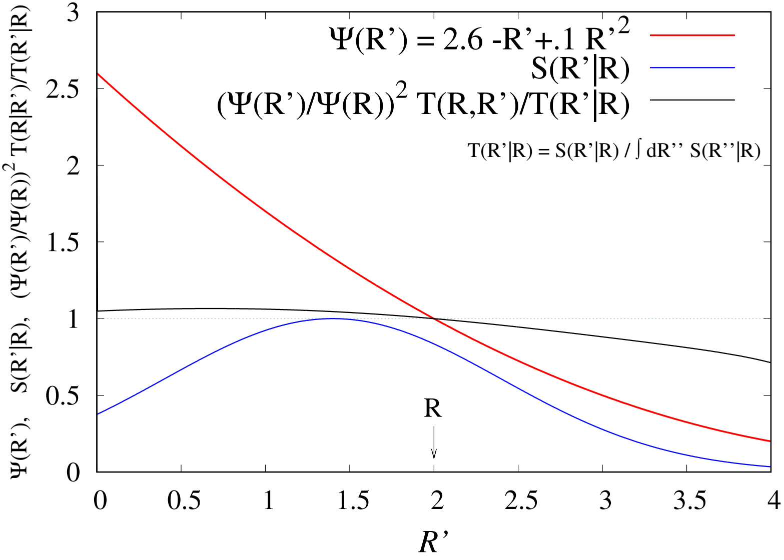
For this $\Psi(R')$, the drifted Gaussian gives perfect acceptance! Not generally true.

Non-symmetrical linear T in Metropolis-Hastings



The force-bias choice works just as well for this different function.

Non-symmetrical drifted Gaussian T in Metropolis-Hastings



For this $\Psi(R')$, the drifted Gaussian deviates from 1 linearly.

Choice of Proposal Matrix T in Metropolis-Hastings (cont)

When will the above not work so well?
What assumptions have we made in both of the non-symmetric choices above?

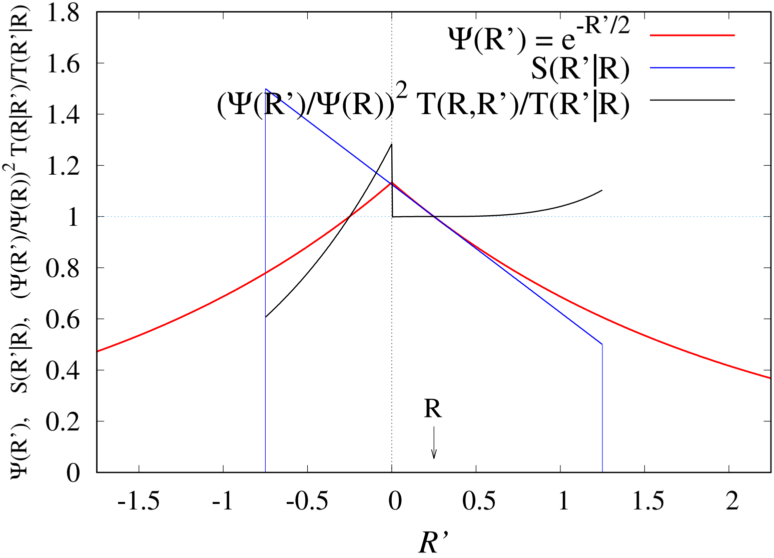
Choice of Proposal Matrix T in Metropolis-Hastings (cont)

When will the above not work so well?
What assumptions have we made in both of the non-symmetric choices above?

Answer: In both cases we are utilizing the gradient of the function to be sampled and are implicitly assuming that it is smooth.

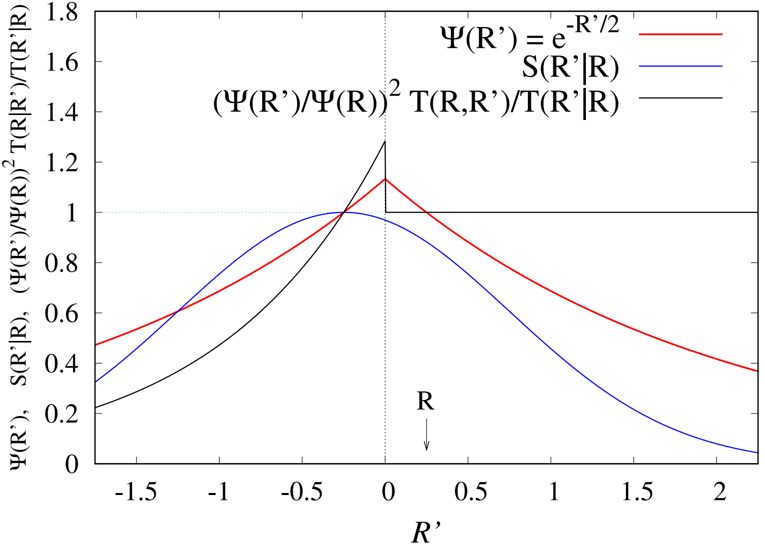
Let's see what happens when it is not.

Non-symmetrical linear T in Metropolis-Hastings



When the gradient has a discontinuity the acceptance goes down.

Non-symmetrical drifted Gaussian T in Metropolis-Hastings



When the gradient has a discontinuity the acceptance goes down.

The drifted-Gaussian even overshoots the nucleus.

Choice of Proposal Matrix T in Metropolis-Hastings (cont)

How to make large moves with high acceptance in spite of wavefunctions that have cusps at nuclei?

1. Make moves in spherical polar coordinates, centered on the nearest nucleus.
2. Radial move is proportional to distance to nucleus, say in interval $[\frac{r}{5}, 5r]$.
3. Angular move gets *larger* as electron approaches nucleus.

Using these ideas an autocorrelation time $T_{\rm corr} \approx 1$ can be achieved!

Details are in: Accelerated Metropolis Method, C. J. Umrigar, PRL **71** 408, (1993).

The point of the above exercise was not the particular problem treated, but rather to provide a concrete example of the ideas that enable making large moves with high acceptance, thereby achieving $T_{\rm corr} \approx 1$.

Estimation of Errors
Autocorrelation time

N Monte Carlo steps $= N_b$ blocks $\times$ N_s steps/block
If N_s is large enough the block averages are nearly independent.

$$\begin{aligned}
\bar{E} &= \text{average of } E_{\rm L} \text{ over the } N \text{ Monte Carlo steps} \\
\sigma &= \text{rms fluctuations of individual } E_{\rm L} \\
\sigma_b &= \text{rms fluctuations of block averages of } E_{\rm L}
\end{aligned}$$

Need to estimate $T_{\rm corr}$ to make sure $N_b \gg T_{\rm corr}$.
$N_{\rm eff} = N/T_{\rm corr}$ independent measurements of $E_{\rm L}$, so get $T_{\rm corr}$ from:

$$\mathrm{err}(\bar{E}) = \frac{\sigma}{\sqrt{N_b \times N_s}} \sqrt{T_{\rm corr}} = \frac{\sigma_b}{\sqrt{N_b}}$$

$$\Rightarrow \boxed{T_{\rm corr} = N_s \left(\frac{\sigma_b}{\sigma}\right)^2}$$

Choose $N_s \gg T_{\rm corr}$, say, 100 $T_{\rm corr}$.
If $N_s \approx 10\,T_{\rm corr}$, $T_{\rm corr}$ underest. $\approx 10\%$.

Blocking Analysis for error of mean of autocorrelated variables

Flyvberg and Peterson, JCP 1979

Compute recursively and plot

$$\frac{1}{N_b(N_b-1)} \sum_{i=1}^{N_b} (m_i - \bar{E})^2$$

for various blocking levels, $N_s = 1,\ 2,\ 2^2,\ 2^3,\ ...,\ N/2$
If the variables were uncorrelated to begin with then these estimates of the error would be equal aside from statistical fluctuations, which would increase with blocking level.
If they are autocorrelated, the estimated error will grow and the flatten out when the block means become uncorrelated, which can only happen if $N \gg T_{\rm corr}$.
Assuming that block means are independent Gaussian variables (they are not at the lower blocking levels), the estimated uncertainty of the error is

$$\frac{\sqrt{2}\ (\text{error estim})}{\sqrt{(N_b-1)}}$$

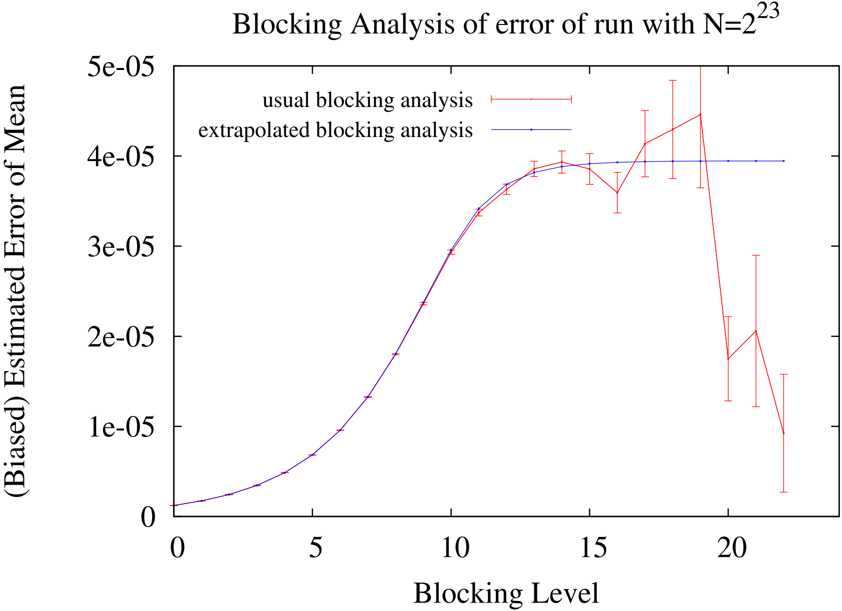
since the PDF of the sum of squares of $N_b - 1$ normal standard deviates is $\chi^2(N_b-1)$ and has variance $2(N_b-1)$. So, cannot go to very large N_s (N_b small).
A reasonable choice of blocking level is the highest one for which the increase in the estimate for the error is larger than the increase in the estimate for the error in the error. It is possible to get a somewhat better estimate by predicting the shape of the curve and extrapolating when say $N < 1000\, T_{\rm corr}$.

Blocking Analysis for error of mean of autocorrelated variables

In variational Monte Carlo, $T_{\rm corr}$ is usually very small if one makes an intelligent choice for the proposal matrix. With the algorithm we typically use $T_{\rm corr} < 2$ even for systems with say 100 electrons!

However, in some of the projector Monte Carlo methods (e.g. FCIQMC), $T_{\rm corr}$ can be much larger, even for much smaller systems. Further, in these methods one needs to use a large population of walkers, so it becomes expensive to have a large number of Monte Carlo steps. In the next viewgraph, a blocking analysis for a run with $T_{\rm corr} \approx 1000$ and $N = 2^{23}$ is shown.

Blocking Analysis for error of mean of autocorrelated variables



Functional form of Trial Wave Function

Other methods: Restrictions on the form of the wavefn.:

1. Many-body wavefn. expanded in determinants of single-particle orbitals.
2. Single-particle orbitals are expanded in planewaves or gaussians.
   occasionally wavelets etc.

QMC: Great freedom in form of the wavefn. – use physics/chemistry intuition:

1. Multideterminant times Jastrow. Ceperley, many others
2. Antisymmetrized Geminal Power times Jastrow. Sorella, Casula
   $\mathcal{A}\left[\Phi(\mathbf{r}_1^{\uparrow},\mathbf{r}_1^{\downarrow})\,\Phi(\mathbf{r}_2^{\uparrow},\mathbf{r}_2^{\downarrow})\,\cdots\,\Phi(\mathbf{r}_{N/2}^{\uparrow},\mathbf{r}_{N/2}^{\downarrow})\right]$
3. Pfaffians times Jastrow. Schmidt, Mitas, Wagner and coworkers
   $\mathcal{A}\left[\Phi(\mathbf{r}_1,s_1;\mathbf{r}_2,s_2)\,\Phi(\mathbf{r}_3,s_3;\mathbf{r}_4,s_4)\;\cdots\;\Phi(\mathbf{r}_{N-1},s_{N-1};\mathbf{r}_N,s_N)\right]$
4. Backflow times Jastrow. Needs and coworkers, Moroni (extension of Feynman)
5. Laughlin and Composite Fermion. Jeon, Güclu, CJU and Jain

Multideterminant $\times$ Jastrow form of Trial Wavefunction

$$\Psi_T = \left(\sum_n d_n \mathrm{D}_n^{\uparrow} \, \mathrm{D}_n^{\downarrow} \right) \times \mathcal{J}(r_i, r_j, r_{ij})$$

• Determinants: $\sum_n d_n \mathrm{D}_n^{\uparrow} \, \mathrm{D}_n^{\downarrow}$

$\mathrm{D}^{\uparrow}$ and $\mathrm{D}^{\downarrow}$ are determinants of single-particle orbitals ϕ for up ($\uparrow$) and down ($\downarrow$) spin electrons respectively.

The single-particle orbitals ϕ are given by:

$$\phi(\mathbf{r}_i) = \sum_{\alpha k_\alpha} c_{k_\alpha} \; N_{k_\alpha} r_{i\alpha}^{n_{k_\alpha}-1} \, e^{-\zeta_{k_\alpha} r_{i\alpha}} \, \mathrm{Y}_{l_{k_\alpha} m_{k_\alpha}}(\widehat{\mathrm{r}}_{i\alpha})$$

• Jastrow: $\mathcal{J}(r_i, r_j, r_{ij}) = \prod_{\alpha i} \exp(A_{\alpha i}) \; \prod_{ij} \exp(B_{ij}) \; \prod_{\alpha ij} \exp(C_{\alpha ij})$

$A_{\alpha i} \Rightarrow$ electron-ion correlation

$B_{ij} \Rightarrow$ electron-electron correlation

$C_{\alpha ij} \Rightarrow$ electron-electron-ion correlation

d_n, c_{k_α}, ζ_{k_α} and parms in $\mathcal{J}$ are optimized.

$\sim N_{\mathrm{atomtype}}$ of $\mathcal{J}$ parms.

$\sim N_{\mathrm{atomtype}}$ of ζ_{k_α} parms.

$\sim N_{\mathrm{atom}}^2$ of c_{k_α} parms.

$\sim e^{N_{\mathrm{atom}}}$ of d_n parms.

Power of QMC:

$\mathcal{J}$ parms. replace many d_n parms.

Cusp-conditions of Trial Wave Functions

Jastrow factor and divergences in the potential

At interparticle coalescence points, the potential diverges as

$$-\frac{Z}{r_{i\alpha}} \quad \text{for the electron-nucleus potential}$$

$$\frac{1}{r_{ij}} \quad \text{for the electron-electron potential}$$

Want local energy $\boxed{\frac{\mathcal{H}\Psi}{\Psi} = -\frac{1}{2}\sum_i \frac{\nabla_i^2\Psi}{\Psi} + \mathcal{V}}$ to be $\boxed{\text{finite}}$ (const. for Ψ_0)

$\Rightarrow$ Kinetic energy must have opposite divergence to the potential $\mathcal{V}$

Cusp-conditions of Trial Wave Functions

Divergence in potential and behavior of the local energy

Consider two particles of masses m_i, m_j and charges q_i, q_j

Assume $r_{ij} \to 0$ while all other particles are well separated

Keep only diverging terms in $\frac{\mathcal{H}\Psi}{\Psi}$ and go to relative coordinates close to $\mathbf{r} = \mathbf{r}_{ij} = 0$

$$-\frac{1}{2\mu_{ij}}\frac{\nabla^2\Psi}{\Psi} + \mathcal{V}(r) \sim -\frac{1}{2\mu_{ij}}\frac{\Psi''}{\Psi} - \frac{1}{\mu_{ij}}\frac{1}{r}\frac{\Psi'}{\Psi} + \mathcal{V}(r)$$

$$\sim \boxed{-\frac{1}{\mu_{ij}}\frac{1}{r}\frac{\Psi'}{\Psi} + \mathcal{V}(r)}$$

where $\mu_{ij} = m_i\, m_j/(m_i + m_j)$

Cusp-conditions of Trial Wave Functions

Divergence in potential and cusp conditions

Diverging terms in the local energy

$$-\frac{1}{\mu_{ij}}\frac{1}{r}\frac{\Psi'}{\Psi}+\mathcal{V}(r)=-\frac{1}{\mu_{ij}}\frac{1}{r}\frac{\Psi'}{\Psi}+\frac{q_i q_j}{r}=\text{ finite}$$

$\Rightarrow \Psi$ must satisfy Kato's cusp conditions:

$$\boxed{\left.\frac{\partial\hat{\Psi}}{\partial r_{ij}}\right|_{r_{ij}=0}=\mu_{ij}q_i\,q_j\Psi(r_{ij}=0)}$$

where $\hat{\Psi}$ is a spherical average

Note: We assumed $\Psi(r_{ij}=0)\neq 0$.
Slightly more involved derivation if $\Psi(r_{ij}=0)=0$ (parallel spins).

Cusp-conditions of Trial Wave Functions

Cusp conditions: example

The condition for the local energy to be finite at $r = 0$ is

$$\frac{\Psi'}{\Psi} = \mu_{ij} q_i \, q_j$$

- Electron-nucleus: $\mu = 1,\ q_i = 1,\ q_j = -Z \quad \Rightarrow \quad \boxed{\left.\frac{\Psi'}{\Psi}\right|_{r=0} = -Z}$
- Electron-electron($\uparrow\downarrow$): $\mu = \frac{1}{2},\ q_i = 1,\ q_j = 1 \quad \Rightarrow \quad \boxed{\left.\frac{\Psi'}{\Psi}\right|_{r=0} = 1/2}$

▷ Electron-nucleus cusps imposed on combination of the determinantal part (using Slater basis functions) and the e-n Jastrow.

Jastrow Factors

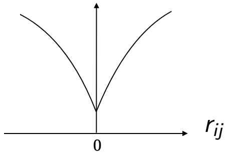
▷ Electron-electron cusps imposed by the Jastrow factor

Example: Simple Jastrow factor

$$\mathcal{J}(r_{ij}) = \prod_{i<j} \exp\left\{ \frac{b_1\, r_{ij}}{1 + b_2\, r_{ij}} \right\}$$

with $\quad b_1^{\uparrow\downarrow} = \dfrac{1}{2} \quad$ or $\quad b_1^{\uparrow\uparrow} = b_1^{\downarrow\downarrow} = \dfrac{1}{4}$

$$\boxed{\begin{array}{c}\text{Imposes cusp conditions} \\ + \\ \text{keeps electrons apart}\end{array}}$$



Comments on Jastrow factor

▷ Electron-electron-nucleus terms C

The Jastrow does not change the nodes of the wavefunction.

If the order of the polynomial in the e-e-n terms is infinite, Ψ can exactly describe a two-electron atom or ion in ${}^1S$ ground state

For the He atom, a $6^{\rm th}$-order polynomial gives energies good to better than 1 μHa, or 99.998% of the correlation energy, $E_{\rm corr} = E_{\rm exact} - E_{\rm HF}$

▷ Is this Jastrow factor adequate for multi-electron systems?

