# Zero-temperature Quantum Monte Carlo Methods in Chemistry and Physics

Cyrus Umrigar

Physics Department, Cornell University, Ithaca.

Email: CyrusUmrigar@cornell.edu

(M,W,F) 1:30-3:00, Apr 27 - May 8, 2020

# 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: $\binom{200}{10}^2 = 5.0 \times 10^{32}$

Half-filled 2D Hubbard model on $16 \times 16$ lattice: $\binom{256}{128}^2 = 3.3 \times 10^{151}$

(Partial) Solutions:

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

# 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.

# 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 most of the above mean (except the last line).

# Definitions

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

$$\text{Exact} \quad |\Psi_0\rangle = \sum_i e_i|\phi_i\rangle, \quad \text{where,} \quad e_i = \langle\phi_i|\Psi_0\rangle$$

$$\text{Trial} \quad |\Psi_T\rangle = \sum_i t_i|\phi_i\rangle, \quad \text{where,} \quad t_i = \langle\phi_i|\Psi_T\rangle$$

$$\text{Guiding} \quad |\Psi_G\rangle = \sum_i g_i|\phi_i\rangle, \quad \text{where,} \quad g_i = \langle\phi_i|\Psi_G\rangle$$

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

$\Psi_{\rm T}$ and $\Psi_{\rm G}$ are often chosen to be the same, but they serve different purposes.

$\Psi_T$: used to calculate variational and mixed estimators of operators $\hat{A}$, i.e., $\langle\Psi_{\rm T}|\hat{A}|\Psi_{\rm T}\rangle/\langle\Psi_{\rm T}|\Psi_{\rm T}\rangle$, $\langle\Psi_{\rm T}|\hat{A}|\Psi_0\rangle/\langle\Psi_{\rm T}|\Psi_0\rangle$. Need rapid evaluation of "local energy", $E_{\rm L}(i) = \sum_j H_{ij}t_j/t_i$.

$\Psi_G$: used to alter sampled probability density: $\Psi_{\rm G}^2$ in VMC, $\Psi_{\rm G}\Psi_0$ in PMC. So, must satisfy $g_i \neq 0$ if $e_i \neq 0$. Also, chosen to have finite variance estimators.

To simplify expressions, we sometimes use $\Psi_{\rm G} = \Psi_{\rm T}$ or $\Psi_{\rm G} = 1$.

# Variational MC

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

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

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

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

Value depends only on $\Psi_{\rm T}$. Statistical error depend on $\Psi_{\rm T}$ and $\Psi_{\rm G}$.

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

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

In fact $\Psi_G \neq \Psi_T$ needed when optimizing wavefunctions 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:* $E_0 = \dfrac{\langle \Psi_0 | \hat{H} | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle} = \dfrac{\langle \Psi_0 | \hat{H} | \Psi_T \rangle}{\langle \Psi_0 | \Psi_T \rangle}$

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

$$E_0 = \frac{\langle \Psi_0 | \hat{H} | \Psi_T \rangle}{\langle \Psi_0 | \Psi_T \rangle} = \frac{\sum_{ij}^{N_{st}} \langle \Psi_0 | \phi_i \rangle \langle \phi_i | \hat{H} | \phi_j \rangle \langle \phi_j | \Psi_T \rangle}{\sum_{k}^{N_{st}} \langle \Psi_0 | \phi_k \rangle \langle \phi_k | \Psi_T \rangle}$$

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

$$= \sum_{i}^{N_{st}} \frac{e_i t_i}{\sum_{k}^{N_{st}} e_k t_k} E_L(i) = \frac{\left[\sum_{i}^{N_{MC}} E_L(i)\right]_{\Psi_T \Psi_0}}{N_{MC}} \xrightarrow{\Psi_G \neq \Psi_T} \frac{\left[\sum_{i}^{N_{MC}} \left(\frac{t_i}{g_i}\right) E_L(i)\right]_{\Psi_G \Psi_0}}{\left[\sum_{k}^{N_{MC}} \left(\frac{t_k}{g_k}\right)\right]_{\Psi_G \Psi_0}}$$

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

Statistical error vanishes as $\Psi_T \to \Psi_0$.

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

e.g. FCIQMC is a PMC method where $\Psi_G = \mathbf{1} \neq \Psi_T$.

# 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., discrete space: $\hat{H}$ is sparse, continuous space: $V(\mathbf{R})$ is local since K.E. requires only local derivs.

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

# Variational Monte Carlo in Real Space

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

Real space $\Longrightarrow$ $|\phi_i\rangle = |\mathbf{R}\rangle$. Monte Carlo is used to perform the many-dimensional integrals needed to calculate quantum mechanical expectation values. e.g.

$$
\begin{aligned}
E_T &= \frac{\int d\mathbf{R}\, \Psi_{\rm T}^*(\mathbf{R})\, \mathcal{H}\, \psi_{\rm T}(\mathbf{R})}{\int d\mathbf{R}\, \psi_{\rm T}^2(\mathbf{R})} \\
&= \int d\mathbf{R}\, \frac{\psi_{\rm T}^2(\mathbf{R})}{\int d\mathbf{R}\, \psi_{\rm T}^2(\mathbf{R})}\, \frac{\mathcal{H}\psi_{\rm T}(\mathbf{R})}{\psi_{\rm T}(\mathbf{R})} \\
&= \frac{1}{N}\sum_i \frac{\mathcal{H}\Psi_{\rm T}(\mathbf{R}_i)}{\Psi_{\rm T}(\mathbf{R}_i)} = \frac{1}{N}\sum_i E_L(\mathbf{R}_i)
\end{aligned}
$$

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

# Rest of this lecture

Now that you know the essence of quantum Monte Carlo methods, for the rest of this lecture we will discuss basic concepts that underlie both classical and quantum Monte Carlo methods, e.g., the central limit theorem, techniques for sampling various distributions, importance sampling for reducing statistical error, calculation of unbiased estimators, ...

Then in the rest of the lectures we will continue our study of quantum Monte Carlo methods.