The e-e-n terms are the most important: due to the exclusion principle, it is rare for 3 or more electrons to be close, since at least 2 electrons must necessarily have the same spin

Higher-order Jastrow factors

Jastrow factor with e-e, e-e-n and e-e-e-n terms

		$\mathcal{J}$	E_{VMC}	$E_{\mathrm{VMC}}^{\mathrm{corr}}$ (%)	σ_{VMC}
Li	E_{HF}		-7.43273	0	
		e-e	-7.47427(4)	91.6	0.240
		+ e-e-n	-7.47788(1)	99.6	0.037
		+ e-e-e-n	-7.47797(1)	99.8	0.028
	E_{exact}		-7.47806	100	0
Ne	E_{HF}		-128.5471	0	
		e-e	-128.713(2)	42.5	1.90
		+ e-e-n	-128.9008(1)	90.6	0.90
		+ e-e-e-n	-128.9029(3)	91.1	0.88
	E_{exact}		-128.9376	100	0

Huang, Umrigar, Nightingale, J. Chem. Phys. **107**, 3007 (1997)

Static and Dynamic Correlation

Dynamic and static correlation

Ψ = Jastrow $\times$ Determinants $\rightarrow$ Two types of correlation

- Dynamic correlation

  Due to inter-electron repulsion

  Always present

  Efficiently described by Jastrow factor

- Static correlation

  Due to near-degeneracy of occupied and unoccupied orbitals

  Not always present

  Efficiently described by a linear combination of determinants (change nodes of Ψ)

Static and Dynamic Correlation

Example: Be atom has $2s$-$2p$ near-degeneracy, prototypical example of static correlation

HF ground state configuration $\boxed{1s^2 2s^2}$

Additional important configuration $\boxed{1s^2 2p^2}$

Ground state has ^1S symmetry $\Rightarrow$ 4 determinants:

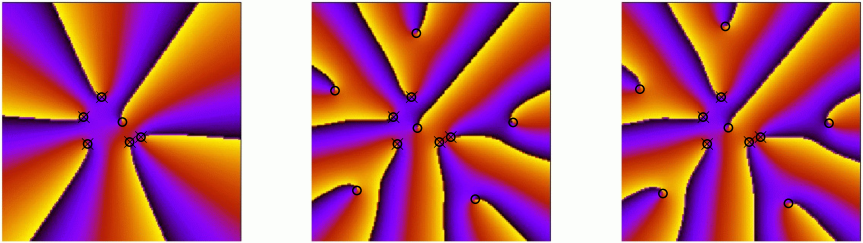
$$D = (1s^\uparrow, 2s^\uparrow, 1s^\downarrow, 2s^\downarrow) \; +$$
$$c \left[(1s^\uparrow, 2p_x^\uparrow, 1s^\downarrow, 2p_x^\downarrow) + (1s^\uparrow, 2p_y^\uparrow, 1s^\downarrow, 2p_y^\downarrow) + (1s^\uparrow, 2p_z^\uparrow, 1s^\downarrow, 2p_z^\downarrow) \right]$$

$$\boxed{\begin{array}{lll} 1s^2 2s^2 & \times\; \mathcal{J}(r_{ij}) & \rightarrow E_{\rm VMC}^{\rm corr} = 61\% \\ 1s^2 2s^2 & \times\; \mathcal{J}(r_{ij}, r_{\alpha i}, r_{\alpha j}) & \rightarrow E_{\rm VMC}^{\rm corr} = 80\% \\ 1s^2 2s^2 \oplus 1s^2 2p^2 & \times\; \mathcal{J}(r_{ij}, r_{\alpha i}, r_{\alpha j}) & \rightarrow E_{\rm VMC}^{\rm corr} = 99.3\% \end{array}}$$

Wavefunctions for Hole in a Filled Landau Level

6 electrons in a harmonic well and a magnetic field.

5 electrons are fixed. Phase of the wavefunction is plotted as the 6^{th} moves.



Vortex hole	Composite-fermion	Exact diagonalization
E = 4.293	E = 4.265	E = 4.264

Jeon, Güclu, CJU, Jain, PRB 2005

Optimization of many-body wavefunctions

Almost all errors reduced by optimizing trial wavefunctions

1. Statistical error (both the rms fluctuations of E_{L} and the autocorrelation time)
2. E_{VMC}
3. Fixed-node error in E_{DMC} (nodes move during optimization). Fixed node errors can be LARGE. For C_2, FN error for 1-det wavefn is 1.3 eV for total energy and 0.7 eV for well-depth. However, optimized multidet. wavefn has FN error that is better than chemical accuracy (1 kcal/mole $= 0.043$ eV/molecule).
4. Time-step error in DMC
5. Population control error in PMC
6. Pseudopotential locality error in DMC when using nonlocal pseudopotentials
7. Error of observables that do not commute with the Hamiltonian (mixed estimators, $\langle \Psi_0 | \hat{A} | \Psi_T \rangle$ not exact even for nodeless ψ_0, ψ_{T}) if one does not use forward/side walking.

Choices to be made when optimizing trial wavefunctions

1. What precisely do we want to optimize – the objective function or measure of goodness?
2. What method do we use to do the optimization? If more than one method is applied to the same objective function, they will of course give the same wavefunction, but the efficiency with which we arrive at the solution may be much different.

Measures of goodness of variational wave functions

$$
\begin{aligned}
\min\ E_{\rm VMC} &= \frac{\langle\psi_{\rm T}|H|\psi_{\rm T}\rangle}{\langle\psi_{\rm T}|\psi_{\rm T}\rangle} = \langle E_{\rm L}\rangle_{|\psi_{\rm T}|^2} \\
\min\ \sigma^2_{\rm VMC} &= \frac{\langle\psi_{\rm T}|(H-E_{\rm T})^2|\psi_{\rm T}\rangle}{\langle\psi_{\rm T}|\psi_{\rm T}\rangle} = \left\langle E^2_{\rm L}(\mathbf{R}_i)\right\rangle_{|\psi_{\rm T}|^2} - \langle E_{\rm L}(\mathbf{R}_i)\rangle^2_{|\psi_{\rm T}|^2} \\
\max\ \Omega^2 &= \frac{|\langle\psi_{\rm FN}|\psi_{\rm T}\rangle|^2}{\langle\psi_{\rm FN}|\psi_{\rm FN}\rangle\,\langle\psi_{\rm T}|\psi_{\rm T}\rangle} = \frac{\left\langle \frac{\psi_{\rm FN}}{\psi_{\rm T}}\right\rangle^2_{|\psi_{\rm T}|^2}}{\left\langle \left|\frac{\psi_{\rm FN}}{\psi_{\rm T}}\right|^2\right\rangle_{|\psi_{\rm T}|^2}} \\
\min\ E_{\rm DMC} &= \frac{\langle\psi_{\rm FN}|H|\psi_{\rm T}\rangle}{\langle\psi_{\rm FN}|\psi_{\rm T}\rangle} = \langle E_{\rm L}\rangle_{|\psi_{\rm FN}\psi_{\rm T}|}
\end{aligned}
$$

For an infinitely flexible wave function all optimizations will yield the exact wavefunction (except that minimizing σ could yield an excited state) but for the imperfect functional forms used in practice they differ.

Progress in optimization of Many-Body Wavefunctions

Naive energy optim. $\rightarrow$ Variance optim. $\rightarrow$ Efficient energy optim.

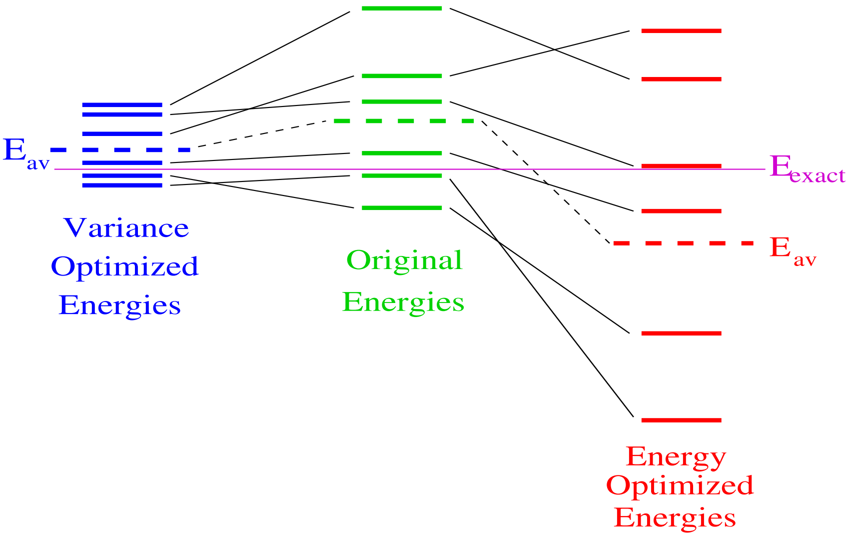
	–	1988	naive energy optimization, few (~ 3) parameters
1988	–	2001	variance optimization, ~ 100 parameters could be used for more, but, variance does not couple strongly to some parameters
2001	–	2012	efficient energy optimization, ~ 1000's of parameters
2012	–	2017	efficient energy optimization, $\sim 100,000$'s of parameters

Most recent advance enables calculation of energy derivatives wrt wavefn. and geometric derivatives for < 4 times cost of just the energy:
Filippi, Assaraf, Moroni, JCP 2016. Assaraf, Moroni, Filippi, JCTC 2017.

Variance vs. Energy

$$\sigma^2 = \sum_{i=1}^{N_{\rm conf}} \left(\frac{\mathcal{H}\Psi_{\rm T}(\mathbf{R}_i)}{\Psi_{\rm T}(\mathbf{R}_i)} - \bar{E} \right)^2$$

$$\bar{E} = \sum_{i=1}^{N_{\rm conf}} \frac{\mathcal{H}\Psi_{\rm T}(\mathbf{R}_i)}{\Psi_{\rm T}(\mathbf{R}_i)}$$



Take-home Message

Energy optimization methods that minimize the energy evaluated on finite sample will yield poor energies on other samples, unless the sample used to do the minimization is very large.

So, efficient energy optimization methods do NOT optimize the energy evaluated on a finite sample, although they do minimize the energy in the limit of an infinite sample.

Advantages of Energy (or Mixed) Optim. vs. Variance Optim.

1. Want lowest energy; fluctuations are of secondary importance. Energy and variance do not always go hand-in-hand enough.
2. Some parameters couple more strongly to energy than variance.
3. Some variance-optimized parameters make wave function too extended.
4. Hellman-Feynman theorem can be used for forces (when combined with variance reduction methods).

Variational energy optimization methods

1. **Newton method** CJU, Filippi, PRL 94, 150201 (2005):
   Add terms to the Hessian that contribute nothing in the limit of an infinite MC sample, but cancel much of the fluctuations for a finite MC sample.
   Gain in efficiency: 3 orders of magnitude for NO_2, more for $C_{10}H_{12}$ compared to Newton of Lin-Zhang-Rappe.

2. **Linear method (generalized eigenvalue problem):**

   1 Linear parameters: Nightingale, et al., PRL, **87**, 043401 (2001)
   Use asymmetric H to have zero variance property in the limit that the basis functions span an invariant subspace.

   2 Nonlinear parameters: Toulouse, CJU, J. Chem. Phys., (2007, 2008).
   CJU, Toulouse, Filippi, Sorella, Hennig, PRL **98**, 110201 (2007).
   Choose freedom of normalization $\Psi(\mathbf{p}, \mathbf{R}) = N(\mathbf{p})\ \Phi(\mathbf{p}, \mathbf{R})$ to make a near optimal change in the parameters.

3. **Perturbation theory in an arbitrary nonorthog. basis:**
   Toulouse, CJU, J. Chem. Phys., **126**, 084102 (2007).
   (Small modification of Scemama-Filippi (2006) perturbative EFP, modification of the Fahy-Filippi-Prendergast-Schautz EFP method.)

4. **Stochastic Reconfiguration:**
   Sorella, Casula, Rocca, J. Chem. Phys., **127**, 014105 (2007).
   Although it requires more iterations than 1) and 2), it is well suited for very large numbers of parameters.

Newton Method

Calculate gradient **g** and Hessian **h** of objective function and update parameters:

$$\mathbf{p}_{\rm next} - \mathbf{p}_{\rm current} \quad \equiv \quad \delta\mathbf{p} \ = \ -\mathbf{h}^{-1}\mathbf{g}$$

or more efficiently ($\mathcal{O}(N_p^2)$ vs. $\mathcal{O}(N_p^3)$) find parameter changes, $\delta\mathbf{p}$, by solving linear equations:

$$\mathbf{h}\ \delta\mathbf{p} = -\mathbf{g},$$

Optimization of Jastrow and determinantal parameters encounter different problems.

Jastrow: For the form of the Jastrow we use and the systems we study the eigenvalues of the Hessian span 10-12 orders of magnitude. So using steepest descent is horribly slow and using the Hessian, or a reasonable approximation to it, is essential even if there were no statistical noise.

determinantal: The eigenvalues of the Hessian span only 1-2 orders of magnitude. However, the Hessian has terms involving

$$\frac{\frac{\partial\psi}{\partial p_i}}{\psi}$$

that diverge as $\psi \to 0$. The strongest divergence among various terms cancels.

Energy Minimization via Newton

Lin, Zhang, Rappe, JCP 2000;   CJU, Filippi, PRL 2005

$$\bar{E} = \frac{\langle\psi|H|\psi\rangle}{\langle\psi|\psi\rangle} = \langle E_{\rm L}\rangle_{\psi^2}; \qquad E_{\rm L}(\mathbf{R}) = \frac{H\psi(\mathbf{R})}{\psi(\mathbf{R})}$$

Energy gradient components, $\bar{E}_i$:

$$\begin{aligned}
\bar{E}_i &= \frac{\langle\psi_i|H\psi\rangle + \langle\psi|H\psi_i\rangle}{\langle\psi|\psi\rangle} - 2\frac{\langle\psi|H|\psi\rangle\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle^2} \\
&= \frac{\langle\psi_i|H\psi\rangle + \langle\psi|H\psi_i\rangle}{\langle\psi|\psi\rangle} - 2\frac{\bar{E}\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle} = 2\frac{\langle\psi_i|H\psi\rangle - \bar{E}\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle} \quad \text{(by Hermiticity)} \\
&= \left\langle \frac{\psi_i}{\psi}E_{\rm L} + \frac{H\psi_i}{\psi} - 2\bar{E}\frac{\psi_i}{\psi}\right\rangle_{\psi^2} = 2\left\langle\frac{\psi_i}{\psi}(E_{\rm L} - \bar{E})\right\rangle_{\psi^2} \quad \text{(MC expression)}
\end{aligned}$$

Is blue or green expression better for MC?

Energy Minimization via Newton

Lin, Zhang, Rappe, JCP 2000;   CJU, Filippi, PRL 2005

$$\bar{E} = \frac{\langle\psi|H|\psi\rangle}{\langle\psi|\psi\rangle} = \langle E_{\rm L}\rangle_{\psi^2}; \qquad E_{\rm L}(\mathbf{R}) = \frac{H\psi(\mathbf{R})}{\psi(\mathbf{R})}$$

Energy gradient components, $\bar{E}_i$:

$$\begin{aligned}
\bar{E}_i &= \frac{\langle\psi_i|H\psi\rangle + \langle\psi|H\psi_i\rangle}{\langle\psi|\psi\rangle} - 2\frac{\langle\psi|H|\psi\rangle\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle^2} \\
&= \frac{\langle\psi_i|H\psi\rangle + \langle\psi|H\psi_i\rangle}{\langle\psi|\psi\rangle} - 2\frac{\bar{E}\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle} = 2\frac{\langle\psi_i|H\psi\rangle - \bar{E}\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle} \quad \text{(by Hermiticity)} \\
&= \left\langle \frac{\psi_i}{\psi}E_{\rm L} + \frac{H\psi_i}{\psi} - 2\bar{E}\frac{\psi_i}{\psi} \right\rangle_{\psi^2} = 2\left\langle \frac{\psi_i}{\psi}(E_{\rm L} - \bar{E}) \right\rangle_{\psi^2} \quad \text{(MC expression)}
\end{aligned}$$

Is blue or green expression better for MC?

Green is better because it is a zero-variance expression in the limit that ψ is the exact ground state (CJU, Filippi, PRL 2005) Moreover it is simpler and faster.

Energy Minimization via Newton

CJU, Filippi, PRL 2005

Energy hessian components, E_{ij}:

$$\bar{E}_i = 2\frac{\langle\psi_i|H\psi\rangle - \bar{E}\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle} \equiv 2\frac{\left\langle\psi_i\psi(E_\mathrm{L}-\bar{E})\right\rangle}{\langle\psi^2\rangle}$$

$$\bar{E}_{ij} = 2\left[\frac{\left\langle(\psi_{ij}\psi+\psi_i\psi_j)(E_\mathrm{L}-\bar{E})\right\rangle + \left\langle\psi_i\psi(E_{\mathrm{L},j}-\bar{E}_j)\right\rangle - \bar{E}_i\left\langle\psi\psi_j\right\rangle}{\langle\psi^2\rangle}\right]$$

$$= 2\left[\left\langle\left(\frac{\psi_{ij}}{\psi}+\frac{\psi_i\psi_j}{\psi^2}\right)(E_\mathrm{L}-\bar{E})\right\rangle_{\psi^2} - \left\langle\frac{\psi_i}{\psi}\right\rangle_{\psi^2}\bar{E}_j - \left\langle\frac{\psi_j}{\psi}\right\rangle_{\psi^2}\bar{E}_i + \left\langle\frac{\psi_i}{\psi}E_{\mathrm{L},j}\right\rangle_{\psi^2}\right].$$

What can be done to improve this expression?

Energy Minimization via Newton

CJU, Filippi, PRL 2005

Energy hessian components, E_{ij}:

$$
\begin{aligned}
\bar{E}_i &= 2\frac{\langle\psi_i|H\psi\rangle - \bar{E}\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle} \equiv 2\frac{\left\langle\psi_i\psi(E_\mathrm{L}-\bar{E})\right\rangle}{\langle\psi^2\rangle} \\
\bar{E}_{ij} &= 2\left[\frac{\left\langle(\psi_{ij}\psi+\psi_i\psi_j)(E_\mathrm{L}-\bar{E})\right\rangle + \left\langle\psi_i\psi(E_{\mathrm{L},j}-\bar{E}_j)\right\rangle - \bar{E}_i\left\langle\psi\psi_j\right\rangle}{\langle\psi^2\rangle}\right] \\
&= 2\left[\left\langle\left(\frac{\psi_{ij}}{\psi}+\frac{\psi_i\psi_j}{\psi^2}\right)(E_\mathrm{L}-\bar{E})\right\rangle_{\psi^2} - \left\langle\frac{\psi_i}{\psi}\right\rangle_{\psi^2}\bar{E}_j - \left\langle\frac{\psi_j}{\psi}\right\rangle_{\psi^2}\bar{E}_i + \left\langle\frac{\psi_i}{\psi}E_{\mathrm{L},j}\right\rangle_{\psi^2}\right].
\end{aligned}
$$

What can be done to improve this expression?

1) Symmetrize – but this does not reduce fluctuations much

Energy Minimization via Newton

CJU, Filippi, PRL 2005

Energy hessian components, E_{ij}:

$$\bar{E}_i = 2\frac{\langle\psi_i|H\psi\rangle - \bar{E}\,\langle\psi|\psi_i\rangle}{\langle\psi|\psi\rangle} \equiv 2\frac{\left\langle\psi_i\psi(E_\mathrm{L}-\bar{E})\right\rangle}{\langle\psi^2\rangle}$$

$$\bar{E}_{ij} = 2\left[\frac{\left\langle(\psi_{ij}\psi+\psi_i\psi_j)(E_\mathrm{L}-\bar{E})\right\rangle + \left\langle\psi_i\psi(E_{\mathrm{L},j}-\bar{E}_j)\right\rangle - \bar{E}_i\,\langle\psi\psi_j\rangle}{\langle\psi^2\rangle}\right]$$

$$= 2\left[\left\langle\left(\frac{\psi_{ij}}{\psi}+\frac{\psi_i\psi_j}{\psi^2}\right)(E_\mathrm{L}-\bar{E})\right\rangle_{\psi^2} - \left\langle\frac{\psi_i}{\psi}\right\rangle_{\psi^2}\bar{E}_j - \left\langle\frac{\psi_j}{\psi}\right\rangle_{\psi^2}\bar{E}_i + \left\langle\frac{\psi_i}{\psi}E_{\mathrm{L},j}\right\rangle_{\psi^2}\right].$$

What can be done to improve this expression?

1) Symmetrize – but this does not reduce fluctuations much

2) Noting that $\langle E_{\mathrm{L},j}\rangle_{\psi^2} = \frac{\left\langle\psi^2\left(\frac{H\psi}{\psi}\right)_j\right\rangle}{\langle\psi^2\rangle} = \frac{\left\langle\psi^2\left(\frac{H\psi_j}{\psi}-\frac{\psi_j}{\psi^2}H\Psi\right)\right\rangle}{\langle\psi^2\rangle} = \frac{\langle\psi H\psi_i-\psi_i H\psi\rangle}{\langle\psi^2\rangle} = 0$

by hermiticity of $\hat{H}$, and, that the fluctuations of the covariance $\langle ab\rangle-\langle a\rangle\langle b\rangle$ are smaller than those of the product $\langle ab\rangle$, when $\sqrt{\langle a^2\rangle-\langle a\rangle^2} \ll |\langle a\rangle|$ and $\langle b\rangle=0$ on ∞ sample but $\langle b\rangle\neq 0$ on finite sample, replace

$$\left\langle\frac{\psi_i}{\psi}E_{\mathrm{L},j}\right\rangle_{\psi^2} \rightarrow \frac{1}{2}\left(\left\langle\frac{\psi_i}{\psi}E_{\mathrm{L},j}\right\rangle_{\psi^2} - \left\langle\frac{\psi_i}{\psi}\right\rangle_{\psi^2}\langle E_{\mathrm{L},j}\rangle_{\psi^2} + \left\langle\frac{\psi_j}{\psi}E_{\mathrm{L},i}\right\rangle_{\psi^2} - \left\langle\frac{\psi_j}{\psi}\right\rangle_{\psi^2}\langle E_{\mathrm{L},i}\rangle_{\psi^2}\right)$$

3) Too hard to describe here.

Energy Minimization via Newton

$$
\begin{aligned}
\bar{E}_{ij} = & \, 2\left[\left\langle\left(\frac{\psi_{ij}}{\psi}+\frac{\psi_i\psi_j}{\psi^2}\right)(E_{\rm L}-\bar{E})\right\rangle_{\psi^2} - \left\langle\frac{\psi_i}{\psi}\right\rangle_{\psi^2}\bar{E}_j - \left\langle\frac{\psi_j}{\psi}\right\rangle_{\psi^2}\bar{E}_i\right] \\
& + \left\langle\frac{\psi_i}{\psi}E_{{\rm L},j}\right\rangle_{\psi^2} - \left\langle\frac{\psi_i}{\psi}\right\rangle_{\psi^2}\left\langle E_{{\rm L},j}\right\rangle_{\psi^2} + \left\langle\frac{\psi_j}{\psi}E_{{\rm L},i}\right\rangle_{\psi^2} - \left\langle\frac{\psi_j}{\psi}\right\rangle_{\psi^2}\left\langle E_{{\rm L},i}\right\rangle_{\psi^2} \\
= & \, 2\left[\left\langle\left(\frac{\psi_{ij}}{\psi}-\frac{\psi_i\psi_j}{\psi^2}\right)(E_{\rm L}-\bar{E})\right\rangle_{\psi^2} \quad \text{(0 for } p_i \text{ linear in exponent)}\right. \\
& \left. +2\left\langle\left(\frac{\psi_i}{\psi}-\left\langle\frac{\psi_i}{\psi}\right\rangle_{\psi^2}\right)\left(\frac{\psi_j}{\psi}-\left\langle\frac{\psi_j}{\psi}\right\rangle_{\psi^2}\right)(E_{\rm L}-\bar{E})\right\rangle_{\psi^2}\right] \\
& + \left\langle\frac{\psi_i}{\psi}E_{{\rm L},j}\right\rangle_{\psi^2} - \left\langle\frac{\psi_i}{\psi}\right\rangle_{\psi^2}\left\langle E_{{\rm L},j}\right\rangle_{\psi^2} + \left\langle\frac{\psi_j}{\psi}E_{{\rm L},i}\right\rangle_{\psi^2} - \left\langle\frac{\psi_j}{\psi}\right\rangle_{\psi^2}\left\langle E_{{\rm L},i}\right\rangle_{\psi^2}.
\end{aligned}
$$

1) Blue and green terms are zero variance estimators.
2) Red terms are not, but, the terms we added in =0 for infinite sample and cancel most of the fluctuations for a finite sample.

Linear method for linear parameters

If all parameters are linear, i.e., $\psi = \sum_i p_i \psi_i$, then optimize using generalized eigenvalue equation, $\mathbf{Hp} = E\mathbf{Sp}$.

Symmetric or nonsymmetric H?

1) true **H** is symmetric:

$$\text{True} \quad H_{ij} = \int \mathrm{d}^{3N}R \; \psi_i(\mathbf{R}) \; \hat{H} \; \psi_j(\mathbf{R}) \quad \text{symmetric}$$

$$\text{MC estim.} \quad H_{ij} = \sum_{n=1}^{N_{\mathrm{MC}}} \frac{\psi_i(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \left(\frac{\hat{H}\psi_j(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \right) \quad \text{nonsymmetric}$$

$$\text{MC estim.} \quad H_{ij} = \frac{1}{2} \sum_{n=1}^{N_{\mathrm{MC}}} \left(\frac{\psi_i(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \; \frac{\hat{H}\psi_j(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} + \frac{\hat{H}\psi_i(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \; \frac{\psi_j(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \right) \quad \text{symmetric}$$

Linear method for linear parameters

If all parameters are linear, i.e., $\psi = \sum_i p_i \psi_i$, then optimize using generalized eigenvalue equation, $\mathbf{Hp} = E\mathbf{Sp}$.

Symmetric or nonsymmetric H?

1) true **H** is symmetric:

$$\begin{aligned}
\text{True} \quad H_{ij} &= \int \mathrm{d}^{3N}R \; \psi_i(\mathbf{R}) \; \hat{H} \; \psi_j(\mathbf{R}) \quad \text{symmetric} \\
\text{MC estim.} \quad H_{ij} &= \sum_{n=1}^{N_{\rm MC}} \frac{\psi_i(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \left(\frac{\hat{H}\psi_j(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \right) \quad \text{nonsymmetric} \\
\text{MC estim.} \quad H_{ij} &= \frac{1}{2} \sum_{n=1}^{N_{\rm MC}} \left(\frac{\psi_i(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \, \frac{\hat{H}\psi_j(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} + \frac{\hat{H}\psi_i(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \, \frac{\psi_j(\mathbf{R}_n)}{\psi(\mathbf{R}_n)} \right) \quad \text{symmetric}
\end{aligned}$$

2) Minimizing the energy evaluated on a finite sample, i.e., minimizing the Rayleigh quotient, $\partial E / \partial p_k = 0$, even with nonsymmetric H evaluated on finite sample, gives generalized eigenvalue equation with symmetric **H**:

$$E = \min_{\mathbf{p}} \frac{\mathbf{p}^{\mathsf{T}}\mathbf{Hp}}{\mathbf{p}^{\mathsf{T}}\mathbf{Sp}} = \min_{\mathbf{p}} \frac{\sum_{ij} p_i H_{ij} p_j}{\sum_{ij} p_i S_{ij} p_j}$$

$$\frac{\partial E}{\partial p_k} = 0 \implies \left(\sum_{ij} p_i S_{ij} p_j \right) \left(\sum_j H_{kj} p_j + \sum_i p_i H_{ik} \right) - \left(\sum_{ij} p_i H_{ij} p_j \right) \left(2 \sum_j S_{kj} p_j \right) = 0$$

$$\frac{(\mathbf{H} + \mathbf{H}^{\mathsf{T}})}{2} \mathbf{p} = E\mathbf{Sp}$$

Nonsymm. H satisfies strong zero-variance principle

M. P. Nightingale and Melik-Alaverdian, PRL, **87**, 043401 (2001).

Nightingale's strong zero-variance principle:
If the states $\psi_i(\mathbf{R})$ are closed under $\hat{H}$ then the values of the optimized parameters using nonsymmetric H_{ij} are independent of the MC sample, provided $N_{\rm MC} \geq N_p$.

Proof: If closed $\exists \; \{p_j\}$ s.t.
$$\hat{H} \sum_{j=1}^{N_p} p_j |\psi_j\rangle = E \sum_{j=1}^{N_p} p_j |\psi_j\rangle$$

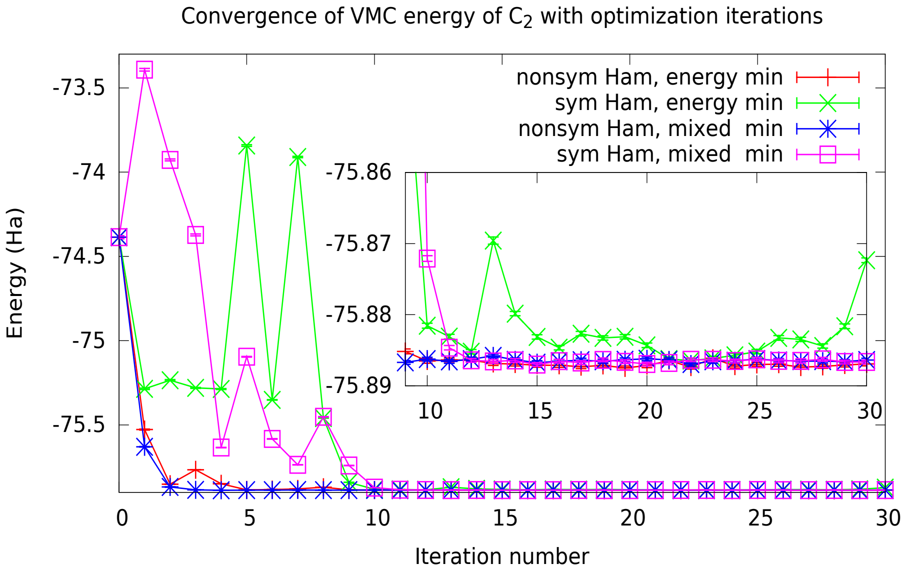
$\times \;\; \langle\psi_i|\mathbf{R}_n\rangle \langle\mathbf{R}_n| / \langle\psi|\mathbf{R}_n\rangle^2$ and sum over $N_{\rm MC}$ pts. (not complete sum over $\mathbf{R}$ states), sampled from $|\psi(\mathbf{R})|^2$:

$$\sum_{j=1}^{N_p} p_j \underbrace{\sum_{n=1}^{N_{\rm MC}} \frac{\langle\psi_i|\mathbf{R}_n\rangle}{\langle\psi|\mathbf{R}_n\rangle} \frac{\langle\mathbf{R}_n|\hat{H}|\psi_j\rangle}{\langle\mathbf{R}_n|\psi\rangle}}_{H_{ij}} = E \sum_{j=1}^{N_p} p_j \underbrace{\frac{\sum_{n=1}^{N_{\rm MC}} \langle\psi_i|\mathbf{R}_n\rangle \langle\mathbf{R}_n|\psi_j\rangle}{\langle\psi|\mathbf{R}_n\rangle \langle\mathbf{R}_n|\psi\rangle}}_{S_{ij}}$$

$$\sum_{i=1}^{N_p} H_{ij} p_j = E \sum_{i=1}^{N_p} S_{ij} p_j$$

H is nonsymmetric H of previous slide. Becomes symmetric when $\sum \to \int$.

Convergence of energy with symmetric and nonsymmetric Hamiltonians



Linear method for nonlinear parameters

Toulouse, CJU, JCP (2007,2008); CJU et al., PRL, **87**, 043401 (2007).

Make linear-order Taylor expansion of Ψ (use $\Psi_i = \partial\Psi/\partial p_i$ as basis):

$$\Psi_{\rm lin} = \Psi_0 + \sum_{i=0}^{N_{\rm parm}} \Delta p_i \, \Psi_i, \qquad \text{(Normalization: } \Delta p_0 = 1\text{)}$$

$\Psi_0 \equiv \Psi(\mathbf{p}_0, \mathbf{R})$ = current wave function
$\Psi_{\rm lin}$ = next linearized wave function
Ψ_i = derivative of Ψ at $\mathbf{p}_0$, wrt i^{th} parameter.
No unique way to obtain new nonlinear parameters.
The simplest procedure: is $p_i^{\rm new} = p_i + \Delta p_i$. Will not work in general. What can one do?
More complicated procedure: fit wave function form to optimal linear combination.
Simpler, yet efficient approach, freedom of norm to make linear approximation better

$$\begin{aligned} \bar\Psi(\mathbf{p}, \mathbf{R}) &= N(\mathbf{p}) \, \Psi(\mathbf{p}, \mathbf{R}), \qquad N(\mathbf{p}_0) = 1 \\ \bar\Psi_i(\mathbf{p}_0, \mathbf{R}) &= \Psi_i(\mathbf{p}_0, \mathbf{R}) + N_i(\mathbf{p}_0)\Psi(\mathbf{p}_0, \mathbf{R}) \end{aligned}$$

Note, $N_i = 0$ for linear parameters by definition. (If normal. depends on p_i, it is not linear.)

Dependence of parameter changes on normalization

Toulouse, CJU, JCP (2007,2008); CJU et al., PRL, **87**, 043401 (2007).

$$\bar{\Psi}(\mathbf{p},\mathbf{R}) = N(\mathbf{p})\,\Psi(\mathbf{p},\mathbf{R}), \quad N(\mathbf{p}_0)=1$$
$$\bar{\Psi}_i = \Psi_i + N_i\Psi_0$$

$$\Psi = \Psi_0 + \sum_{i=1}^{N_{\rm parm}} \delta p_i \Psi_i$$
$$\bar{\Psi} = \Psi_0 + \sum_{i=1}^{N_{\rm parm}} \delta \bar{p}_i \bar{\Psi}_i = \left(1 + \sum_{i=1}^{N_{\rm parm}} N_i \delta \bar{p}_i\right)\Psi_0. + \sum_{i=1}^{N_{\rm parm}} \delta \bar{p}_i \Psi_i$$

Since Ψ and $\bar{\Psi}$ are the optimal linear combin., they are the same aside from normalization

$$\delta p_i = \frac{\delta \bar{p}_i}{1+\sum_{i=1}^{N_{\rm parm}} N_i \delta \bar{p}_i} \quad \Longrightarrow \quad \delta \bar{p}_i = \frac{\delta p_i}{1-\sum_{i=1}^{N_{\rm parm}} N_i \delta p_i}. \tag{1}$$

One can get $\delta \bar{p}_i$ directly from solving the eigenvalue problem in the renormalized basis or get δp_i from eigenvalue problem in the original basis and use the above transformation. In either case, use $\delta \bar{p}_i$ to update the parameters, $p_i^{\rm new} = p_i + \delta \bar{p}_i$.

The denominator in Eq. 1 can be +ve, -ve or zero! So, predicted parameter changes can change sign depending on normalization!! If all parm. linear, $\delta \bar{p}_i = \delta p_i$, since all $N_i = 0$.

General semiorthogonalization

How to choose N_i ?

Toulouse, CJU, JCP (2007,2008); CJU et al., PRL, **87**, 043401 (2007).

$$\begin{aligned}\bar{\Psi}(\mathbf{p},\mathbf{R}) &= N(\mathbf{p})\,\Psi(\mathbf{p},\mathbf{R}), \qquad N(\mathbf{p}_0)=1 \\ \bar{\Psi}_i &= \Psi_i + N_i\Psi_0\end{aligned}$$

Choose N_i such that the $\bar{\Psi}_i$ are orthogonal to a linear combination of Ψ_0 and $\Psi_{\rm lin}$.

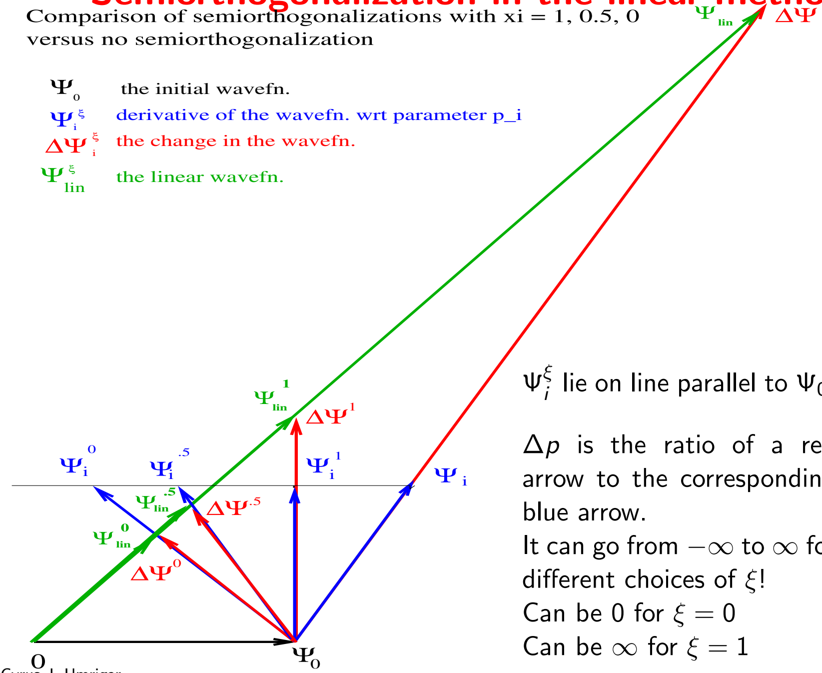
$$\left\langle \xi\frac{\Psi_0}{|\Psi_0|} + s(1-\xi)\frac{\Psi_{\rm lin}}{|\Psi_{\rm lin}|} \right| \left. \Psi_i + N_i\Psi_0 \right\rangle = 0$$

Solving for N_i we get $\left[s = 1(-1) \text{ if } \langle\Psi_0|\Psi_{\rm lin}\rangle = 1+\sum_j S_{0j}\Delta p_j > 0(<0)\right]$,

$$\begin{aligned}N_i &= -\frac{\xi D S_{0i} + s(1-\xi)(S_{0i}+\sum_j S_{ij}\Delta p_j)}{\xi D + s(1-\xi)(1+\sum_j S_{0j}\Delta p_j)} \\ \text{where} \qquad D &= \frac{|\Psi_{\rm lin}|}{|\Psi_0|} = \left(1+2\sum_j S_{0j}\Delta p_j + \sum_{i,j} S_{ij}\Delta p_i \Delta p_j\right)^{1/2}\end{aligned}$$

Semiorthogonalization in the linear method

Comparison of semiorthogonalizations with xi = 1, 0.5, 0
versus no semiorthogonalization



Ψ_i^ξ lie on line parallel to Ψ_0.

Δp is the ratio of a red arrow to the corresponding blue arrow.
It can go from $-\infty$ to ∞ for different choices of ξ!
Can be 0 for $\xi = 0$
Can be ∞ for $\xi = 1$

Semiorthogonalization in the linear method

Ψ_0 is the initial wave function, Ψ_i^ζ is the derivative of the wave function wrt parameter p_i for ζ. If superscript ζ is omitted that denotes that no semiorthogonalization is done. Then

$$\Psi_{\rm lin} \;=\; \Psi_0 + \sum_{i=1}^{N_{\rm parm}} \Delta\Psi_i^\zeta = \Psi_0 + \sum_{i=1}^{N_{\rm parm}} \Delta p_i^\zeta \Psi_i^\zeta, \quad \Delta p_i^\zeta = \frac{\Delta\Psi_i^\zeta}{\Psi_i^\zeta}$$

Note that $||\Delta\Psi^\zeta||$ is smallest for $\zeta=1$ and that $||\Psi_{\rm lin}^{0.5}|| = ||\Psi_0||$.
Also note that when there is just 1 parameter (can be generalized to >1):

1. In the limit that $\Psi_{\rm lin} \parallel \Psi_i$, $\Delta p_i = \pm\infty$
2. In the limit that $\Psi_{\rm lin} \perp \Psi_0$, $\Delta p_i^1 = \pm\infty$ because $\Delta\Psi^1 = \infty$, and, $\Delta p_i^0 = 0$ because $\Psi_i^0 = \infty$
3. $\Delta p_i^{0.5}$ is always finite

Note that Δp_i^ζ decreases as ζ decreases from 1 to 0. In Fig. 1, Δp_i is >1 for $\zeta=1$, and, <1 for $\zeta=0.5, 0$.
Also note that in Fig. 1 if we rotate $\Psi_{\rm lin}$ such that $\frac{\nabla\Psi\cdot\Psi_0}{||\nabla\Psi||||\Psi_0||} > \frac{\Psi_{\rm lin}\cdot\Psi_0}{||\Psi_{\rm lin}||||\Psi_0||}$ then Δp_i has the opposite sign as Δp_i^ζ!

Variance Minimization via Linear method

Toulouse, CJU, J. Chem. Phys., **128**, 174101 (2008)

Can one use the linear method to optimize the variance?

Variance Minimization via Linear method

Toulouse, CJU, J. Chem. Phys., **128**, 174101 (2008)

Can one use the linear method to optimize the variance?
Suppose we have some quadratic model of the energy variance to minimize

$$V_{\rm min} \quad = \quad \min_{\Delta\mathbf{p}} \left\{ V_0 + \mathbf{g}_V^{\rm T}\cdot\Delta\mathbf{p} + \frac{1}{2}\Delta\mathbf{p}^{\rm T}\cdot\mathbf{h}_V\cdot\Delta\mathbf{p} \right\}, \qquad (2)$$

where $V_0 = \langle\overline{\Psi}_0|(\hat{H}-E_0)^2|\overline{\Psi}_0\rangle$ is the energy variance of the current wave function $|\overline{\Psi}_0\rangle$, $\mathbf{g}_V$ is the gradient of the energy variance with components $g_{V,i} = 2\langle\overline{\Psi}_i|(\hat{H}-E_0)^2|\overline{\Psi}_0\rangle$ and $\mathbf{h}_V$ is some approximation to the Hessian matrix of the energy variance. Then, one could instead minimize the following rational quadratic model (*augmented hessian method*)

$$V_{\rm min} \quad = \quad \min_{\Delta\mathbf{p}} \frac{\begin{pmatrix} 1 & \Delta\mathbf{p}^{\rm T} \end{pmatrix} \begin{pmatrix} V_0 & \mathbf{g}_V^{\rm T}/2 \\ \mathbf{g}_V/2 & \mathbf{h}_V/2 + V_0\overline{\mathbf{S}} \end{pmatrix} \begin{pmatrix} 1 \\ \Delta\mathbf{p} \end{pmatrix}}{\begin{pmatrix} 1 & \Delta\mathbf{p}^{\rm T} \end{pmatrix} \begin{pmatrix} 1 & \mathbf{0}^{\rm T} \\ \mathbf{0} & \overline{\mathbf{S}} \end{pmatrix} \begin{pmatrix} 1 \\ \Delta\mathbf{p} \end{pmatrix}},$$

which agrees with the quadratic model in Eq. (2) up to second order in $\mathbf{\Delta p}$, and which leads to the following generalized eigenvalue equation

$$\begin{pmatrix} V_0 & \mathbf{g}_V^{\rm T}/2 \\ \mathbf{g}_V/2 & \mathbf{h}_V/2 + V_0\overline{\mathbf{S}} \end{pmatrix} \begin{pmatrix} 1 \\ \Delta\mathbf{p} \end{pmatrix} = V_{\rm min} \begin{pmatrix} 1 & \mathbf{0}^{\rm T} \\ \mathbf{0} & \overline{\mathbf{S}} \end{pmatrix} \begin{pmatrix} 1 \\ \Delta\mathbf{p} \end{pmatrix}.$$

Hence, we can use linear method to optimize a linear combination of energy and variance!