# 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.

# Why should one be interested in QMC methods?

1. Exact solutions are rarely possible. Even good approximate solutions are only possible for a limited set of problems. These solutions are often obtained by making approximations that reduce a high-dimensional problem to a problem in a smaller number of dimensions.
2. Quantum Monte Carlo methods on the other hand are widely applicable to a wide variety of lattice and continuum systems because the many-dimensional character of the problem is not a big impediment and can be handled without making severe approximations.
3. Simplest application of MC methods is to integration and in fact this is a component of more sophisticated applications also.
4. Becoming increasingly popular with greatly improved algorithms and the advent of massively parallel computers.

# But not a panacea

- Statistical errors: Sometimes small, sometimes prohibitive.
  Quite often a straightforward application of QMC will give large statistical errors but some thought and a minor change in the algorithm can reduce the error dramatically. Knowledge of an approximate solution can reduce statistical errors (*importance sampling*).
- Systematic errors: Often but not always acceptably small.
- Trade offs: Frequently in QMC there is a trade off between systematic and statistical errors.
  Often a happy compromize can be found, *e.g.* in dealing with the *population control error*.
  On the other hand, in dealing with the infamous *Fermion sign problem* the increase in the statistical error from attempts to design algorithms with negligible systematic errors is sufficiently large that the practical route to accurate energies is live with algorithms that have systematic *fixed node* errors but to make the errors small by optimizing trial wavefunctions (and their nodal surfaces).

# Physics/Chemistry applications of Quantum Monte Carlo

Some systems to which they have been applied are:

- strongly correlated systems (Hubbard, Anderson, t-J, ... models)
- quantum spin systems (Ising, Heisenberg, xy, ... models),
- liquid and solid helium, liquid-solid interface, droplets
- energy and response of homogeneous electron gas in 2-D and 3-D
- nuclear structure
- lattice gauge theory
- atomic clusters
- electronic structure calculations of atoms, molecules, solids, quantum dots, quantum wires

- both to zero temperature (pure states) and finite temperature problems, but in these lectures we will mostly discuss zero temperature methods

# 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 few abide 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 $\pi$ (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.

# Comte 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

# Comte 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!

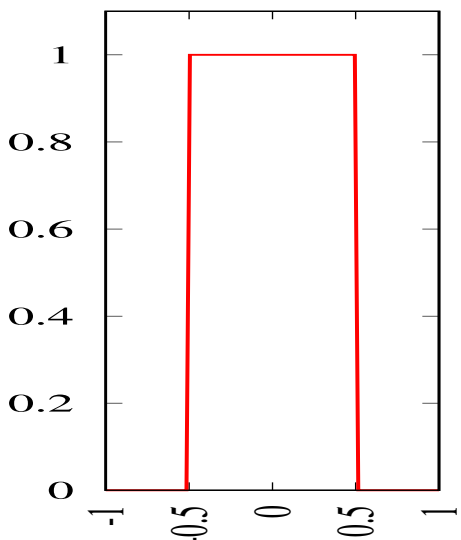
Here he is:

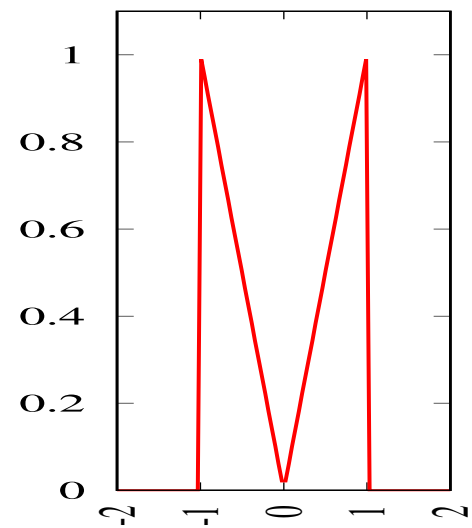
Among other things, he wrote a 36 volume set of books on the Natural History of the Earth!

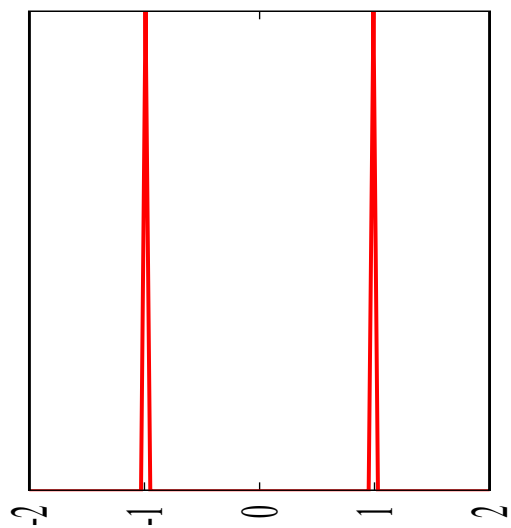
# Central Limit Theorem

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, $\mu$, and variance $\sigma^2$. 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 $\mu$, and variance $\sigma^2/N$, 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, $\mu$, but not necessarily a finite variance $\sigma^2$. 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(Cauchy)

A Lorentzian (also known as Cauchy) probability density function

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

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.

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

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 violates the conditions for the CLT!