Connection between Linear and Newton methods

Toulouse, CJU, J. Chem. Phys., **128**, 174101 (2008)

In semiorthogonal basis with $\xi = 1$, linear eqs. are:

$$\begin{pmatrix} E_0 & \mathbf{g}^{\mathsf{T}}/2 \\ \mathbf{g}/2 & \overline{\mathbf{H}} \end{pmatrix} \begin{pmatrix} 1 \\ \Delta\mathbf{p} \end{pmatrix} = E_{\rm lin} \begin{pmatrix} 1 & \mathbf{0}^{\mathsf{T}} \\ \mathbf{0} & \overline{\mathbf{S}} \end{pmatrix} \begin{pmatrix} 1 \\ \Delta\mathbf{p} \end{pmatrix}, \tag{3}$$

Defining, $\Delta E = E_{\rm lin} - E_0 \le 0$, the 1^{st} and 2^{nd} eqs. are:

$$2\Delta E = \mathbf{g}^{\mathsf{T}} \cdot \Delta\mathbf{p}, \qquad 1^{st}\text{ eq.} \tag{4}$$

$$\frac{\mathbf{g}}{2} + \overline{\mathbf{H}}\Delta\mathbf{p} = E_{\rm lin}\overline{\mathbf{S}}\Delta\mathbf{p} \qquad 2^{nd}\text{ eq.} \tag{5}$$

$$\text{i.e.,}\quad 2\left(\overline{\mathbf{H}} - E_{\rm lin}\overline{\mathbf{S}}\right)\Delta\mathbf{p} = -\mathbf{g}, \tag{6}$$

This can be viewed as the Newton method with an approximate hessian, $\mathbf{h} = 2(\overline{\mathbf{H}} - E_{\rm lin}\overline{\mathbf{S}})$ which is nonnegative definite. (It has all nonnegative eigenvalues since we are subtracting out the lowest eigenvalue.) This also means that the linear method can be stabilized in much the same way as the Newton method.

Note that $2(\overline{\mathbf{H}} - E_{\rm lin}\overline{\mathbf{S}}) = 2(\overline{\mathbf{H}} - E_0\mathbf{S} - \Delta E\,\overline{\mathbf{S}})$ and $2(\overline{\mathbf{H}} - E_0\overline{\mathbf{S}})$ is the approximate hessian of Sorella's stochastic reconfiguration with approximate hessian (SRH) method (which converges more slowly that our linear and Newton methods). The present method provides an automatic stabilization of the SRH method by a positive definite matrix $-\Delta E\,\overline{\mathbf{S}}$ making the hessian nonnegative definite.

Stabilization

If far from the minimum, or, N_{MC}, is small, then the Hessian, $\bar{E}_{ij}$, need not be positive definite (whereas variance-minimization Levenberg-Marquardt $\bar{E}_{ij}$ is positive definite).

Even for positive definite $\bar{E}_{ij}$, the new parameter values may make the wave function worse if quadratic approximation is not good.

Add a_{diag} to the diagonal elements of the Hessian. This shifts the eigenvalues by the added constant. As a_{diag} is increased, the proposed parameter changes become smaller and rotate from the Newtonian direction to the steepest descent direction, but in practice a_{diag} is tiny.

The linear method and the perturbative method can be approximately recast into the Newton method. Consequently we can use the same idea for the linear and perturbative methods too.

Stabilization with Correlated Sampling

Each method has a parameter a_{diag} that automatically adjusts to make the method totally stable:

1. Do a MC run to compute the gradient and the Hessian (or overlap and Hamiltonian).
2. Using the above gradient and Hessian (or overlap and Hamiltonian), use 3 different values of a_{diag} to predict 3 different sets of updated parameters.
3. Do a short correlated sampling run for the 3 different wave functions to compute the energy differences for the 3 wave functions more accurately than the energies themselves.
4. Fit a parabola through the 3 energies to find the optimal a_{diag}.
5. Use this optimal a_{diag} to predict a new wave function, using the gradient and Hessian computed in step 1.
6. Loop back

Comparison of Newton, linear and perturbative methods

Programming effort and cost per iteration:

1. Newton method requires ψ, ψ_i, ψ_{ij}, $\hat{H}\psi$, $\hat{H}\psi_i$. ($\hat{H}\psi_{ij}$ removed by Hermiticity)
2. Linear method requires ψ, ψ_i, $\hat{H}\psi$, $\hat{H}\psi_i$.
3. Perturbative method requires ψ, ψ_i, $\hat{H}\psi$, $\hat{H}\psi_i$.
4. Perturbative method with approx. denom., and, SR require ψ, ψ_i, $\hat{H}\psi$.

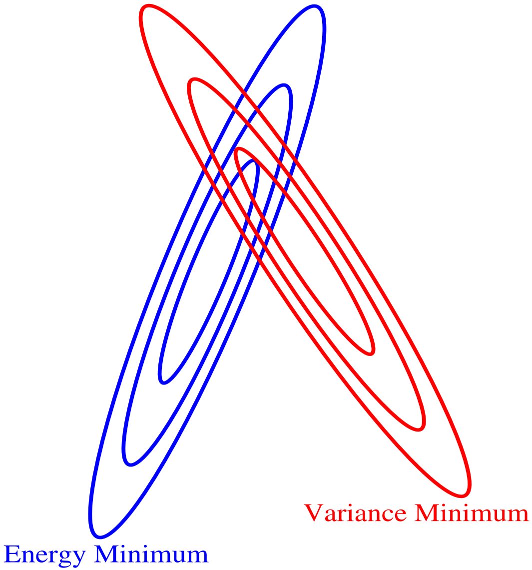
Convergence with number of iterations:

1. Newton and linear methods converge in 2-10 iterations for all parameters (CSF, orbital and Jastrow), but sometimes orbitals and exponents can take much longer.
2. SR method takes many more iterations to converge but can sometimes be more stable when there are many parameters.

Things to note

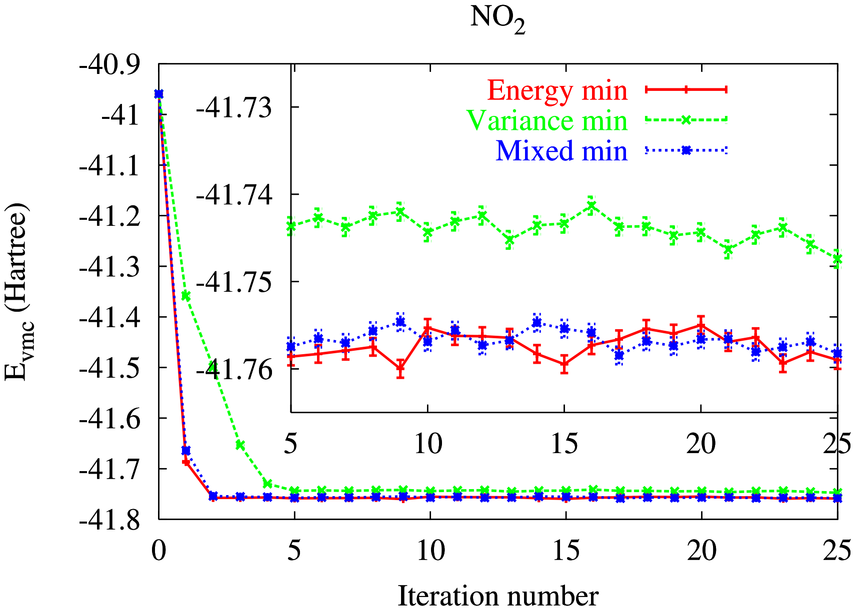
Eigenvalues of $\bar{E}_{ij}$ for Jastrow parameters typically span 10-12 orders of magnitude. So steepest descent would be horribly slow to converge!

Optimization of linear combination of energy and variance

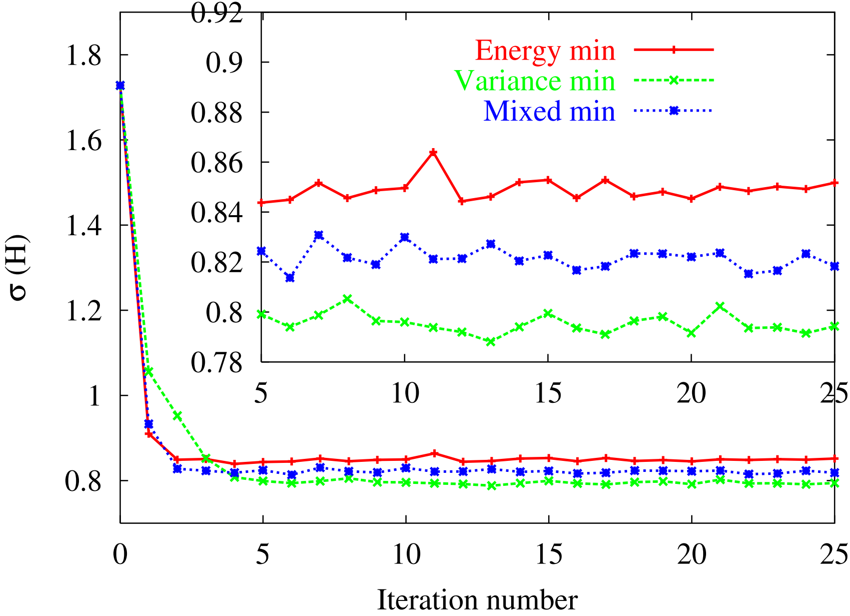


1. Can reduce the variance, without sacrificing appreciably the energy, by minimizing a linear combination, particularly since the ratio of hard to soft directions is 11 orders of magnitude.
2. Easy to do – obvious for Newton. Not obvious, but easy to do for linear method as shown above.
3. Measure of efficiency of the wave function is $\sigma^2 T_{\rm corr}$.

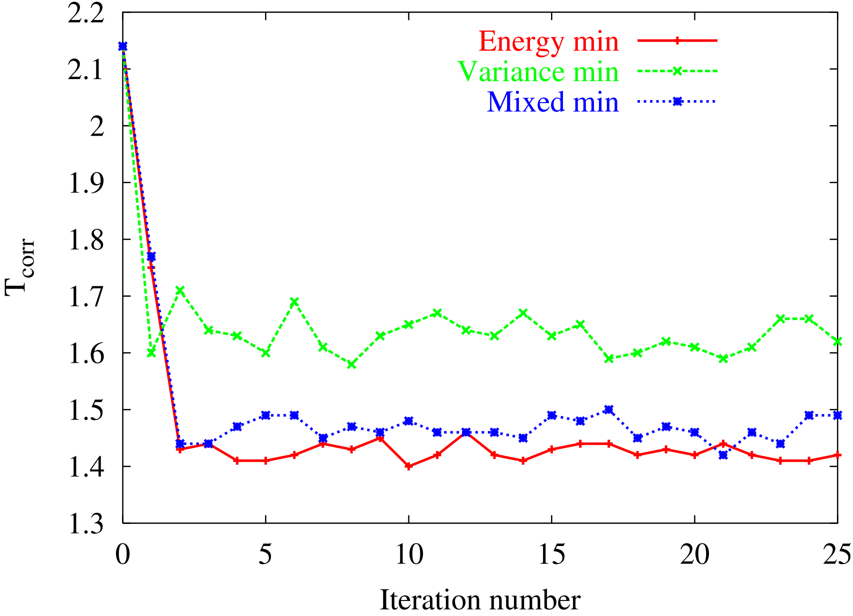
Convergence of energy of NO_2



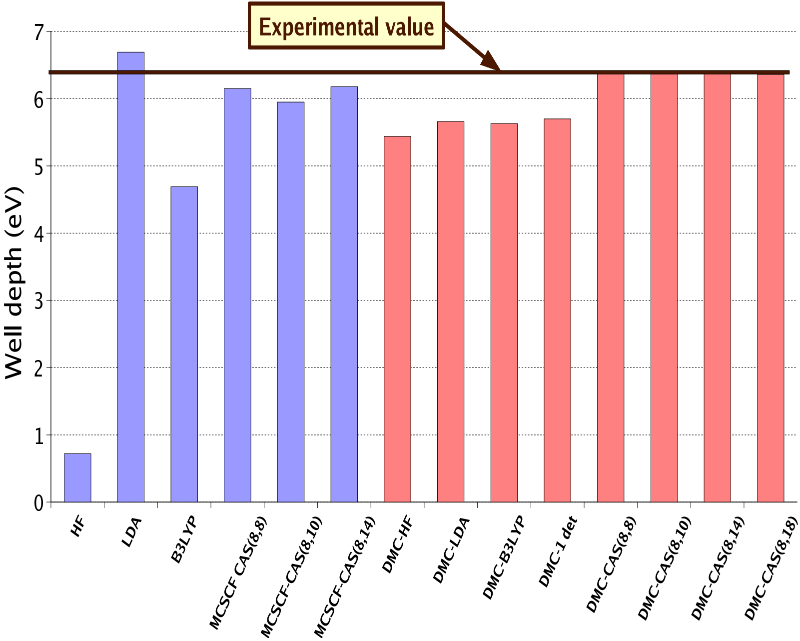
Convergence of energy fluctuations, σ, of NO_2



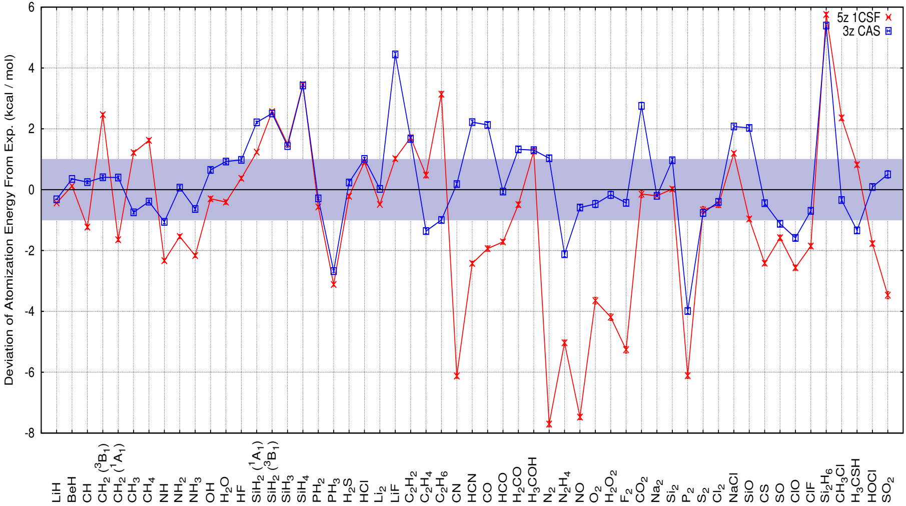
Convergence of autocorrelation time, T_{corr}, of NO_2



Well-depth of C_2

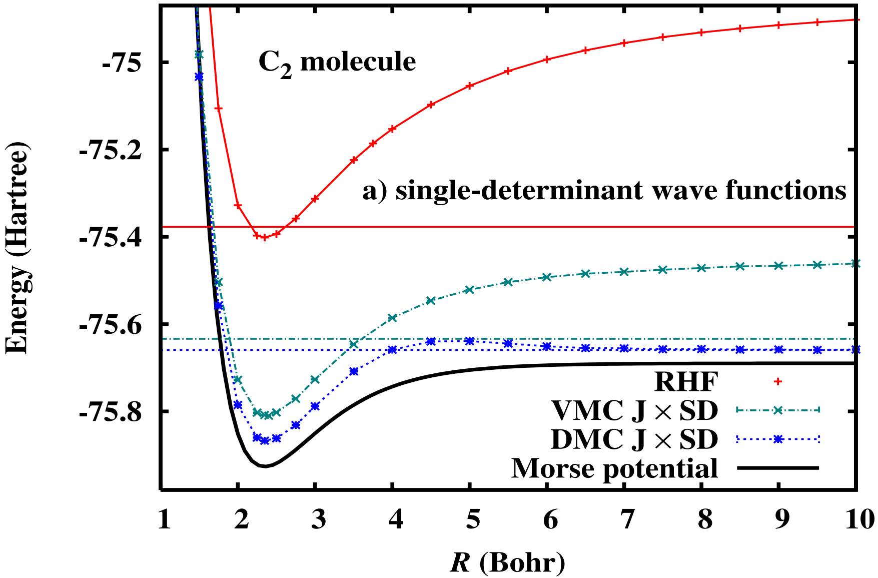


Atomization energies of the G2 set

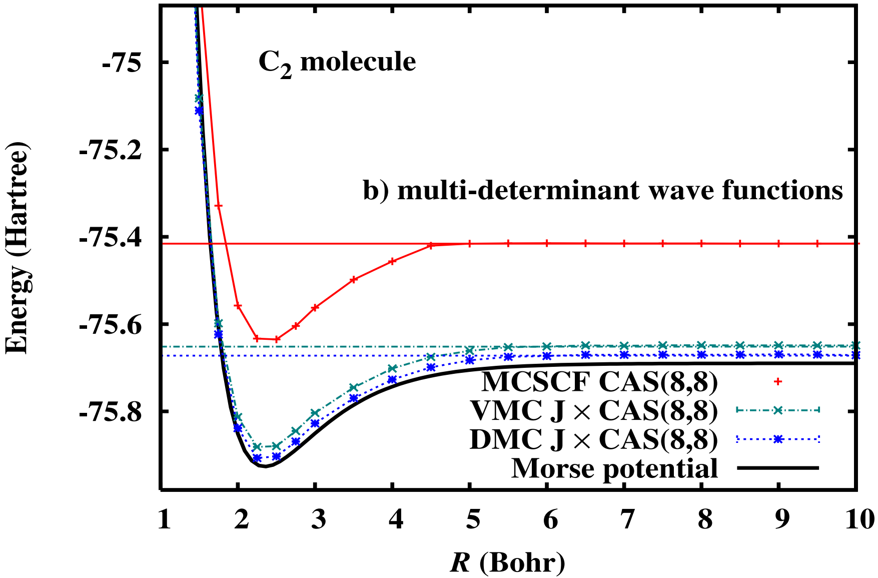


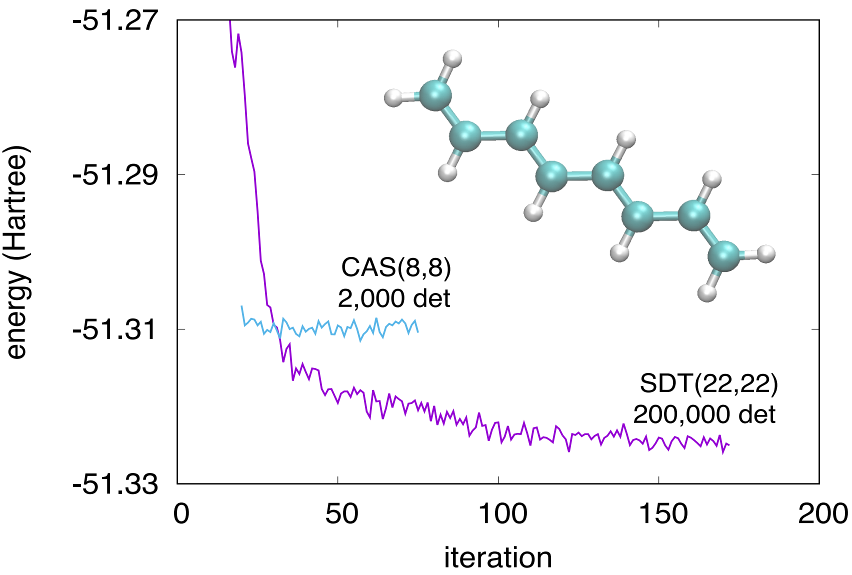
The mean absolute deviation from experiment for the DMC energies using the FV-CAS trial wave functions is 1.2 kcal/mole. Petruzielo, Toulouse, CJU, JCP 2012

C_2 potential energy curve - QMC with single det



C_2 potential energy curve - QMC with CAS-MCSCF





Optimization of wavefunction and geometry of C_8H_{10} (42 electrons) with 201924 determinants (all SDT in a space of 22 electrons in 22 orbitals) and 58652 parameters. 10 hours on 128 cores. Filippi, Assaraf, Moroni, arXiv 2017

Lecture 3: Projector MC and Path-integral MC

Projector Monte Carlo (zero T)

- Diffusion MC
- Full configuration interaction MC
- Phaseless Auxiliary-Field MC

Path Integral Monte Carlo

- Path Integral MC in real space (finite T)
- Path Integral MC in orbital space / Auxiliary Field MC / Determinantal MC (finite T)
- Path integral Ground State (PIGS) / Reptation MC (zero T)

Recap of Projector MC

Pure and Mixed estimators for energy are equal: $E_0 = \dfrac{\langle \Psi_0 | \hat{H} | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle} = \dfrac{\langle \Psi_0 | \hat{H} | \Psi_{\mathrm{T}} \rangle}{\langle \Psi_0 | \Psi_{\mathrm{T}} \rangle}$

Projector: $\quad |\Psi_0\rangle = \hat{P}(\infty)\, |\Psi_{\mathrm{T}}\rangle = \lim_{n\to\infty} \hat{P}^n(\tau)\, |\Psi_{\mathrm{T}}\rangle$

$$
\begin{aligned}
E_0 &= \frac{\langle \Psi_0 | \hat{H} | \Psi_{\mathrm{T}} \rangle}{\langle \Psi_0 | \Psi_{\mathrm{T}} \rangle} = \frac{\sum_{ij}^{N_{\mathrm{st}}} \langle \Psi_0 | \phi_i \rangle \ \langle \phi_i | \hat{H} | \phi_j \rangle \ \langle \phi_j | \Psi_{\mathrm{T}} \rangle}{\sum_k^{N_{\mathrm{st}}} \langle \Psi_0 | \phi_k \rangle \ \langle \phi_k | \Psi_{\mathrm{T}} \rangle} \\
&= \frac{\sum_{ij}^{N_{\mathrm{st}}} e_i H_{ij} t_j}{\sum_k^{N_{\mathrm{st}}} e_k t_k} = \sum_i^{N_{\mathrm{st}}} \frac{e_i t_i}{\sum_k^{N_{\mathrm{st}}} e_k t_k} \ \frac{\sum_j^{N_{\mathrm{st}}} H_{ij} t_j}{t_i} \\
&= \sum_i^{N_{\mathrm{st}}} \frac{e_i t_i}{\sum_k^{N_{\mathrm{st}}} e_k t_k} \ E_{\mathrm{L}}(i) \ = \ \frac{\left[\sum_i^{N_{\mathrm{MC}}} E_{\mathrm{L}}(i) \right]_{\Psi_{\mathrm{T}}\Psi_0}}{N_{\mathrm{MC}}} \xrightarrow{}_{\Psi_G \neq \Psi_T} \ \frac{\left[\sum_i^{N_{\mathrm{MC}}} \left(\frac{t_i}{g_i} \right) E_{\mathrm{L}}(i) \right]_{\Psi_{\mathrm{G}}\Psi_0}}{\left[\sum_k^{N_{\mathrm{MC}}} \left(\frac{t_k}{g_k} \right) \right]_{\Psi_{\mathrm{G}}\Psi_0}}
\end{aligned}
$$

Sample $e_i g_i / \sum_k^{N_{\mathrm{st}}} e_k g_k$ using *importance-sampled* projector.

For exact PMC, value indep. of Ψ_{T}, Ψ_{G}, statistical error depends on Ψ_{T}, Ψ_{G}.
(For FN-PMC, value depends on Ψ_{G}, statistical error on $\Psi_{\mathrm{T}}, \Psi_{\mathrm{G}}$.)
(For FN-DMC, value depends on nodes of Ψ_{G}, statistical error on $\Psi_{\mathrm{T}}, \Psi_{\mathrm{G}}$.)
Statistical error vanishes as $\Psi_{\mathrm{T}} \to \Psi_0$.
For fixed Ψ_T, $\Psi_G = \Psi_T$ does not minimize statistical fluctuations!

Projector MC

Projector: $|\Psi_0\rangle = \hat{P}(\infty)\,|\Psi_{\rm T}\rangle = \lim_{n\to\infty} \hat{P}^n(\tau)\,|\Psi_{\rm T}\rangle$

Projector is any function of the Hamiltonian that maps the ground state eigenvalue of $\hat{H}$ to 1, and the higher eigenvalues of $\hat{H}$ to absolute values that are < 1 (preferably close to 0).

Exponential projector: $\hat{P} = e^{\tau(E_T\hat{\mathbf{1}} - \hat{H})}$ (usually has time-step error)

Linear projector: $\hat{P} = \hat{\mathbf{1}} + \tau(E_T\hat{\mathbf{1}} - \hat{H})$ $\left(\tau < \frac{2}{E_{\rm max} - E_0}\right)$

Green's function projector: $\hat{P} = \dfrac{1}{\hat{\mathbf{1}} - \tau(E_T\hat{\mathbf{1}} - \hat{H})}$

Importance Sampling in Projector Monte Carlo

We want to sample from $g_i e_i = \langle \phi_i | \Psi_{\rm G} \rangle \langle \phi_i | \Psi_0 \rangle$ rather than $e_i = \langle \phi_i | \Psi_0 \rangle$.

If

$$\sum_j P_{ij} e_j \quad = \quad e_i$$

the similarity transformed matrix with elements $\tilde{P}_{ij} = \frac{g_i P_{ij}}{g_j}$ has eigenstate with elements $g_i e_i$:

$$\sum_j \tilde{P}_{ij} (g_j e_j) \quad = \quad \sum_j \left(\frac{g_i P_{ij}}{g_j} \right) (g_j e_j) \;=\; g_i e_i$$

$\tilde{P}_{ij}$ is called the *importance sampled* projector.

Projector vs Path Integral MC Methods

$$E_0 = \frac{\langle \Psi_\mathrm{T}|\hat{H}\hat{P}|\Psi_\mathrm{T}\rangle}{\langle \Psi_\mathrm{T}|\hat{P}|\Psi_\mathrm{T}\rangle}$$

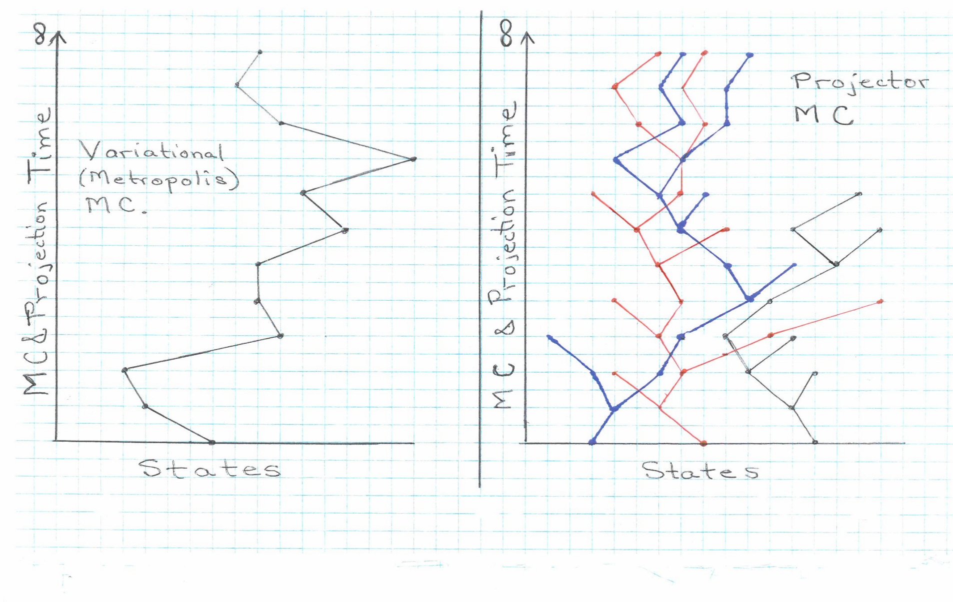
$$= \frac{\sum_{jik}^{N_\mathrm{st}} t_j H_{ji} P_{ik} t_k}{\sum_{ik}^{N_\mathrm{st}} t_i P_{ik} t_k^2} \qquad = \frac{\sum_{jik}^{N_\mathrm{st}} \frac{t_j H_{ji}}{t_i} \tilde{P}_{ik} t_k^2}{\sum_{ik}^{N_\mathrm{st}} \tilde{P}_{ik} t_k^2}$$

$$= \underbrace{\frac{\sum_{ik}^{N_\mathrm{st}} E_\mathrm{L}(i)\ t_i\ P_{ik}\ t_k}{\sum_{ik}^{N_\mathrm{st}} t_i\ P_{ik}\ t_k}}_{\text{not importance sampled}} \qquad = \underbrace{\frac{\sum_{ik}^{N_\mathrm{st}} E_\mathrm{L}(i)\ \tilde{P}_{ik}\ t_k^2}{\sum_{ik}^{N_\mathrm{st}} \tilde{P}_{ik}\ t_k^2}}_{\text{importance sampled}}$$

2 ways to turn these into MC expressions, e.g. for expression on left:

1. Since P is not column stochastic write $P_{ik} = w_{ik} T_{ik}$. Weighted walk. (Projector MC)
   1. Start run by sampling state k from $\langle \phi_k|\Psi_\mathrm{T}\rangle$.
   2. At each step sample i from $\langle \phi_i|\hat{P}|\phi_k\rangle$ and average $E_\mathrm{L}(i) t_i w_{ik}$ and $t_i w_{ik}$.
2. Since we have explicit expressions for $\langle \phi_k|\Psi_\mathrm{T}\rangle$ and $\langle \phi_i|\hat{P}|\phi_k\rangle$ use Metropolis-Hastings to sample entire path $t_i P_{ik} t_k$ and average $E_\mathrm{L}(i)$. Have to store the entire path. (Reptation MC (Path-integral MC for T=0))

Schematic of VMC and PMC

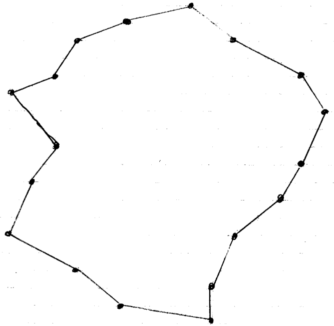


Schematic of PIMC and Reptation MC

Path-Integral M.C.

Average path length ∝ projection time/inverse T

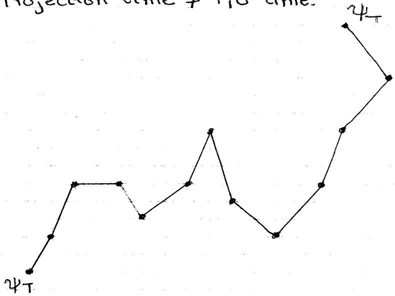
Projection time ≠ M.C. time



Reptation M.C.

Average path length ∝ projection time

Projection time ≠ MC time.



Projector vs Path Integral MC Methods

Projector MC

Open-ended weighted random walks (uses branching and population control for efficiency).
Easy to achieve essentially zero temperature.
Requires "forward walking" to compute expectation values of observables that do not commute with $\hat{H}$.

Path integral MC

No weights needed, but have to store the entire path and can be expensive to get to zero temperature (even with Reptation MC).
Efficient for computing expectation values of observables that do not commute with $\hat{H}$.

Zoo of Quantum Monte Carlo methods

There are a large number of QMC methods with a bewildering array of names, but just like a Chipotle wrap they are comprised of a few ingredients.

Chipotle wrap

white rice or brown rice
mild or medium or hot salsa
steak or carnitas or chicken or sofritas

QMC

zero temperature or finite temperature
linear projector or exponential projector
first quantized or second quantized
discrete time or continuous time
finite basis (site, Gaussian, planewave, ...) or infinite basis (real-space)
fixed-node or release-node
constrained-path or phaseless or free projection
finite path with Metropolis or open-ended walk with branching
pure estimator or mixed estimator or extrapolated estimator
single site or cluster or loop or worm updates

In these lectures we will see what all of the above mean (except the last line).

Taxonomy of Projector Monte Carlo Methods

The amplitudes of Ψ_0 in the chosen basis are obtained by using a "Projector", $\hat{P}$, that is a function of the Hamiltonian, $\hat{H}$, and has Ψ_0 as its dominant state.

Various Projector Monte Carlo Methods differ in:
a) form of the projector, and,
b) space in which the walk is done (single-particle basis and quantization).
(1^{st}-quantized $\equiv$ unsymmetrized basis, 2^{nd}-quantized $\equiv$ antisymmetrized basis.)

Method	Projector	SP Basis	Quantiz
Diffusion Monte Carlo	$e^{\tau(E_T\hat{\mathbf{1}}-\hat{H})}$	$\mathbf{r}$	1^{st}
GFMC (Kalos, Ceperley, Schmidt)	$e^{\tau(E_T\hat{\mathbf{1}}-\hat{H})}$ (samp. τ)	$\mathbf{r}$	1^{st}
LRDMC (Sorella, Casula)	$e^{\tau(E_T\hat{\mathbf{1}}-\hat{H})}$ (samp. τ)	$\mathbf{r}_i$	1^{st}
FCIQMC (Alavi, Booth)	$\hat{\mathbf{1}}+\tau(E_T\hat{\mathbf{1}}-\hat{H})$	$\phi_i^{\rm orthog}$	2^{nd}
phaseless AFQMC (Zhang, Krakauer)	$e^{\tau(E_T\hat{\mathbf{1}}-\hat{H})}$	$\phi_i^{\rm nonorthog}$	2^{nd}

$1+\tau(E_T\hat{\mathbf{1}}-\hat{H})$ can be used only if the spectrum of $\hat{H}$ is bounded, $\tau < \frac{2}{E_{\rm max}-E_0}$.

Linear Projector in a Discrete Space

$\hat{P} = \hat{\mathbf{1}} + \tau(E_T\hat{\mathbf{1}} - \hat{H})$, space is: 2^{nd}-quant. space of $\phi_i^{\rm orthog}$, i.e., determinants

e.g. Full Configuration Interaction Quantum Monte Carlo (FCIQMC)
Booth, Thom, Alavi, JCP (2009), Cleland, Booth, Alavi, JCP (2010)

States are represented as bit-packed orbital occupation numbers.

Although Hilbert space can be huge, since $\hat{H}$ and therefore $\hat{P}$ is sparse in the chosen basis, it is possible to sample from all connected states.

1. Starting from state i, sample state $j \neq i$ with probability T_{ji}.
   ($T_{ji} \neq 0$, if $P_{ji} \neq 0$)
2. Reweight state j by P_{ji}/T_{ji}
3. Reweight state i by P_{ii}
4. Branch states with appropriate probabilities to have unit weight walkers.

If this were the entire algorithm, there would be a fatal sign problem.
Discuss this later.

Diffusion Monte Carlo

i.e., $\hat{P}(\tau) = \exp(\tau(E_T - \hat{H}))$, $\quad |\phi_i\rangle = |\mathbf{R}\rangle$ **, walkers are 1^{st}-quantized**

$$G(\mathbf{R}', \mathbf{R}, t) \equiv P(\mathbf{R}', \mathbf{R}, t) = \langle \mathbf{R}' | e^{-\hat{H}t} | \mathbf{R} \rangle.$$

Diffusion Monte Carlo

Do not know $\langle \mathbf{R}'|e^{-\hat{H}t}|\mathbf{R}\rangle$ exactly, but can evaluate $G_{\rm rewt}(\mathbf{R}',\mathbf{R},t) \equiv \langle \mathbf{R}'|e^{-\hat{V}t}|\mathbf{R}\rangle$ exactly, if $\hat{V}$ is local in real space, $G_{\rm diffusion}(\mathbf{R}',\mathbf{R},t) \equiv \langle \mathbf{R}'|e^{-\hat{T}t}|\mathbf{R}\rangle$ exactly by introducing momentum representation
But, $e^{-t\hat{H}} \neq e^{-t\hat{T}}e^{-t\hat{V}}$, so repeatedly use short-time approx., $t = M\tau$.

$$\begin{aligned} G(\mathbf{R}',\mathbf{R},t) &= \langle \mathbf{R}'| \left(e^{-t\hat{H}/M}\right)^M |\mathbf{R}\rangle \\ &= \int d\mathbf{R}_1 \cdots d\mathbf{R}_{M-1} \, \langle \mathbf{R}'|e^{-\tau\hat{H}}|\mathbf{R}_{M-1}\rangle \, \langle \mathbf{R}_{M-1}|e^{-\tau\hat{H}}|\mathbf{R}_{M-2}\rangle \, \cdots \, \langle \mathbf{R}_1|e^{-\tau\hat{H}}|\mathbf{R}\rangle \end{aligned}$$

$$\begin{aligned} e^{-\tau\hat{H}} &\approx e^{-\tau\hat{T}}e^{-\tau\hat{V}} + \mathcal{O}(\tau^2) \qquad \text{Trotter breakup} \\ e^{-\tau\hat{H}} &\approx e^{-\tau\hat{V}/2}e^{-\tau\hat{T}}e^{-\tau\hat{V}/2} + \mathcal{O}(\tau^3) \end{aligned}$$

Since the potential energy is diagonal in position space, introducing 2 resolutions of the identity,

$$\langle \mathbf{R}'|e^{-\tau\hat{H}}|\mathbf{R}\rangle = e^{-\tau(V(\mathbf{R}')+V(\mathbf{R}))/2} \underbrace{\langle \mathbf{R}'|e^{-\tau\hat{T}}|\mathbf{R}\rangle}_{\text{next viewgr}} + \mathcal{O}(\tau^3)$$

Short-time Green's function – Kinetic term

Since the kinetic energy is diagonal in momentum space, we can evaluate $\langle \mathbf{R}' | e^{-\tau \hat{T}} | \mathbf{R} \rangle$ by introducing complete sets of momentum eigenstates

$$
\begin{aligned}
\langle \mathbf{R}' | e^{-\tau \hat{T}} | \mathbf{R} \rangle &= \int d\mathbf{P}'\, d\mathbf{P}\ \left\langle \mathbf{R}' | \mathbf{P}' \right\rangle\ \langle \mathbf{P}' | e^{-\tau \hat{T}} | \mathbf{P} \rangle\ \langle \mathbf{P} | \mathbf{R} \rangle \\
&= \frac{1}{(2\pi\hbar)^{3N}} \int d\mathbf{P}'\, d\mathbf{P}\ e^{\frac{-i\mathbf{P}'\cdot\mathbf{R}'}{\hbar}}\ \delta(\mathbf{P}' - \mathbf{P})\ e^{\frac{-\tau P^2}{2m}}\ e^{\frac{i\mathbf{P}\cdot\mathbf{R}}{\hbar}} \\
&= \frac{1}{(2\pi\hbar)^{3N}} \int d\mathbf{P}\ e^{\frac{i\mathbf{P}\cdot(\mathbf{R}-\mathbf{R}')}{\hbar}}\ e^{\frac{-\tau P^2}{2m}} \\
&= \frac{1}{(2\pi\hbar)^{3N}} e^{\frac{-m(\mathbf{R}-\mathbf{R}')^2}{2\hbar^2\tau}} \int d\mathbf{P}\ e^{\frac{\tau}{2m}(i\mathbf{P}+\frac{m}{\hbar\tau}(\mathbf{R}-\mathbf{R}'))^2} \\
&= \left(\frac{m}{2\pi\hbar^2\tau}\right)^{\frac{3N}{2}}\ e^{-\frac{m}{2\hbar^2\tau}(\mathbf{R}-\mathbf{R}')^2} \\
&= \frac{e^{-\frac{1}{2\tau}(\mathbf{R}-\mathbf{R}')^2}}{(2\pi\tau)^{\frac{3N}{2}}} \qquad 3N\text{-dim gaussian of width } \sqrt{\tau} \text{ in a.u.}
\end{aligned}
$$

Diffusion Monte Carlo – Short-time Green's function

Putting the two pieces together

$$G(\mathbf{R}',\mathbf{R},\tau) \;=\; \langle \mathbf{R}' | e^{\tau(E_{\mathrm{T}}-\hat{H})} | \mathbf{R} \rangle \approx \frac{1}{(2\pi\tau)^{3N/2}} e^{\left[-\frac{(\mathbf{R}'-\mathbf{R})^2}{2\tau} + \left\{E_{\mathrm{T}} - \frac{(\mathcal{V}(\mathbf{R}')+\mathcal{V}(\mathbf{R}))}{2}\right\}\tau\right]}$$

Diffusion Monte Carlo – Short-time Green's function

Can get the same result directly from the imaginary time Schrödinger Eq:

$$-\frac{1}{2}\nabla^2\psi(\mathbf{R},t) \;+\; (\mathcal{V}(\mathbf{R})-E_{\mathrm{T}})\psi(\mathbf{R},t) \;=\; -\frac{\partial\psi(\mathbf{R},t)}{\partial t}$$

Combining the diffusion Eq. and the rate Eq. Green's functions:

$$G(\mathbf{R}^{'},\mathbf{R},\tau) \;\approx\; \frac{1}{(2\pi\tau)^{3N/2}}e^{\left[-\frac{(\mathbf{R}^{'}-\mathbf{R})^2}{2\tau}+\left\{E_{\mathrm{T}}-\frac{(\mathcal{V}(\mathbf{R}^{'})+\mathcal{V}(\mathbf{R}))}{2}\right\}\tau\right]}$$

The wavefunction, $\psi(\mathbf{R}^{'},t+\tau)$, evolves according to the integral eqquation,

$$\psi(\mathbf{R}^{'},t+\tau) \;=\; \int d\mathbf{R}\, G(\mathbf{R}^{'},\mathbf{R},\tau)\psi(\mathbf{R},t).$$

Columns of $G(\mathbf{R}^{'},\mathbf{R},\tau)$ are not normalized to 1, so weights and/or branching are needed.
The potential energy $\mathcal{V}$ can diverge to $\pm\infty$, so the fluctuations in the weights and/or population are huge!