# Lorentzian(Cauchy)

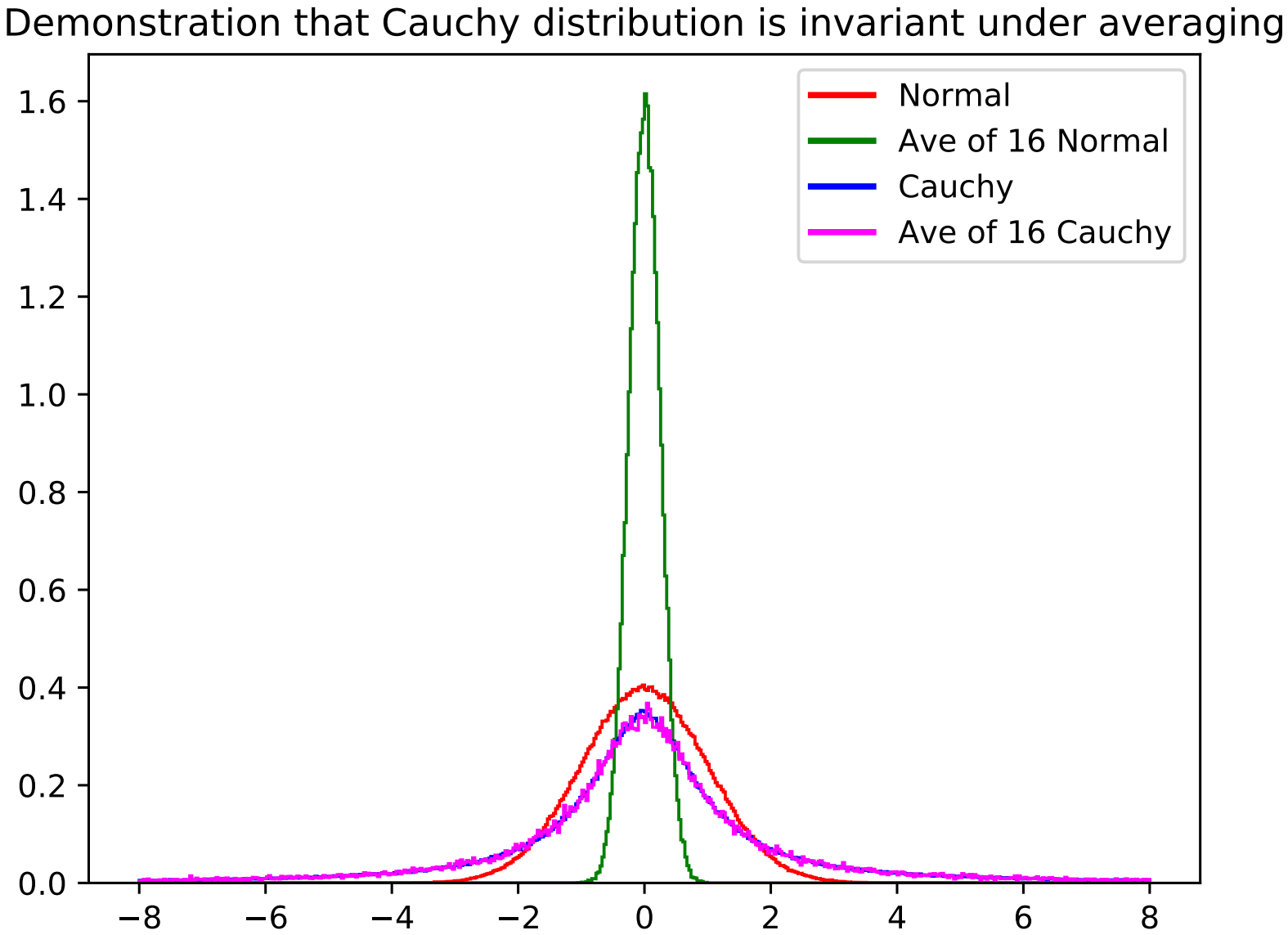
We are all brought up to believe that if we average numbers drawn from some probability density then the distribution of the averages will be narrower than the distribution of the individual values.

In fact this is not true in general!

Any amount of averaging of a Lorentzian gives the **same** Lorentzian!!

We can easily verify this. Using the transformation method (discussed a few viewgraphs later), we can sample from a Lorentzian with the prescription $x = \tan(\pi(\xi - 1/2))$.

# Lorentzian(Cauchy)



# Chebychev Inequality

The Central Limit Theorem tells you that if $\sigma^2$ is finite, the distribution of sample averages will converge to a gaussian, but it 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 $\mu$ and finite variance $\sigma^2$, 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 erf$(n/\sqrt{2})$ for a Gaussian.

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

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

# Infinite variance estimators

What if the population variance $\sigma^2 = \infty$ but we do not know that beforehand? The computed sample variance will of course always be finite. The practical signature of an infinite variance estimator is that the estimated $\sigma$ 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, $\epsilon$, 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 $d = 100$, even 2 grid points per dimensions gives $N_{\rm int} \approx 10^{30}$, 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

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

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

## Monte Carlo integration

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

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

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

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

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

## Importance sampling (put most of the fluctuations in sampled distribution)

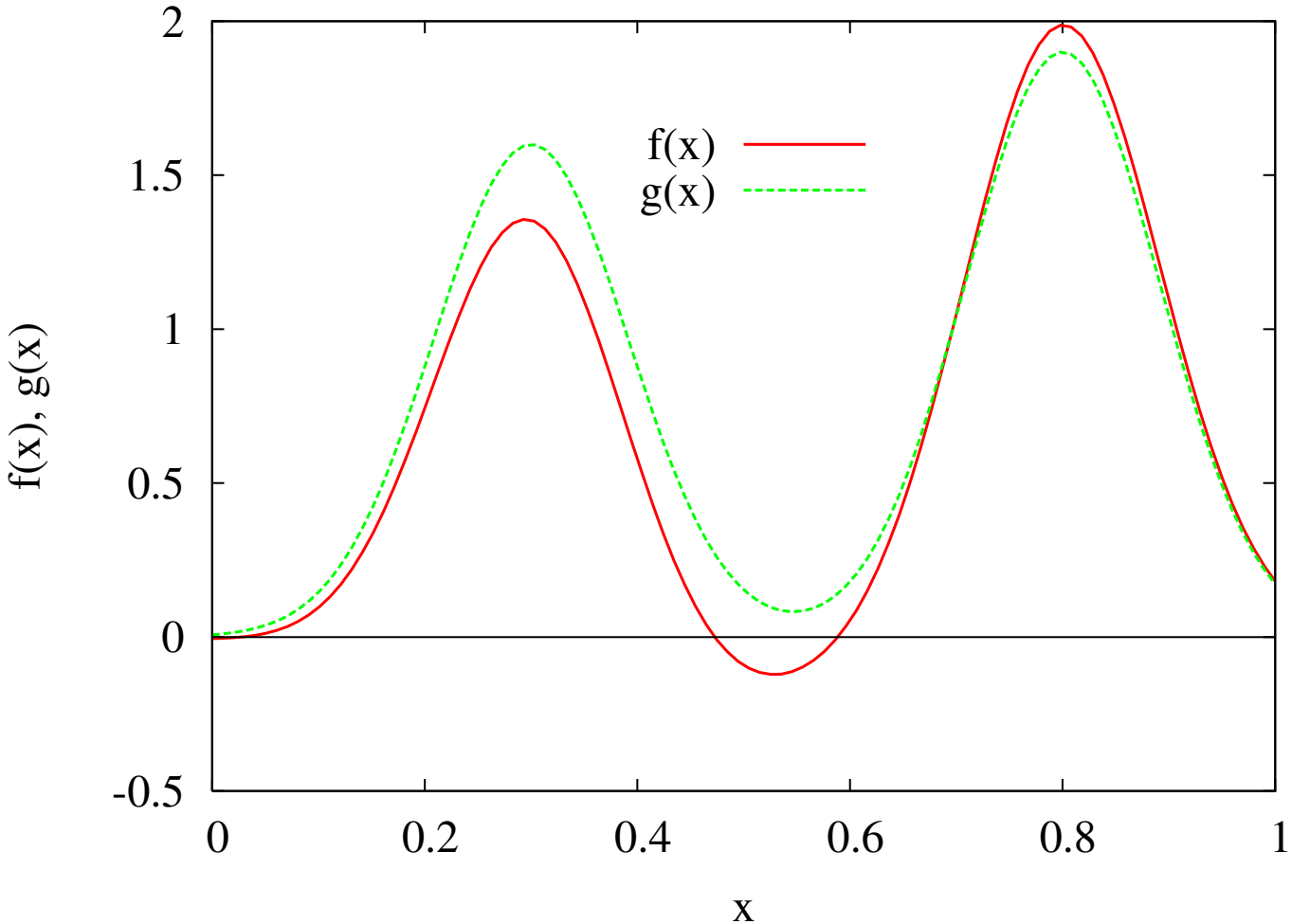
$$I = \int_V g(x)\frac{f(x)}{g(x)}dx = \overline{\left(\frac{f}{g}\right)} \pm \sqrt{\frac{\overline{\left(\frac{f}{g}\right)^2} - \overline{\left(\frac{f}{g}\right)}^2}{N-1}}$$