Diffusion Monte Carlo – Importance Sampled Green's Function

Importance sampling: Multiply imaginary-time the Schrödinger equation

$$-\frac{1}{2}\nabla^2\Psi(\mathbf{R},t) \;+\; (\mathcal{V}(\mathbf{R}) - E_{\mathrm{T}})\Psi(\mathbf{R},t) \;=\; -\frac{\partial\Psi(\mathbf{R},t)}{\partial t}$$

by $\Psi_{\mathrm{T}}(\mathbf{R})$ and rearranging terms we obtain

$$-\frac{\nabla^2}{2}(\Psi\Psi_{\mathrm{T}}) \;+\; \nabla\cdot\left(\frac{\nabla\Psi_{\mathrm{T}}}{\Psi_{\mathrm{T}}}\Psi\Psi_{\mathrm{T}}\right) \;+\; \Bigg(\underbrace{\frac{-\nabla^2\Psi_{\mathrm{T}}}{2\Psi_{\mathrm{T}}} + \mathcal{V}}_{E_{\mathrm{L}}(\mathbf{R})} - E_{\mathrm{T}}\Bigg)(\Psi\Psi_{\mathrm{T}}) \;=\; -\frac{\partial(\Psi\Psi_{\mathrm{T}})}{\partial t}$$

defining $f(\mathbf{R},t) = \Psi(\mathbf{R},t)\Psi_{\mathrm{T}}(\mathbf{R})$, this is

$$\underbrace{-\frac{1}{2}\nabla^2 f}_{\text{diffusion}} \;+\; \underbrace{\nabla\cdot\left(\frac{\nabla\Psi_{\mathrm{T}}}{\Psi_{\mathrm{T}}}f\right)}_{\text{drift}} \;+\; \underbrace{(E_{\mathrm{L}}(\mathbf{R}) - E_{\mathrm{T}})\,f}_{\text{growth/decay}} \;=\; -\frac{\partial f}{\partial t}$$

Since we know the exact Green function for any one term on LHS, an approximation is:

$$\tilde{G}(\mathbf{R}',\mathbf{R},\tau) \approx \frac{1}{(2\pi\tau)^{3N/2}}\, e^{\left[-\frac{(\mathbf{R}'-\mathbf{R}-\mathbf{V}(\mathbf{R})\tau)^2}{2\tau} + \left\{E_{\mathrm{T}} - \frac{(E_{\mathrm{L}}(\mathbf{R}')+E_{\mathrm{L}}(\mathbf{R}))}{2}\right\}\tau\right]}$$

Short-time Green's function – Drift term

$$-\frac{\partial f(\mathbf{R},t)}{\partial t} = \nabla\cdot(\mathbf{V}(\mathbf{R})f(\mathbf{R},t)) \approx \mathbf{V}\cdot\nabla f(\mathbf{R},t)$$

for $\mathbf{V}(\mathbf{R}) = \frac{\nabla\psi_{\mathrm{T}}(\mathbf{R})}{\psi_{\mathrm{T}}(\mathbf{R})} \approx \mathbf{V}$ (indep. of $\mathbf{R}$). Then, eigenfns. of $\mathbf{V}\cdot\nabla$ are

$$\frac{e^{i\mathbf{P}\cdot\mathbf{R}}}{(2\pi)^{\frac{3N}{2}}} \qquad \text{momentum eigenstates as for K.E.}$$

with eigenvalue $i\mathbf{V}\cdot\mathbf{P}$, so that

$$\begin{aligned} G_{\mathrm{drift}}(\mathbf{R}',\mathbf{R},\tau) &\approx \frac{1}{(2\pi)^{3N}}\int d\mathbf{P}\; e^{i\mathbf{P}\cdot\mathbf{R}'}\; e^{-i\mathbf{V}\cdot\mathbf{P}\tau}\; e^{-i\mathbf{P}\cdot\mathbf{R}} \\ &= \frac{1}{(2\pi)^{3N}}\int d\mathbf{P}\; e^{i\mathbf{P}\cdot(\mathbf{R}'-\mathbf{R}-\mathbf{V}\tau)} \\ &= \delta(\mathbf{R}'-\mathbf{R}-\mathbf{V}\tau) \end{aligned}$$

Assuming $\mathbf{V}$ is constant (the velocity at the initial point, $\mathbf{V}(\mathbf{R})$), is a bad approximation near singularities. The exact G_{drift} replaces $\mathbf{V}\equiv\mathbf{V}(\mathbf{R})$ by $\int_0^\tau dt\; \mathbf{V}(\mathbf{R}(t))$ over the path and this is done approximately, at no extra computational cost, in UNR93.

Diffusion Monte Carlo with Importance Sampling

Putting the drift, diffusion and reweighting Green's functions together,

$$\tilde{G}(\mathbf{R}', \mathbf{R}, \tau) \approx \int d\mathbf{R}'' \; G_{\rm rew}(\mathbf{R}', \frac{\tau}{2}) \; G_{\rm dif}(\mathbf{R}', \mathbf{R}'', \tau) \; G_{\rm dri}(\mathbf{R}'', \mathbf{R}, \tau) \; G_{\rm rew}(\mathbf{R}', \frac{\tau}{2})$$

$$\boxed{\tilde{G}(\mathbf{R}', \mathbf{R}, \tau) \approx \frac{1}{(2\pi\tau)^{3N/2}} e^{\left[-\frac{(\mathbf{R}'-\mathbf{R}-\mathbf{V}(\mathbf{R})\tau)^2}{2\tau} + \left\{E_{\rm T} - \frac{(E_{\rm L}(\mathbf{R}') + E_{\rm L}(\mathbf{R}))}{2}\right\}\tau\right]}}$$

The importance-sampled Green function has $E_{\rm L}(\mathbf{R})$ in the reweighting factor, which behaves MUCH better than the potential, $V(\mathbf{R})$. $V(\mathbf{R})$ diverges to $\pm\infty$ at particle coincidences whereas $E_{\rm L}(\mathbf{R})$ goes to a constant, E_0, as $\Psi_{\rm T} \to \Psi_0$. In addition it has a drift term that keeps the particles in the important regions, rather than relying on the reweighting to achieve that.

Even this does not always work. Why?

Diffusion Monte Carlo with Importance Sampling

Putting the drift, diffusion and reweighting Green's functions together,

$$\tilde{G}(\mathbf{R}', \mathbf{R}, \tau) \approx \int d\mathbf{R}''\ G_{\rm rew}(\mathbf{R}', \frac{\tau}{2})\ G_{\rm dif}(\mathbf{R}', \mathbf{R}'', \tau)\ G_{\rm dri}(\mathbf{R}'', \mathbf{R}, \tau)\ G_{\rm rew}(\mathbf{R}', \frac{\tau}{2})$$

$$\boxed{\tilde{G}(\mathbf{R}', \mathbf{R}, \tau) \approx \frac{1}{(2\pi\tau)^{3N/2}} e^{\left[-\frac{(\mathbf{R}'-\mathbf{R}-\mathbf{V}(\mathbf{R})\tau)^2}{2\tau} + \left\{E_{\rm T} - \frac{(E_{\rm L}(\mathbf{R}')+E_{\rm L}(\mathbf{R}))}{2}\right\}\tau\right]}}$$

The importance-sampled Green function has $E_{\rm L}(\mathbf{R})$ in the reweighting factor, which behaves MUCH better than the potential, $V(\mathbf{R})$. $V(\mathbf{R})$ diverges to $\pm\infty$ at particle coincidences whereas $E_{\rm L}(\mathbf{R})$ goes to a constant, E_0, as $\Psi_{\rm T} \to \Psi_0$. In addition it has a drift term that keeps the particles in the important regions, rather than relying on the reweighting to achieve that.

Even this does not always work. Why?
The above importance sampled Green function leads to an "infinite variance" estimate for systems other than Bosonic ground states!!

Singularities of Green's function

CJU, Nightingale, Runge, JCP 1993

Region	Local energy $E_{\rm L}$	Velocity $\mathbf{V}$
Nodes	$E_{\rm L} \sim \pm\frac{1}{R_\perp}$ for $\Psi_{\rm T}$ $E_{\rm L} = E_0$ for Ψ_0	$V \sim \frac{1}{R_\perp}$ for both $\Psi_{\rm T}$ and Ψ_0
e-n and e-e coincidences	$E_{\rm L} \sim \frac{1}{x}$ if cusps not imposed $E_{\rm L}$ finite if cusps are imposed $E_{\rm L} = E_0$ for Ψ_0	V has a discontinuity for both $\Psi_{\rm T}$ and Ψ_0

All the above infinities and discontinuities cause problems, e.g.,

$$\begin{aligned}\int_0^a dx\, E_{\rm L} &= \int_0^a dx \left(\frac{1}{x}\right) = \pm\infty \\ \int_0^a dx\, E_{\rm L}^2 &= \int_0^a dx \left(\frac{1}{x}\right)^2 = \infty\end{aligned}$$

Modify Green's function, by approximately integrating $E_{\rm L}$ and $\mathbf{V}$ over path, taking account of the singularities, at no additional computational cost.

Nonanalyticity of velocity near a node

CJU, Nightingale, Runge, JCP 1993

Linear approximation to $\Psi_{\rm T}$ (knowing $\mathbf{V} = \nabla\Psi_{\rm T}/\Psi_{\rm T}$):

$$\begin{aligned} \Psi_{\rm T}(\mathbf{R}') &= \Psi_{\rm T}(\mathbf{R}) + \nabla\Psi_{\rm T}(\mathbf{R})\cdot(\mathbf{R}'-\mathbf{R}) \\ &\propto 1 + \mathbf{V}\cdot(\mathbf{R}'-\mathbf{R}) \end{aligned}$$

The average velocity over the time-step τ is:

$$\bar{\mathbf{V}} = \frac{-1+\sqrt{1+2V^2\tau}}{V^2\tau}\mathbf{V} \to \begin{cases} \mathbf{V} & \text{if } V^2\tau \ll 1 \\ \sqrt{\frac{2}{\tau}}\hat{\mathbf{V}} & \text{if } V^2\tau \gg 1 \end{cases}$$

Discontinuity of velocity at particle coincidences

The e-N coincidence is more important than e-e coincidences because the wavefunction is larger in magnitude there.
Sample from linear combination of drifted Gaussians and exponential centered on nearest nucleus.

Infinite local energy near particle coincidences

Kato, Pure Appl. Math (1957), Pack and Byers-Brown, JCP, (1966), 2^{nd} order, Tew, JCP (2008)

Impose e-N and e-e cusp conditions on the wavefunction, so that divergence in potential energy is exactly canceled by divergence in kinetic energy.

$$\Psi = \sum_{l=0}^{\infty} \sum_{m=-l}^{l} r^l \, f_{lm}(r) \, Y_l^m(\theta, \phi)$$

$$f_{lm}(r) \approx f_{lm}^0 \left[1 + \frac{q_i q_j \mu_{ij} r}{l+1} + O(r^2) \right]$$

with f_{lm}^0 being the first term in the expansion of $f_{lm}(r)$.

Familiar example: e-N cusp for s-state of Hydrogenic atom is $-Z$. e-e cusps are $1/2$ and $1/4$ for $\uparrow\downarrow$ and $\uparrow\uparrow$ respectively. (This is why we chose two of the parameters in the wavefunction in the lab to be -2 and $1/2$.)

Combining with Metropolis to reduce time-step error

Reynolds, Ceperley, Alder, Lester, JCP 1982

$$\underbrace{-\frac{1}{2}\nabla^2 f}_{\text{diffusion}} + \underbrace{\nabla \cdot \left(\frac{\nabla \psi_{\mathrm{T}}}{\psi_{\mathrm{T}}} f\right)}_{\text{drift}} + \underbrace{\left(E_{\mathrm{L}}(\mathbf{R}) - E_{\mathrm{T}}\right) f}_{\text{growth/decay}} = -\frac{\partial f}{\partial t}$$

If we omit the growth/decay term then $|\Psi_{\mathrm{T}}|^2$ is the solution.

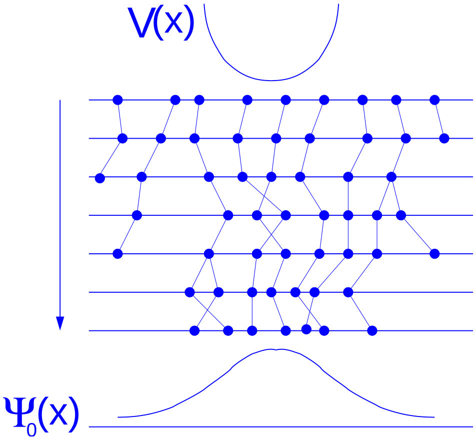
$$-\frac{1}{2}\nabla^2 \psi_{\mathrm{T}}^2(\mathbf{R}) + \nabla \cdot \left(\frac{\nabla \psi_{\mathrm{T}}}{\psi_{\mathrm{T}}} \psi_{\mathrm{T}}^2(\mathbf{R})\right) = 0$$

We can sample $|\Psi_{\mathrm{T}}|^2$ exactly using Metropolis-Hastings! So, view $G(\mathbf{R}', \mathbf{R}, t)$ as being the proposal matrix $T(\mathbf{R}', \mathbf{R})$ and introduce accept-reject step after drift and diffusion steps.

Since some of the moves are rejected, account for that approximately by reducing the time step in the reweighting factor from τ to τ_{eff}.
If accept/reject is done after each 1-electron move, then

$$\tau_{\mathrm{eff}} = \tau \frac{R^2_{\mathrm{accep}}}{R^2_{\mathrm{prop}}}$$

Branching random walks



Walkers multiply/die in regions of lower/higher V than E_{T} (no imp. sampling)
Walkers multiply/die in regions of lower/higher E_{L} than E_{T} (with imp. sampling)

Sign Problem in PMC

The sign problem differs for the various PMC methods.
However, in all PMC methods the underlying cause is that a state other than the desired state grows exponentially compared to the desired state, combined with the fact that since we are sampling states, cancellations of opposite sign contributions are relatively ineffective.

In DMC we saw that we sample not $\Psi_0(\mathbf{R})$ but $\Psi_{\rm T}(\mathbf{R})\Psi_0(\mathbf{R})$ using the importance-sampled projector.
This sneaks in the fixed-node approximation, since we are projecting onto the lowest state that has the same nodes $\Psi_{\rm T}(\mathbf{R})$ rather than the global ground state.

Sign Problem

Except for some special cases, there is a sign problem, and there is a FN error.

Diffusion Monte Carlo

Physical dimension of space	d
Number of parallel-spin electrons	N
Dimension of wavefunction	dN
Dimension of nodal surface	$dN-1$
Dimension of particle coincidences	$dN-d$

So, in 1-d the nodal surface is known (when particles cross) and DMC does not have a sign problem.

Another special case: AFQMC does not have a sign problem for the 1/2-filled Hubbard model.

What if we use the projector without importance sampling?

Fermion Nodes - a simple case

Consider a He atom in its 1^1S ground state. What are its nodes?

Fermion Nodes - a simple case

Consider a He atom in its 1^1S ground state. What are its nodes?

It has none!

Consider a He atom in its 1^3S state. What are its nodes?

Fermion Nodes - a simple case

Consider a He atom in its 1^1S ground state. What are its nodes?
It has none!

Consider a He atom in its 1^3S state. What are its nodes?
$r_1 = r_2$ $\quad$ ($\mathbf{r}_1 = \mathbf{r}_2$ is co-dimension 2 from $r_1 = r_2$.)
Proof: Suppose $r_1 = r_2$.
If we rotate by 180deg about line joining nucleus to the midpoint of the 2 electrons, $\Psi \to \Psi$.
If we exchange the electrons, $\Psi \to -\Psi$.
So, $\Psi = 0$ when $r_1 = r_2$.

Sign Problem in DMC

$\hat{P}(\tau) \;=\; e^{\tau(E_T\hat{\mathbf{1}}-\hat{H})}$. $|\phi_i\rangle = |\mathbf{R}\rangle$

$$\langle \mathbf{R}|\hat{P}(\tau)|\mathbf{R}'\rangle \approx \frac{e^{\frac{-\left(\mathbf{R}-\mathbf{R}'\right)^2}{2\tau}+\left(E_T-\frac{\mathcal{V}(\mathbf{R})+\mathcal{V}(\mathbf{R}')}{2}\right)\tau}}{(2\pi\tau)^{3N/2}}$$ is nonnegative.

So, where does the sign problem come from?

Sign Problem in DMC

$\hat{P}(\tau) = e^{\tau(E_T \hat{\mathbf{1}} - \hat{H})}$. $|\phi_i\rangle = |\mathbf{R}\rangle$

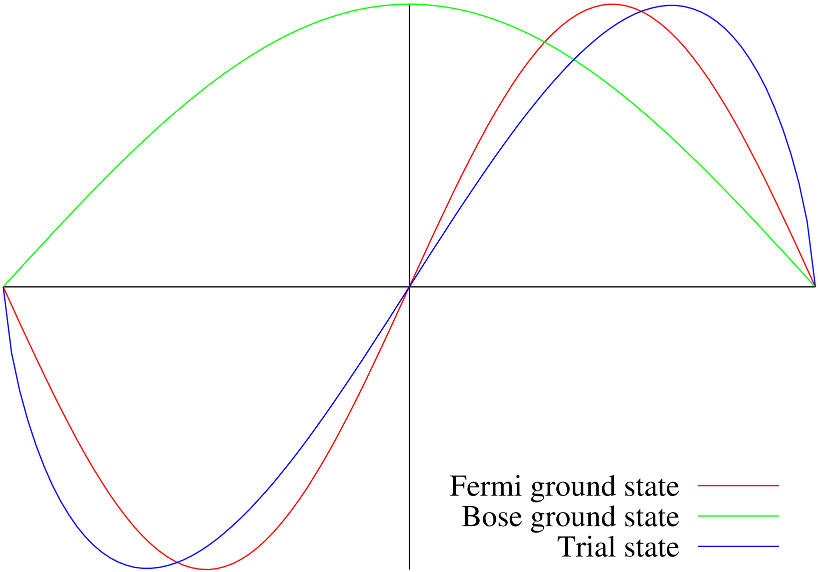
$$\langle \mathbf{R}|\hat{P}(\tau)|\mathbf{R}'\rangle \approx \frac{e^{\frac{-\left(\mathbf{R}-\mathbf{R}'\right)^2}{2\tau} + \left(E_T - \frac{\mathcal{V}(\mathbf{R})+\mathcal{V}(\mathbf{R}')}{2}\right)\tau}}{(2\pi\tau)^{3N/2}}$$ is nonnegative.

So, where does the sign problem come from?
Problem: Since the Bosonic energy is always lower than the Fermionic energy, the projected state is the Bosonic ground state.

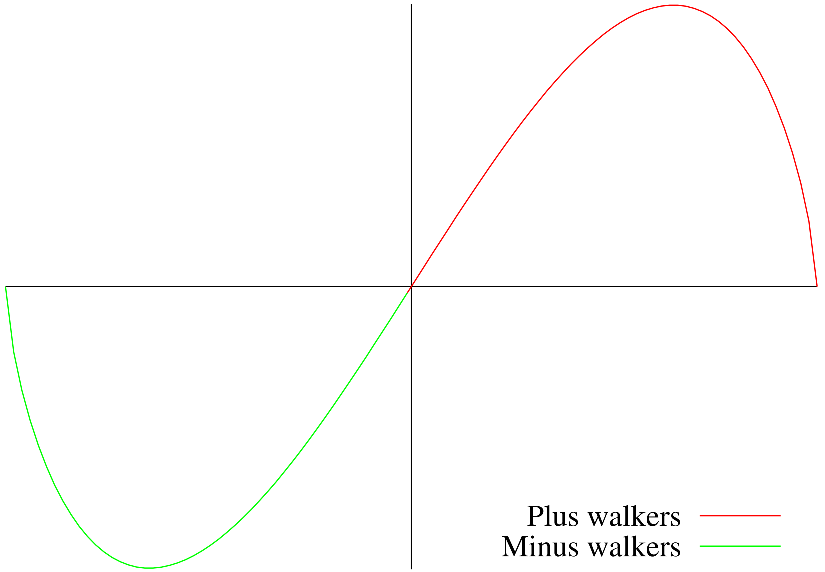
Fixed-node approximation

All except a few calculations (release-node, Ceperley) are done using FN approximation. Instead of doing a free projection, impose the boundary condition that the projected state has the same nodes as the trial state $\Psi_{\rm T}(\mathbf{R})$.
This gives an upper bound to the energy and becomes exact in the limit that $\Psi_{\rm T}$ has the same nodes as Ψ_0.