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!

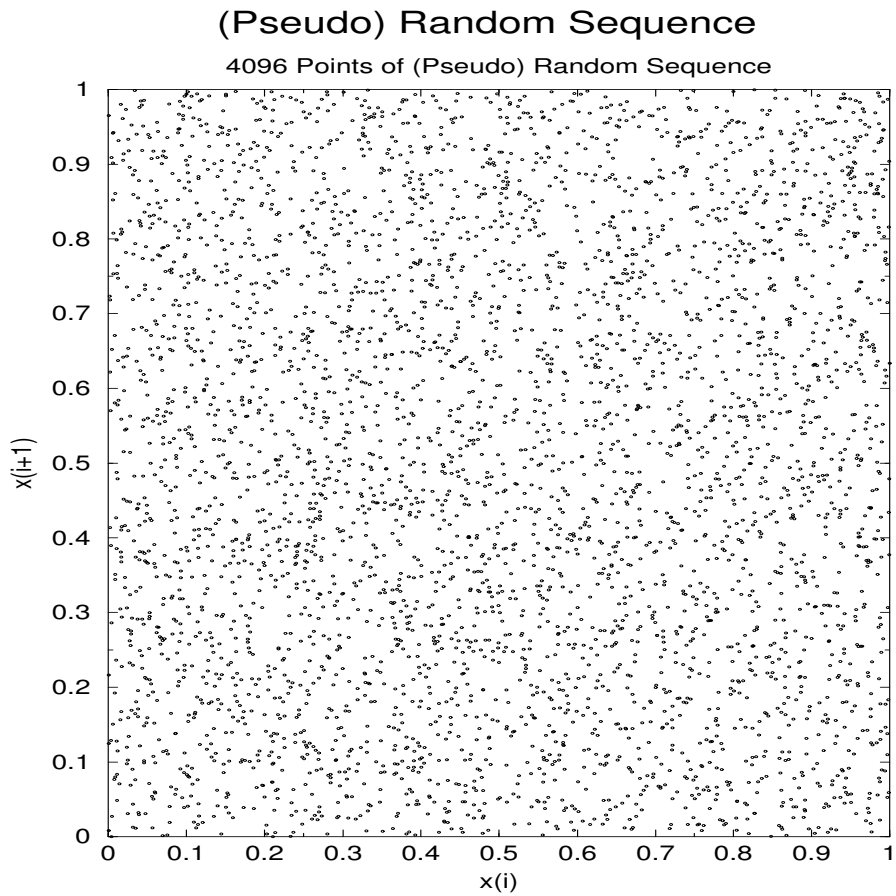
# Illustration of Importance Sampling

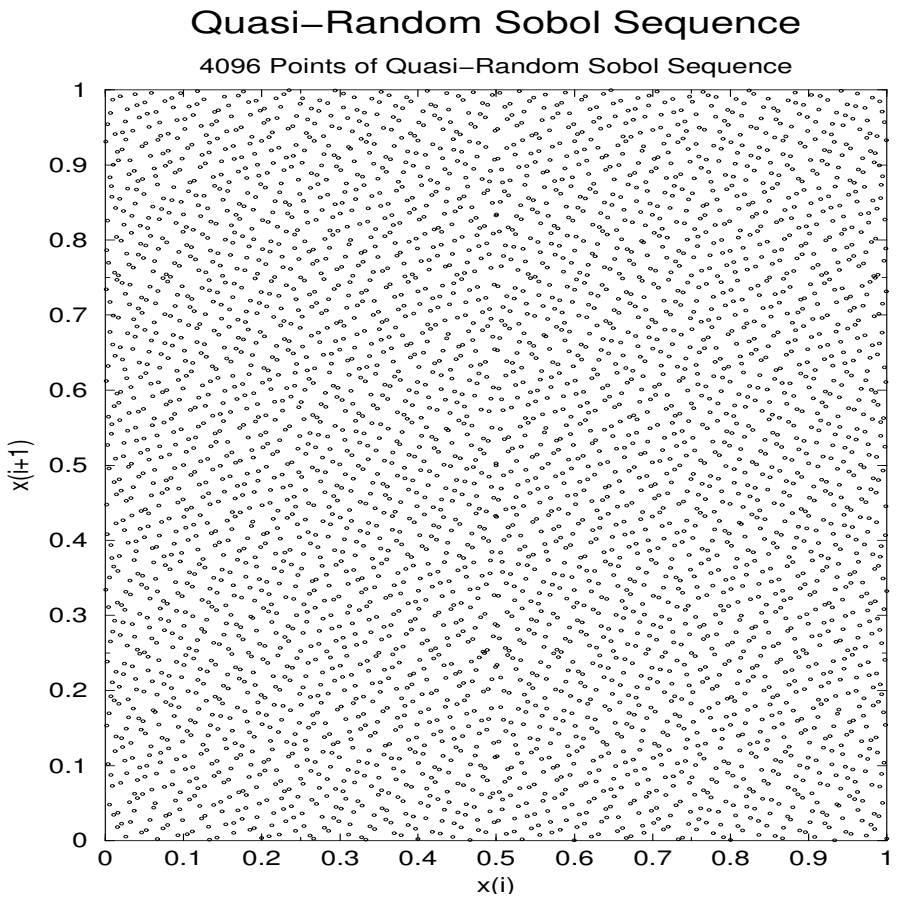
$f(x)$ is the function to be integrated. $g(x)$ is a function that is "similar" to $f(x)$ and has the required properties: a) $g(x) \geq 0$, b) $\int dx\, g(x) = 1$, and, c) we know how to sample it. $\int f(x)dx$ can be evaluated efficiently by sampling $g(x)$ and averaging $f(x)/g(x)$.



# 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.

# Expectation Values

Interested in calculating expection values (ensemble averages) of a variable $X$ with respect to a probability density $\rho$, (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 Monte Carlo time average.
Equality when $T \to \infty$.
We need a means to sample $\rho$.

# Sampling of arbitrary probability density functions

Infinite-variance estimators can be replaced by finite-variance estimators by sampling the MC points from an appropriate probability density functions.

Techniques for sampling arbitrary probability density functions employ standard random numbers generators that sample a uniform distribution in $[0, 1]$. We study 3 techniques for sampling nonuniform distributions:

1. transformation method
2. rejection method
3. Metropolis-Hastings method (may use transformation method for proposal probability)

but first we say a few words about random number generators.

# 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. If $N$ bits are used to represent the random numbers, then the number of different numbers generated can by no larger than $2^N$.
Also, the random numbers must repeat after a finite (though hopefully very large) period. 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 somewhat more complicated functions)
3) Metropolis-Hastings method (For any function)

1) Transformation method: Perform a transformation $x(\xi)$ on a uniform deviate $\xi$, 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 $\xi$ 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)

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

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

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

($x_1$ 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)

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

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

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

($x_1$ 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 uniform in $[-1/2, 1/2]$. Since $\sigma^2$ for $\xi - 0.5$ 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 $\pm 6$)

# Rejection Method

We wish to sample $f(x)$.
Find a function $g(x)$ that can be sampled by another method (say transformation) that preferably mimics the behaviour of $f(x)$, and for which we know that $C \geq \max(f(x)/g(x))$.
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 $\mathcal{O}(N)$, $\mathcal{O}(\log_2(N))$, $\mathcal{O}(1)$?

# 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 $\mathcal{O}(N)$, $\mathcal{O}(\log_2(N))$, $\mathcal{O}(1)$?

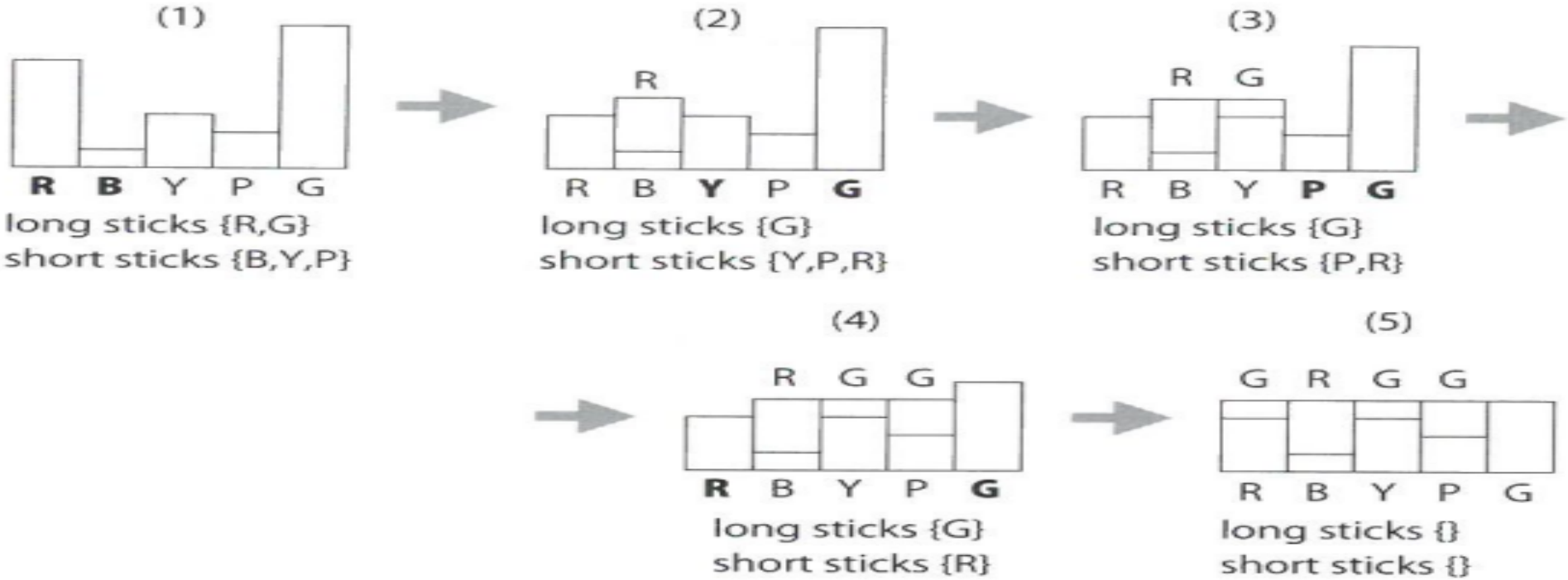
Straightforward $\mathcal{O}(\log_2(N))$ method with binary search:

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

So, one time $\mathcal{O}(N)$ cost to construct cumulative array, but then only $\mathcal{O}(\log_2(N))$ for each sample.

Can we draw each sample in $\mathcal{O}(1)$ time (again after $\mathcal{O}(N)$ one-time cost)?
Yes, using the Alias Method!

# 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 $A_i$.

Requires 2 random numbers, but no binary search!