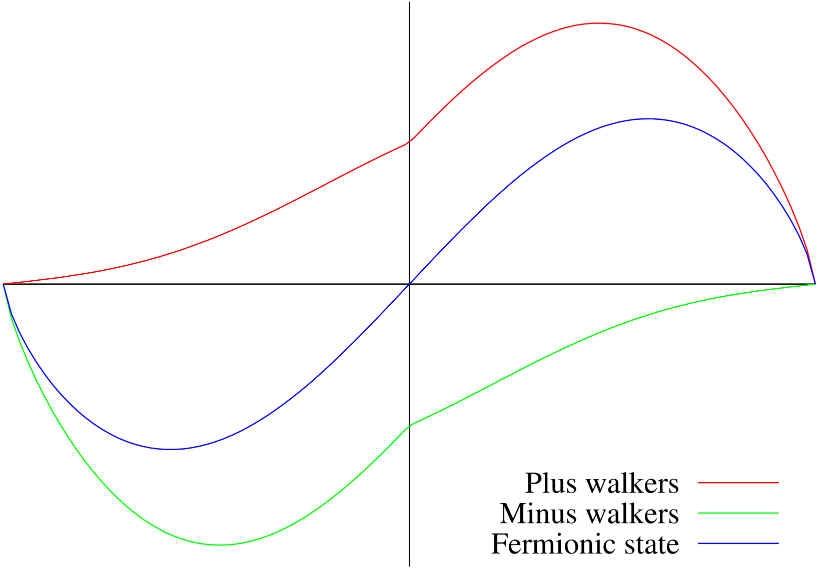
Sign Problem in 1^{st} Quantization and R space



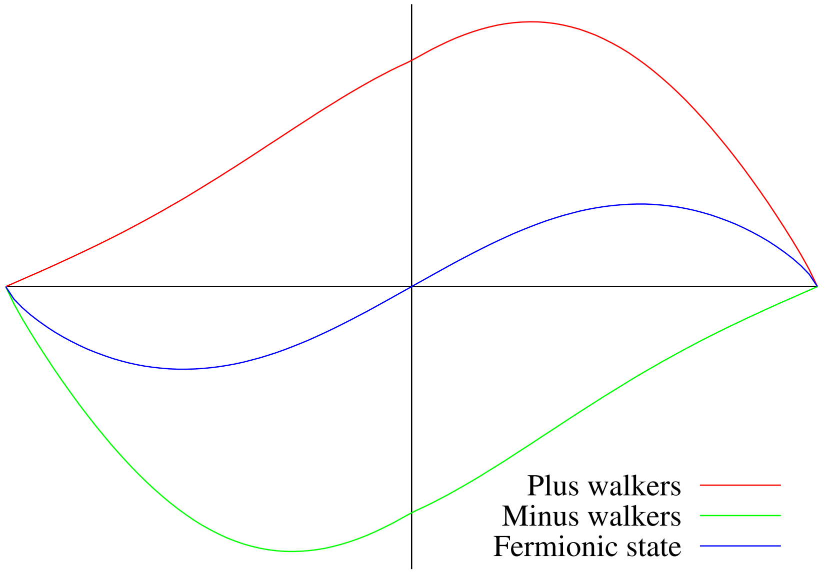
Sign Problem in 1^{st} Quantization and R space



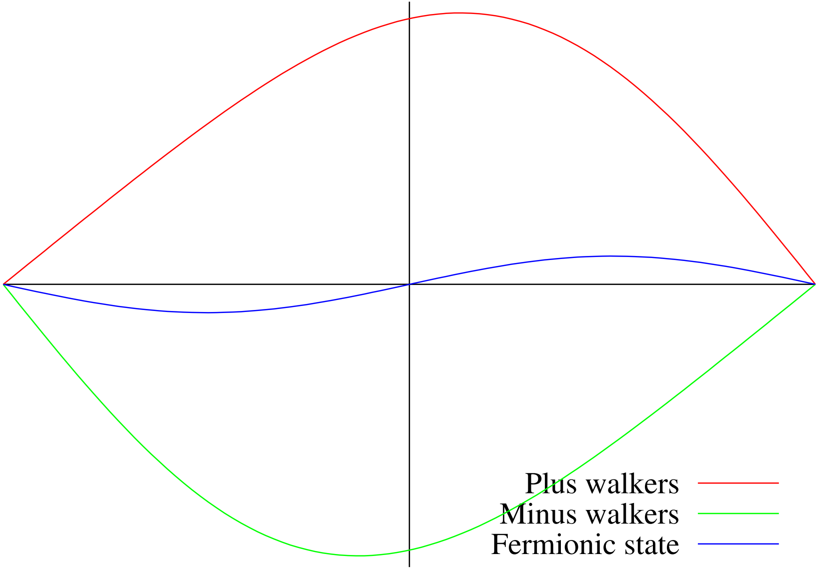
Sign Problem in 1^{st} Quantization and R space



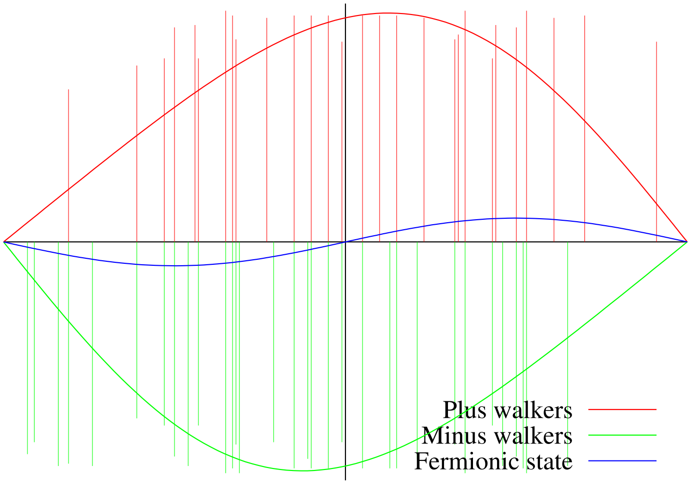
Sign Problem in 1^{st} Quantization and R space



Sign Problem in 1^{st} Quantization and R space



Sign Problem in 1^{st} Quantization and R space



Problem: In large space walkers rarely meet and cancel, so tiny signal/noise! Further, if there are many cancellations, eventually there will be exclusively walkers of one sign only and a purely Bosonic distribution.

Sign Problem in 2^{nd} quantization

It would appear from the above discussion that one could eliminate the sign problem simply by using an antisymmetrized basis. In that case there are no Bosonic states or states of any other symmetry than Fermionic, so there is no possibility of getting noise from non-Fermionic states. Is that the case?

Sign Problem in 2^{nd} quantization

It would appear from the above discussion that one could eliminate the sign problem simply by using an antisymmetrized basis. In that case there are no Bosonic states or states of any other symmetry than Fermionic, so there is no possibility of getting noise from non-Fermionic states. Is that the case?

No!

Sign Problem in 2^{nd} quantization

Walk is done in the space of determinants.

Since Bosonic and other symmetry states are eliminated, there is some hope of having a stable signal to noise, but there is still a sign problem.

Problem: Paths leading from state i to state j can contribute with opposite sign. Further, Ψ and $-\Psi$ are equally good.

The projector in the chosen 2^{nd}-quantized basis does not have a sign problem if:
The columns of the projector have the same sign structure aside from an overall sign, e.g.

$$P\Psi = \begin{bmatrix} + & - & + & + \\ - & + & - & - \\ + & - & + & + \\ + & - & + & + \end{bmatrix} \begin{bmatrix} + \\ - \\ + \\ + \end{bmatrix} = \begin{bmatrix} + \\ - \\ + \\ + \end{bmatrix}$$

or equivalently:
It is possible to find a set of sign changes of the basis functions such that all elements of the projector are nonnegative.

The sign problem is an issue only because of the stochastic nature of the algorithm. Walkers of different signs can be spawned onto a given state in different MC generations.

Comparison of DMC with S-FCIQMC

DMC (walk in electron coordinate space)	S-FCIQMC (walk in determin. space)
Severe Fermion sign problem due to growth of Bosonic component relative to Fermionic.	Less severe Fermion sign problem due to opposite sign walkers being spawned on the same determinant
Fixed-node approximation needed for stable algorithm. Exact if Ψ_{T} nodes exact.	Walker cancellation plus initiator approximation needed for stable algorithm. Exact in ∞-population limit.
Infinite basis.	Finite basis. (Same basis set dependence as in other quantum chemistry methods.
Computational cost is low-order polynomial in N	Computational cost is exponential in N but with much smaller exponent than full CI
Energy is variational	Energy not variational but DM variant is
Need to use pseudopotentials for large Z.	Can easily do frozen-core
Moderate T_{corr}	For large basis sets T_{corr} is very large

Reweighting, Branching/Reconfiguration, Population Control

1. If we have a single weighted walker, then a few generations of the walk will dominate and the computational effort expended on the rest of the walk would be largely wasted. So, efficiency demands a population of walkers, such that all walkers within a generation have roughly the same weight, say within a factor of 2.
2. Even so, the weights of different generations will vary a lot in a sufficiently long run. So, efficiency demands that we exercise population control to make the weights of each generation approximately the same. This results in a positive *population control bias*, but there is a simple method by Nightingale and Blote for removing most of it.

Reconfiguration methods (for making weights approximately equal)

Integerization (commonly used)

Replace each walker with weight w_i with $\lfloor w_i \rfloor$ or $\lceil w_i \rceil$ walkers with probabilities $\lceil w_i \rceil - w_i$ and $w_i - \lfloor w_i \rfloor$.

Advantage (minor): All walkers have unit weight
Disadvantage: Unnecessary fluctuations

Split-join method (CJU, Nightingale, Runge, JCP 1993)

Split walkers with $w > 2$ into 2 walkers each with weight $w_i/2$.
Join 2 walkers with weights w_1, w_2, keeping walker 1 or 2 with probabilities $w_1 = w_1/(w_1 + w_2)$, $w_2 = w_1/(w_1 + w_2)$ and giving the resultant walker weight $w_1 + w_2$.

Advantage: Avoids unnecessary fluctuations

Stochastic reconfiguration (Sorella)

Construct an array of cumulative probabilities, W_i.
Throw 1 random number in the interval $[0, W_n]$ and superimpose a comb with N evenly spaced tines on the cumulative probabilities.
Take as many copies of each state as the number of tines in that cumulative probability interval.
Each walker gets the same weight (the average).

Advantage: Constant number of walkers, and, avoids unnecessary fluctuations

In practice the 2^{nd} and 3^{rd} methods work equally well.

Population control error

The log of the weights of the generations will undergo a random walk and so some generations will have very small or very large weights. So, we have to exercise *population control* by dividing the weights by a generation-dependent fluctuating factor f.

$$f(t) = e^{\tau(E_T - E_{\rm est})} \left(\frac{W(t-1)}{W_{\rm target}} \right)^{1/g}$$

$e^{\tau(E_T - E_{\rm est})}$ compensates for an inaccurate E_T and $\left(\frac{W(t-1)}{W_{\rm target}} \right)^{1/g}$ tries to restore the population weight g generations later.
If we are using the exponential projector, $\exp((E_{\rm T} - \hat{H})\tau)$, this is equivalent to adjusting $E_{\rm T}$, but for the purpose of removing the population control error it is better to think in terms of a fixed $E_{\rm T}$ and a fluctuating factor f that needs to be corrected for.
Now the naive estimator for the energy is biased:

$$E_{\rm mix} \quad = \quad \frac{\sum_t^{N_{\rm gen}} \sum_i^{N_w(t)} w_i(t) E_{\rm L}(\mathbf{R}_i(t))}{\sum_t^{N_{\rm gen}} \ \sum_i^{N_w(t)} w_i(t)}$$

Population control error

The population control error is proportional to the inverse of the target population size $N_{\rm walk}$. The error arises because of a negative correlation between the energy averaged over the generation and the weight of the generation. When the energy is low, the weight tends to be large and population control artificially reduces the weight and thereby creates a positive bias in the energy. Similarly, when the energy is high, the weight tends to be small and population control artificially increases the weight and this too creates a positive bias in the energy. Since the relative fluctuations in the energy and in the weight go as $1/\sqrt{N_{\rm walk}}$, the relative fluctuations in their covariance goes as $1/N_{\rm walk}$.

So, one way to reduce the population control error is to simply use a large population, and this is what most people do. If one wishes to be sure that the error is sufficiently small, plot the energy versus $1/N_{\rm walk}$ and take the limit $1/N_{\rm walk} \to 0$. But there exists a better way that allows us to estimate and remove most of the population control error within a single run, as described next.

Removing the population control error

Nightingale and Bloete, PRB 1986;   CJU, Nightingale, Runge, JCP 1993

The basic idea for correcting the population control error is the following. When we do population control we have a generation-dependent factor by which we change the weights of all the walkers in that generation relative to what the mathematics tells us is correct. So, we keep track of these factors and when computing expectation values we undo these factors for the last several generations. If we undid them for the entire run then we would be back to our original problem, i.e., very large fluctuations in the weights. However, we only need to undo them for a number of generations corresponding to a few times the autocorrelation time to get rid of almost all of the population control bias. In the next viewgraph we explain how to do this and then in the following one we explain a continuous version of the algorithm that is even simpler to implement, though a bit harder to explain (which is why we do them in this order).

Removing the population control error

Nightingale and Bloete, PRB 1986; CJU, Nightingale, Runge, JCP 1993

Introduce popul. control factor $f(t) = e^{\tau(E_T - E_{\rm est})} \left(\frac{W(t-1)}{W_{\rm target}}\right)^{1/g}$ &

$P_i(t, T_p) = \prod_{p=0}^{T_p} f(t-p)$, where $E_{\rm est}$ is best current estim. of the energy, $W(t-1) = \sum_j w_j(t-1)$.

($e^{\tau(E_T - E_{\rm est})}$ compensates for an inaccurate E_T and $\left(\frac{W(t-1)}{W_{\rm target}}\right)^{1/g}$ tries to restore the population weight g generations later.)

The modified expressions for $E_{\rm mix}$ is:

$$E_{\rm mix} = \frac{\sum_t^{N_{\rm gen}} P_i(t, T_p) \ \sum_i^{N_w(t)} w_i(t) E_{\rm L}(\mathbf{R}_i(t))}{\sum_t^{N_{\rm gen}} P_i(t, T_p) \ \sum_i^{N_w(t)} w_i(t)}$$

This requires us to store a circular buffer of $T_{\rm p}$ population control factors $f(t)$ and iteratively compute the product $T_{\rm p}$.

Removing the population control error

Continuous method

A simpler procedure that does not require a circular buffer is to replace $P_i(t, T_{\rm p})$ by

$$P_i(t,p) = \cdots f(t-2)^{p^2}\ f(t-1)^{p^1}\ f(t)^{p^0} = \prod_{n=0}^{t-1} f(t-n)^{p^n}$$

with p a bit less than 1, calculated recursively at each generation using

$$P_i(t,p) = P_i(t-1,p)^p\ f(t),$$

$$E_{\rm mix} = \frac{\sum_t^{N_{\rm gen}} P_i(t,p)\ \sum_i^{N_w(t)} w_i(t) E_{\rm L}(\mathbf{R}_i(t))}{\sum_t^{N_{\rm gen}} P_i(t,p)\ \sum_i^{N_w(t)} w_i(t)}$$

A rough correspondence between $T_{\rm p}$ in the discrete method and p in the continuous method is established by requiring that $p^{T_{\rm p}} = 1/e$, i.e., $p = e^{-1/T_{\rm p}}$.

Expectation values of operators

We wish to compute the pure (as opposed to mixed) expectation value

$$\langle A \rangle_{\mathrm{pure}} = \frac{\langle \Psi_0 | \hat{A} | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle}$$

We consider various cases in order of increasing difficulty:

M.P. Nightingale, in Quantum Monte Carlo Methods in Physics and Chemistry, edited by M.P. Nightingale and CJU

1. $\hat{A}$ commutes with with $\hat{H}$, or equivalently $\hat{G}$, and is near-diagonal in chosen basis. (mixed expectation value)
2. $\hat{A}$ is diagonal in chosen basis. (forward/future walking) Liu, Kalos, and Chester, PRA (1974)
3. $\hat{A}$ is not diagonal in chosen basis, but, $A_{ij} \neq 0$ only when $G_{ij} \neq 0$. (forward/future walking)
4. $\hat{A}$ is not diagonal in chosen basis. (side walking) Barnett, Reynolds, Lester, JCP (1992)

Expectation values of operators

Factor the elements of the importance-sampled projector, $\tilde{G}(\mathbf{R}^{'},\mathbf{R})$, as products of elements of a stochastic matrix/kernel (elements are nonnegative and elements of column sum to 1), $\tilde{T}(\mathbf{R}^{'},\mathbf{R})$, and a reweight factor, $w(\mathbf{R}^{'},\mathbf{R})$.

$$\tilde{G}(\mathbf{R}^{'},\mathbf{R}) \;=\; \tilde{T}(\mathbf{R}^{'},\mathbf{R})w(\mathbf{R}^{'},\mathbf{R})$$

In the case of DMC

$$\begin{aligned}
\tilde{T}(\mathbf{R}^{'},\mathbf{R}) &= G_{\rm dif}(\mathbf{R}^{'},\mathbf{R}^{''})\; G_{\rm drift}(\mathbf{R}^{''},\mathbf{R}) \;=\; \frac{1}{(2\pi\tau)^{3N/2}} e^{-\frac{(\mathbf{R}^{'}-\mathbf{R}-\mathbf{V}\tau)^2}{2\tau}} \\
w(\mathbf{R}^{'},\mathbf{R}) &= e^{\left\{E_{\rm T}-\frac{(E_{\rm L}(\mathbf{R}^{'})+E_{\rm L}(\mathbf{R}))}{2}\right\}\tau}
\end{aligned}$$

For discrete state space and sparse H, define

$$\begin{aligned}
\tilde{T}(\mathbf{R}^{'},\mathbf{R}) &= \frac{\tilde{G}(\mathbf{R}^{'},\mathbf{R})}{\sum_{\mathbf{R}^{''}} \tilde{G}(\mathbf{R}^{''},\mathbf{R})} \\
w(\mathbf{R}^{'},\mathbf{R}) &= w(\mathbf{R}) \;=\; \sum_{\mathbf{R}^{''}} \tilde{G}(\mathbf{R}^{''},\mathbf{R})
\end{aligned}$$

1) $\hat{A}$ commutes with with $\hat{H}$ and is near-diagonal in chosen basis

By *near diagonal* we mean that either:

1. In discrete space $\hat{A}$ is sufficiently sparse that when walker is at state i, $A_{\mathrm{L},i} = \sum_j g_j Aji/g_i$ can be computed sufficiently quickly, or
2. In continuous space $\hat{A}$ has only local and local-derivative terms, e.g., $\frac{-1}{2}\sum_i \nabla_i^2 + V(\mathbf{R})$.

Since $\hat{A}$ commutes with with $\hat{H}$ the mixed estimator equals the pure estimator

$$\langle A \rangle_{\mathrm{mix}} = \frac{\langle \Psi_{\mathrm{T}} | \hat{A} | \Psi_0 \rangle}{\langle \Psi_{\mathrm{T}} | \Psi_0 \rangle} = \frac{\langle \Psi_0 | \hat{A} | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle} = \langle A \rangle_{\mathrm{pure}}$$

Next write out explicitly to set the stage for more complicated cases.