Figure taken from book by Gubernatis, Kawashima and Werner

# 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.

# Computer lab on VMC and DMC (Tyler)

Get computer lab:

git clone https://github.com/CyrusUmrigar/QMC_module.git QMC_module

# Example of Importance Sampling to Calculate Integrals More Efficiently

Suppose we wish to compute

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

Note that

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

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

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

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

$$\int_0^1 dx\, f(x) = \int_0^1 dx\, g(x)\frac{f(x)}{g(x)} = \int_0^1 dx\, \frac{1-p}{x^p}\frac{1}{(1-p)(1+x^{1-p})}$$

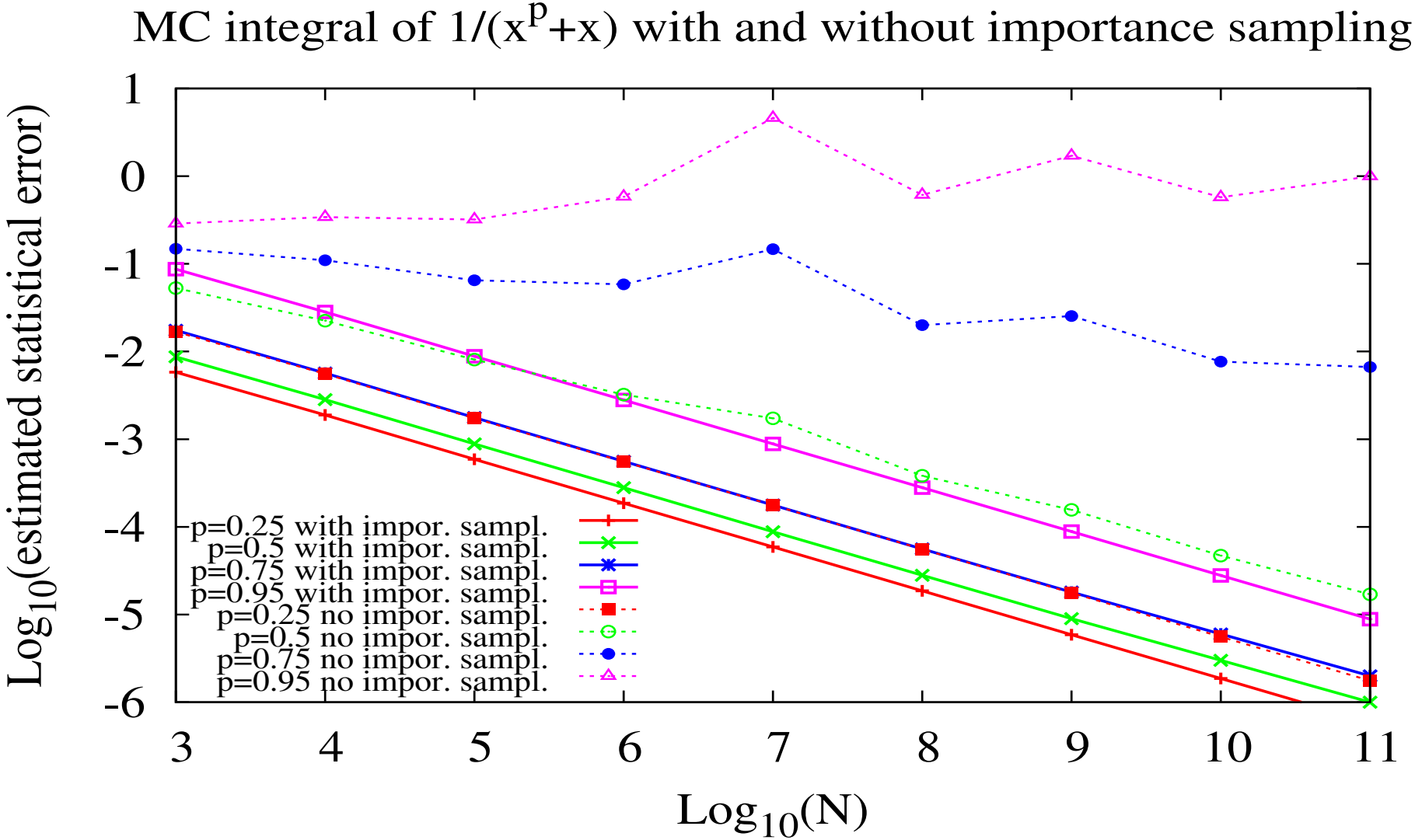
$$\approx \frac{1}{N_{\rm MC}(1-p)} \sum_{k=1}^{N_{\rm MC}} \frac{1}{(1+x_k^{1-p})} = \frac{1}{N_{\rm MC}(1-p)} \sum_{k=1}^{N_{\rm MC}} \frac{1}{(1+\xi_k)}$$

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 $\log(N_{\rm MC})$.

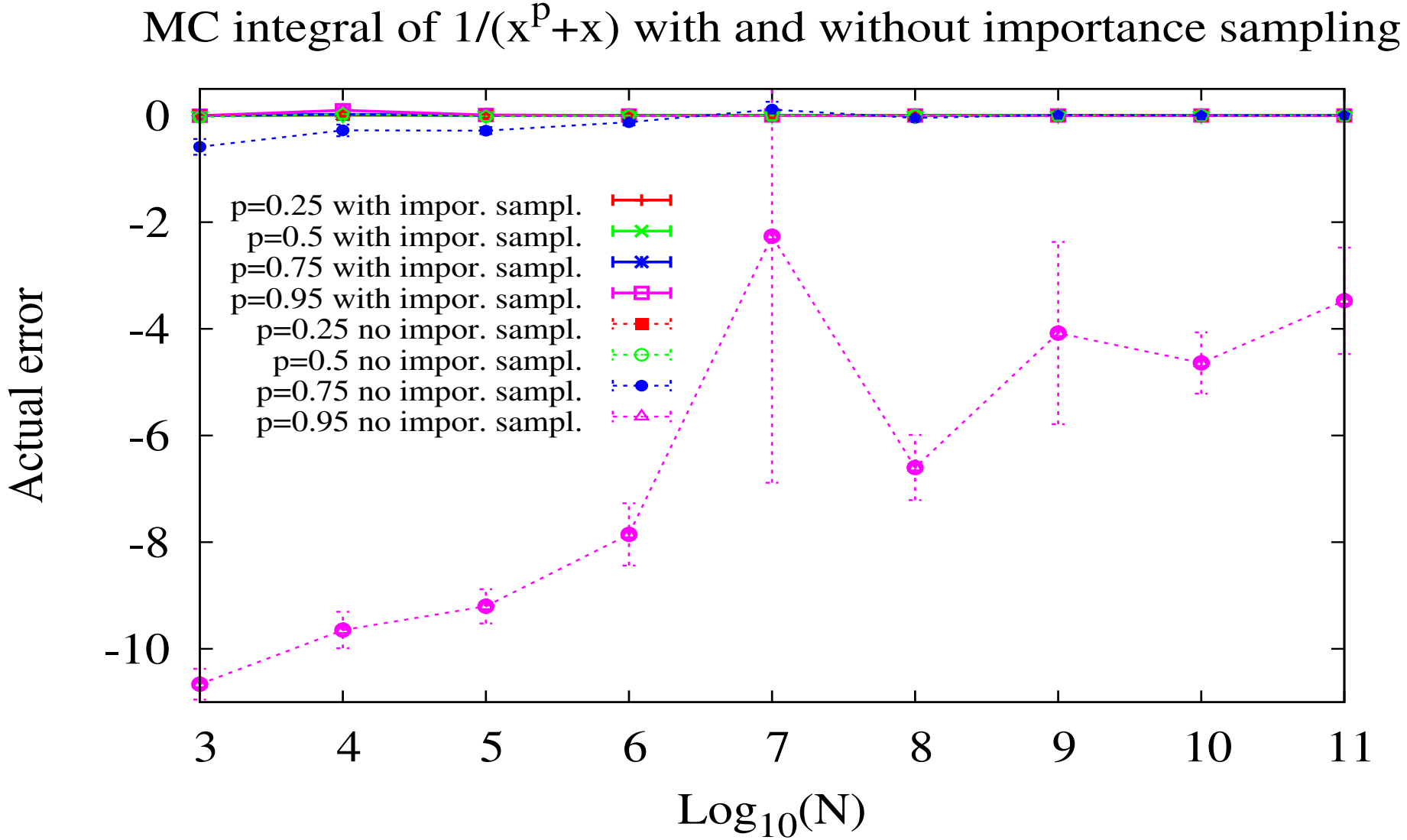
Cyrus J. Umrigar

# Homework Solution 1a



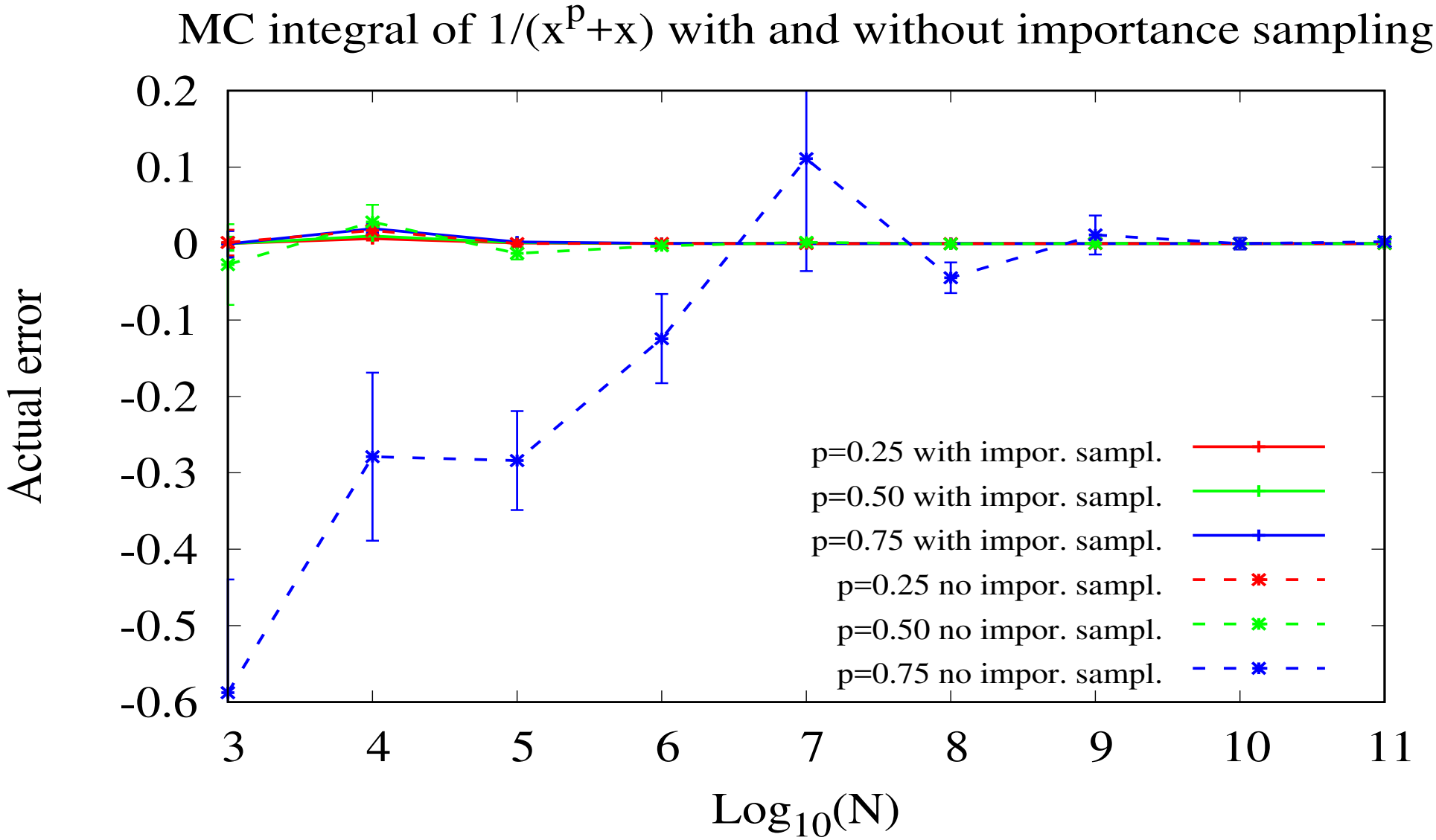

Statistical errors $\sim {N_{\rm MC}}^{-1/2}$ for all $p$ with importance sampling but only for $p = 0.25$ without importance sampling. For $p = 1$ even the integral is infinite. For $p = 0.95$ no sign of convergence. Theorem about asymptotic convergence of little practical utility.