1) $\hat{A}$ commutes with with $\hat{H}$ and is near-diagonal in chosen basis

$$
\begin{aligned}
\langle A \rangle &= \frac{\langle \Psi_{\rm T} | \hat{A} | \Psi_0 \rangle}{\langle \Psi_{\rm T} | \Psi_0 \rangle} = \frac{\langle \Psi_{\rm T} | \hat{A} G^p(\tau) | \Psi_{\rm T} \rangle}{\langle \Psi_{\rm T} | G^p(\tau) | \Psi_{\rm T} \rangle} \\
&= \frac{\sum_{\mathbf{R}_p \cdots \mathbf{R}_0} \; A \Psi_{\rm T}(\mathbf{R}_p) \left(\prod_{i=0}^{p-1} G(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) \Psi_{\rm T}(\mathbf{R}_0)}{\sum_{\mathbf{R}_p \cdots \mathbf{R}_0} \; \Psi_{\rm T}(\mathbf{R}_p) \left(\prod_{i=0}^{p-1} G(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) \Psi_{\rm T}(\mathbf{R}_0)} \\
&= \frac{\sum_{\mathbf{R}_p \cdots \mathbf{R}_0} \; \frac{A \Psi_{\rm T}(\mathbf{R}_p)}{\Psi_{\rm T}(\mathbf{R}_p)} \left(\prod_{i=0}^{p-1} \tilde{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) (\Psi_{\rm T}(\mathbf{R}_0))^2}{\sum_{\mathbf{R}_p \cdots \mathbf{R}_0} \; \left(\prod_{i=0}^{p-1} \tilde{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) (\Psi_{\rm T}(\mathbf{R}_0))^2} \\
&= \frac{\sum_{t=T_{\rm eq}+1}^{T_{\rm eq}+T} A_L(\mathbf{R}_t) W_t}{\sum_{t=T_{\rm eq}+1}^{T_{\rm eq}+T} W_t} \quad \text{since MC pts. from } \left(\prod_{i=0}^{p-1} \tilde{T}(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) (\Psi_{\rm T}(\mathbf{R}_0))^2
\end{aligned}
$$

$W_t = \prod_{i=0}^{p-1} w(\mathbf{R}_{t-i}, \mathbf{R}_{t-i-1})$ or better $W_t = \prod_{i=0}^{T_{\rm eq}+t-1} w(\mathbf{R}_{T_{\rm eq}+t-i}, \mathbf{R}_{T_{\rm eq}+t-i-1})$.

Branching (described later) is used to prevent inefficiency due wide disparity in weight products.

2) Expectation values of diagonal operators that do not commute with $\hat{H}$

DMC straightforwardly gives us

$$\langle A\rangle_{\rm mix} = \frac{\langle\Psi_0|\hat{A}|\Psi_{\rm T}\rangle}{\langle\Psi_0|\Psi_{\rm T}\rangle} = \frac{\int d\mathbf{R}\ \langle\Psi_0|\mathbf{R}\rangle\ \langle\mathbf{R}|\hat{A}|\mathbf{R}\rangle\ \langle\mathbf{R}|\Psi_{\rm T}\rangle}{\int d\mathbf{R}\ \langle\Psi_0|\mathbf{R}\rangle\ \langle\mathbf{R}|\Psi_{\rm T}\rangle} = \frac{\int d\mathbf{R}\ \Psi_0(\mathbf{R})\ A(R)\ \Psi_{\rm T}(\mathbf{R})}{\int d\mathbf{R}\ \Psi_0(\mathbf{R})\ \Psi_{\rm T}(\mathbf{R})}$$

but we want

$$\langle A\rangle_{\rm pure} = \frac{\langle\Psi_0|\hat{A}|\Psi_0\rangle}{\langle\Psi_0|\Psi_0\rangle} = \frac{\int d\mathbf{R}\ \langle\Psi_0|\mathbf{R}\rangle\ \langle\mathbf{R}|\hat{A}|\mathbf{R}\rangle\ \langle\mathbf{R}|\Psi_0\rangle}{\int d\mathbf{R}\ \langle\Psi_0|\mathbf{R}\rangle\ \langle\mathbf{R}|\Psi_0\rangle} = \frac{\int d\mathbf{R}\ \Psi_0(\mathbf{R})\ A(R)\ \Psi_0(\mathbf{R})}{\int d\mathbf{R}\ \Psi_0(\mathbf{R})\ \Psi_0(\mathbf{R})}$$

Two possibilities: Extrapolated estimator and forward walking

<u>1) Extrapolated estimator</u>

$$\begin{aligned}
\langle A\rangle_{\rm DMC} &= \langle A\rangle_{\rm pure} + \mathcal{O}(||\Psi_{\rm T}-\Psi_0||) \\
\langle A\rangle_{\rm VMC} &= \langle A\rangle_{\rm pure} + \mathcal{O}(||\Psi_{\rm T}-\Psi_0||) \\
2\langle A\rangle_{\rm DMC} - \langle A\rangle_{\rm VMC} &= \langle A\rangle_{\rm pure} + \mathcal{O}(||\Psi_{\rm T}-\Psi_0||)^2
\end{aligned}$$

2) Expectation values of diagonal operators that do not commute with $\hat{H}$

Forward or Future Walking

$$
\begin{aligned}
\langle A\rangle &= \frac{\langle\Psi_{\rm T}|G^p(\tau)\hat{A}G^{p'}(\tau)|\Psi_{\rm T}\rangle}{\langle\Psi_{\rm T}|G^{p+p'}(\tau)|\Psi_{\rm T}\rangle} \quad \text{need projectors on both sides of } \hat{A}\\
&= \frac{\sum_{\mathbf{R}_{p+p'}\cdots\mathbf{R}_0} A(\mathbf{R}_{p'})\left(\prod_{i=0}^{p+p'-1}\tilde{G}(\mathbf{R}_{i+1},\mathbf{R}_i)\right)(\Psi_{\rm T}(\mathbf{R}_0))^2}{\sum_{\mathbf{R}_{p+p'}\cdots\mathbf{R}_0}\left(\prod_{i=0}^{p+p'-1}\tilde{G}(\mathbf{R}_{i+1},\mathbf{R}_i)\right)(\Psi_{\rm T}(\mathbf{R}_0))^2}\\
&= \frac{\sum_{t=T_{\rm eq}+1}^{T_{\rm eq}+T} A(\mathbf{R}_t)W_{t+p}}{\sum_{t=T_{\rm eq}+1}^{T_{\rm eq}+T} W_{t+p}}
\end{aligned}
$$

$W_{t+p}=\prod_{i=0}^{p+p'-1} w(\mathbf{R}_{t+p-i},\mathbf{R}_{t+p-i-1})$ (product over p' past and p future) or better $W_{t+p}=\prod_{i=0}^{T_{\rm eq}+t+p-1} w(\mathbf{R}_{T_{\rm eq}+t+p-i},\mathbf{R}_{T_{\rm eq}+t+p-i-1})$, (product over entire past and p future generations).

The contribution to the expectation value is: the local operator at time t, multiplied by the weight at a future time $t+p$. Need to store $A(\mathbf{R}_t)$ for p generations.

Usual tradeoff: If p is small, there is some residual bias since $\Psi_{\rm T}$ has not been fully projected onto Ψ_0, whereas, if p is large the fluctuations of the descendent weights increases the statistical noise. (Since we use branching, weight factors from past are not a problem.) For very large p all walkers will be descended from the same ancestor. (Mitochondrial Eve!)

$\hat{A}$ is not diagonal in chosen basis, but, $A_{ij} \neq 0$ only when $G_{ij} \neq 0$

Forward or Future Walking

$$
\begin{aligned}
\langle A \rangle &= \frac{\langle \Psi_{\rm T} | G^{p-1}(\tau) \hat{A} G^{p'}(\tau) | \Psi_{\rm T} \rangle}{\langle \Psi_{\rm T} | G^{p+p'}(\tau) | \Psi_{\rm T} \rangle} \\
&= \frac{\sum_{\mathbf{R}_{p+p'} \cdots \mathbf{R}_0} \left(\prod_{i=p'+1}^{p+p'-1} \tilde{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) \tilde{A}(\mathbf{R}_{p'+1}, \mathbf{R}_{p'}) \left(\prod_{i=0}^{p'-1} \tilde{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) (\Psi_{\rm T}(\mathbf{R}_0))^2}{\sum_{\mathbf{R}_{p+p'} \cdots \mathbf{R}_0} \left(\prod_{i=0}^{p+p'-1} \tilde{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) (\Psi_{\rm T}(\mathbf{R}_0))^2} \\
&= \frac{\sum_{t=T_{\rm eq}+1}^{T_{\rm eq}+T} W_{t+p-1,t+1} \; a(\mathbf{R}_{t+1}, \mathbf{R}_t) \; W_{t,t-p'}}{\sum_{t=T_{\rm eq}+1}^{T_{\rm eq}+T} W_{t+p}}
\end{aligned}
$$

$$
a(\mathbf{R}_{t+1}, \mathbf{R}_t) = \frac{\tilde{A}(\mathbf{R}_{t+1}, \mathbf{R}_t)}{\tilde{T}(\mathbf{R}_{t+1}, \mathbf{R}_t)} = \frac{A(\mathbf{R}_{t+1}, \mathbf{R}_t)}{T(\mathbf{R}_{t+1}, \mathbf{R}_t)}
$$

$$
W_{t_2,t_1} = \prod_{i=t_1}^{t_2-1} w(\mathbf{R}_{i+1}, \mathbf{R}_i)
$$

Again, the product of p' past weights can be replaced by products of weights over entire past.

$\hat{A}$ is not diagonal in chosen basis, and, $\exists$ some $A_{ij} \neq 0$ where $G_{ij} = 0$

Side Walking

Now it becomes necessary to have side walks that start from the backbone walk.

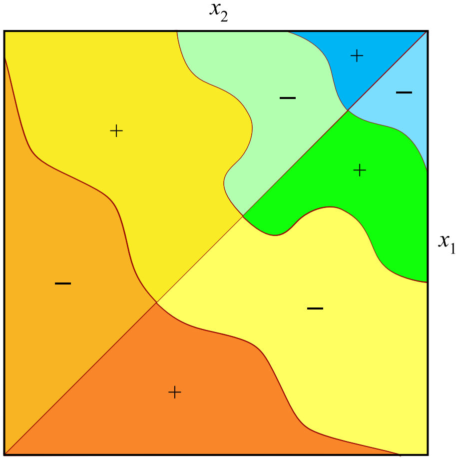
Just as we did for the importance-sampled projector, we factor $\tilde{A}$ into a Markov matrix and a reweighting factor.

The first transition of the side walk is made using this Markov matrix and and the rest of the side-walk using the usual Markov matrix.

The ends of the side-walks contribute to the expectation values.

This method is even more computationally expensive than forward walking, because one has to do an entire side walk long enough to project onto the ground state to get a single contribution to the expectation value.

Excited states



Pure State versus Finite Temperature MC methods

So far we have talked about pure-state MC methods. We now give a very brief introduction to a finite-temperature MC method, the path-integral Monte Carlo (PIMC) method. Sometimes PIMC is used to approximate ground states, but that gets expensive since the length of the polymer grows as the inverse temperature. We also very briefly discuss the essence of reptation MC, which is in some sense a hybrid between PMC and PIMC. First we give schematics of the walks in VMC, PMC, PIMC and reptation MC.

Schematic of VMC, PMC, PIMC and Reptation MC

1. Metropolis MC requires just a single unweighted walker, since the sum of the elements of a column of the Markov matrix add up to 1. PMC requires (for reasons of efficiency) a population of weighted walkers.
   (One could view the Markov matrix in Metropolis MC as a projector. Instead of projecting onto the ground state as in PMC, it instead projects onto the known guiding/trial state.)
2. In VMC and PMC the projection time and the MC time are the same. In PIMC and reptation MC, they are different – the projection time or inverse temperature is finite, whereas the MC time can of course be as large as patience permits.
3. The object being evolved has an extra time dimension in PIMC and reptation MC, compared to VMC and PMC, and so this adds to the computational cost. On the other hand one can compensate for this because the probability density is known, and so one can use Metropolis-Hastings and one has great freedom in the choice of the moves. (Algorithm is correct so long as detailed balance is satisfied.)
4. VMC, PMC and reptation MC are ground state methods, PIMC a finite-temperature method. People do use PIMC to approach the ground state but it becomes very inefficient since the length of the paths become very long.
5. VMC, PMC and reptation MC take advantage of accurate guiding/trial wavefunctions to greatly enhance their efficiency. PIMC does not.

Path-integral Monte Carlo, Finite Temperature

Detailed and excellent review: Ceperley, Rev. Mod. Phys. (1995)

In quantum statistical mechanics the expectation value of an operator A is given by

$$\langle \hat{A} \rangle = \frac{\mathrm{tr}(\rho \hat{\mathrm{A}})}{\mathrm{tr}(\rho)}$$

where $\hat{\rho} = e^{-\beta \hat{H}}$ is the quantum density matrix operator, $Z = \mathrm{tr}(\hat{\rho})$ is the partition function, $\hat{H}$ is the Hamiltonian and $\beta = 1/k_B T$.
The density operator in quantum statistical mechanics is identical to the quantum mechanical time evolution operator if we make the identification $t = -i\hbar\beta$.
If trace is evaluated in energy representation then energy is

$$E(\beta) = \frac{\mathrm{tr}(\rho \hat{\mathrm{H}})}{\mathrm{tr}(\rho)} = \frac{\sum_{i=0}^{\infty} \langle \Psi_i | e^{-\beta \hat{H}} \hat{H} | \Psi_i \rangle}{\sum_{i=0}^{\infty} \langle \Psi_i | e^{-\beta \hat{H}} | \Psi_i \rangle} = \frac{\sum_{i=0}^{\infty} E_i e^{-\beta E_i}}{\sum_{i=0}^{\infty} e^{-\beta E_i}} \underset{\beta \to \infty}{\longrightarrow} E_0$$

where E_i are the eigenvalues of $\hat{H}$. Not useful since we do not know the energy eigenvalues or eigenvectors. On the other hand we do have approximate representations of $e^{-\beta \hat{H}}$ in coordinate representation.

Path-integral Monte Carlo, Finite Temperature

The expectation value of $\hat{A}$ in coordinate representation is

$$\langle A \rangle = \frac{\int d\mathbf{R}\ \langle \mathbf{R}|Ae^{-\beta H}|\mathbf{R}\rangle}{\int d\mathbf{R}\ \langle \mathbf{R}|e^{-\beta H}|\mathbf{R}\rangle}$$

Introducing a complete set of position eigenstates, we obtain,

$$\langle A \rangle = \frac{\int d\mathbf{R}\ d\mathbf{R}'\ \langle \mathbf{R}|A|\mathbf{R}'\rangle\ \langle \mathbf{R}'|e^{-\beta H}|\mathbf{R}\rangle}{\int d\mathbf{R}\ \langle \mathbf{R}|e^{-\beta H}|\mathbf{R}\rangle}$$

If A is diagonal in the coordinate representation then this reduces to

$$\langle A \rangle = \frac{\int d\mathbf{R}\ A(\mathbf{R})\ \rho(R,R,\beta)}{\int d\mathbf{R}\ \rho(R,R,\beta)}$$

where $\rho(R',R,\beta) \equiv \langle \mathbf{R}'|e^{-\beta H}|\mathbf{R}\rangle$ is the density matrix in coordinate representation. This is the same as in diffusion MC, but now imaginary time is interpreted as inverse temperature.

Eq. 7 can be evaluated by the Metropolis method since we derived an explicit, though approximate, expression for $\rho(R',R,\beta)$ when discussing diffusion MC.

$$\rho(R',R,\beta) \propto$$

$$\int d\mathbf{R}_1 \ldots d\mathbf{R}_{M-1} \quad e^{-\tau\left[\left\{\left(\frac{\mathbf{R}'-\mathbf{R}_{M-1}}{2\tau}\right)^2 + \left(\frac{\mathbf{R}_{M-1}-\mathbf{R}_{M-2}}{2\tau}\right)^2 + \ldots + \left(\frac{\mathbf{R}_1-\mathbf{R}}{2\tau}\right)^2\right\} \ + \ \left(\frac{V(\mathbf{R}')}{2} + V(\mathbf{R}_{M-1}) + \ldots + V(\mathbf{R}_1) + \frac{V(\mathbf{R})}{2}\right)\right]}$$

Path-integral Ground State (PIGS) / Reptation Monte Carlo

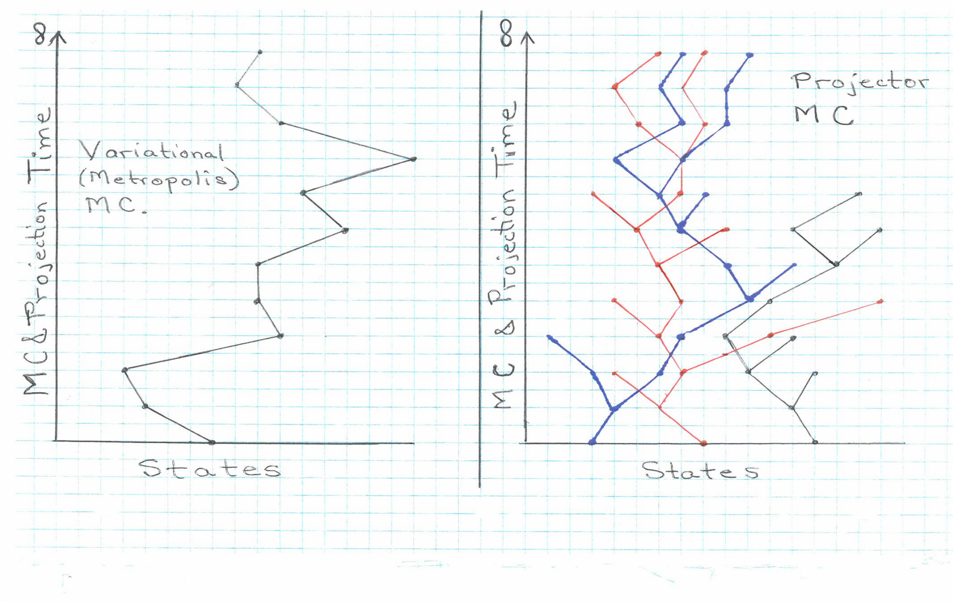
A hybrid between DMC and PIMC

Baroni and Moroni, in NATO book, ed. by M.P, Nightingale and CJU, (1999)

$$\langle A \rangle = \frac{\int d\mathbf{R}\; d\mathbf{R}'\; d\mathbf{R}''\, \langle \Psi_{\rm T}|\mathbf{R}\rangle\; \langle \mathbf{R}|e^{-t\hat{H}}|\mathbf{R}'\rangle\; \langle \mathbf{R}'|A|\mathbf{R}'\rangle\; \langle \mathbf{R}'|e^{-t\hat{H}}|\mathbf{R}''\rangle\; \langle \mathbf{R}''|\Psi_{\rm T}\rangle}{\int d\mathbf{R}\; d\mathbf{R}'\; d\mathbf{R}''\, \langle \Psi_{\rm T}|\mathbf{R}\rangle\; \langle \mathbf{R}|e^{-t\hat{H}}|\mathbf{R}'\rangle\; \langle \mathbf{R}'|e^{-t\hat{H}}|\mathbf{R}''\rangle\; \langle \mathbf{R}''|\Psi_{\rm T}\rangle}$$

Compared to PIMC – Instead of having a closed polymer, which needs to be very long in order to get the ground state, have an open polymer terminated by $\Psi_{\rm T}$. As $\Psi_{\rm T}$ gets better, the length of open polymer can get shorter.
Compared to DMC – Tradeoff between having the complication of moving an entire polymer versus the freedom of using clever Metropolis moves. Less efficient for the energy, but, more efficient for operators that do not commute with $\hat{H}$ if extrapolated estimators are not accurate enough.

Schematic of VMC and PMC

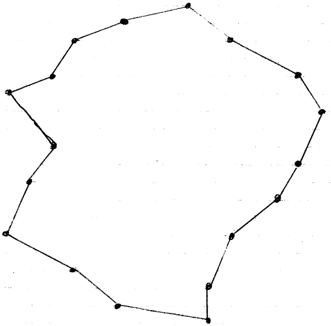


Schematic of PIMC and Reptation MC

Path-Integral M.C.

Average path length ∝ projection time/inverse T

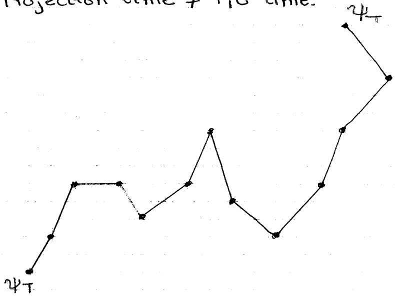
Projection time ≠ M.C. time



Reptation M.C.

Average path length ∝ projection time

Projection time ≠ MC time.



Some topics in classical MC and QMC we did not discuss

1. Cluster algorithms – Swendsen-Wang and Wolff
2. Multiple-try Metropolis
3. Multilevel sampling
4. Correlated sampling, umbrella sampling, Wang-Landau
5. Lattice-regularized DMC (Sorella, Casula)
6. Nonlocal pseudopotentials in QMC (Fahy; Mitas, Shirley, Ceperley; Casula)
7. Extended systems (periodic BC, wavefns, finite-size errors, ...) (Foulkes, Needs, ...)
8. Penalty method and coupled electron-ion MC (Ceperley, Dewing, Pierleoni)
9. Zero-variance zero bias (ZVZB) method (Assaraf and Caffarel)
10. Extension of fixed-node to fixed-phase method (Bolton; Ortiz, Jones, Ceperley)
11. Domain Green's function MC. (Kalos)
12. Stochastic series expansion (Sandvik)
13. Loop and worm algorithms
14. Diagrammatic Monte Carlo
15. Impurity solvers for DMRG