# Homework Solution 1b




For $p = 0.95$ all of the errors are negative. Occasional large positive errors will bring mean to correct value. Actual errors may be MANY standard deviations. For infinite variance highly skewed variables the averages with the larger actual errors tend to have the smaller estimated errors! So, weighting estimates by inverse variances is bad!!

# Homework Solution 1b (expanded scale)




Beware of infinite-variance estimators, particularly if the distribution of estimates is highly skewed!

# Unbiased Estimators

Population mean: $\langle f \rangle$

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

In general, $\left\langle F(\bar{f}) \right\rangle \neq F(\langle f \rangle)$, so $F(\bar{f})$ is a biased estimator.

$\tilde{F}(\bar{f})$ 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: $\bar{f}$

In general, $\left\langle F(\bar{f}) \right\rangle \neq F(\langle f \rangle)$, so $F(\bar{f})$ is a biased estimator.

$\tilde{F}(\bar{f})$ 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 $\left\langle \bar{f}^2 - \bar{f}^2 \right\rangle = \langle f^2 \rangle - \langle f \rangle^2$ ? no. Correct: $\frac{N}{N-1} \left\langle \bar{f}^2 - \bar{f}^2 \right\rangle = \langle f^2 \rangle - \langle f \rangle^2$

# Examples of Unbiased and Biased Estimators

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

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

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

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

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, $\{\langle f_i \rangle\}$. (Here $f_i$ are different variables, not samples)
$F\left(\{\bar{f}_i\}\right)$ is unbiased estimator for $F(\{\langle f_i \rangle\})$ iff $F$ is linear function of $\{\langle f_i \rangle\}$.

In general use:

$$F(\{\langle f_i \rangle\}) = F\left(\{\bar{f}_i\}\right) - \frac{1}{2}\sum_{i,j} \left.\frac{\partial^2 F}{\partial f_i \partial f_j}\right|_{\bar{f}_i \bar{f}_j} \frac{\mathrm{cov}(f_i, f_j)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

$$\mathrm{var}\big(F(\{\langle f_i \rangle\})\big) = \sum_{i,j} \left.\frac{\partial F}{\partial f_i}\frac{\partial F}{\partial f_j}\right|_{\bar{f}_i \bar{f}_j} \mathrm{cov}(f_i, f_j) + \mathcal{O}\left(\frac{1}{N}\right)$$

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

Proof:

Taylor expand $F\left(\{\bar{f}_i\}\right)$ about $\{\langle f_i\rangle\}$ and take average,

$$\left\langle F\left(\{\bar{f}_i\}\right)\right\rangle = F(\{\langle f_i\rangle\}) + \sum_i \frac{\partial F}{\partial f_i}\underbrace{\left\langle \bar{f}_i - f_i\right\rangle}_{\to 0}$$

$$+\frac{1}{2}\sum_{i,j}\left.\frac{\partial^2 F}{\partial f_i \partial f_j}\right|_{\langle f_i\rangle\langle f_j\rangle}\left\langle(\bar{f}_i - \langle f_i\rangle)(\bar{f}_j - \langle f_j\rangle)\right\rangle + \cdots$$

$$= F(\{\langle f_i\rangle\}) + \frac{1}{2}\sum_{i,j}\left.\frac{\partial^2 F}{\partial f_i \partial f_j}\right|_{\langle f_i\rangle\langle f_j\rangle}\mathrm{cov}(\bar{f}_i, \bar{f}_j) + \cdots$$

$$F(\{\langle f_i\rangle\}) = \left\langle F\left(\{\bar{f}_i\}\right)\right\rangle - \frac{1}{2}\sum_{i,j}\left.\frac{\partial^2 F}{\partial f_i \partial f_j}\right|_{\bar{f}_i\bar{f}_j}\frac{\mathrm{cov}(f_i, f_j)}{N} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

$$= \left\langle F\left(\{\bar{f}_i\}\right)\right\rangle - \frac{1}{2}\sum_{i,j}\left.\frac{\partial^2 F}{\partial f_i \partial f_j}\right|_{\bar{f}_i\bar{f}_j}\frac{\overline{f_i f_j} - \overline{f}_i\,\overline{f}_j}{N-1} + \mathcal{O}\left(\frac{1}{N^2}\right)$$

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

$$
\begin{aligned}
\mathrm{Var}(F(\{\bar{f}_i\})) &= \left\langle \left( F(\{\bar{f}_i\}) - F(\{\langle f_i \rangle\}) \right)^2 \right\rangle \\
&\approx \sum_{i,j} \frac{\partial F}{\partial f_i} \frac{\partial F}{\partial f_j} \bigg|_{\bar{f}_i \bar{f}_j} \left\langle (\bar{f}_i - \langle f_i \rangle)(\bar{f}_j - \langle f_j \rangle) \right\rangle \\
&\approx \sum_{i,j} \frac{\partial F}{\partial f_i} \frac{\partial F}{\partial f_j} \bigg|_{\bar{f}_i \bar{f}_j} \mathrm{Cov}(\overline{f_i}, \overline{f_j}) \\
&\approx \sum_{i,j} \frac{\partial F}{\partial f_i} \frac{\partial F}{\partial f_j} \bigg|_{\bar{f}_i \bar{f}_j} \frac{\overline{f_i f_j} - \overline{f_i}\,\overline{f_j}}{N-1}
\end{aligned}
$$

Since the biases go down with increasing $N$, even when we break a long MC run into blocks for estimating the statistical error, in order to calculate the least biased estimate of $F$ one should use $F$ applied to the global averages of $\{f_i\}$ rather than averaging over $F$ applied to the block averages of $\{f_i\}$.

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

$$\begin{aligned}
\text{Estim. of mean} \quad \langle f \rangle_\rho &= \overline{f} \\
\text{Estim. of variance} \quad \langle f^2 \rangle - \langle f \rangle_\rho^2 &= \frac{N}{N-1}\left(\overline{f^2} - \overline{f}^2\right) \\
\text{Estim. of error of sample mean} &= \sqrt{\frac{1}{N-1}\left(\overline{f^2} - \overline{f}^2\right)} \\
\text{Estim. of covar.} \quad \mathrm{cov}(f,g) \equiv \langle fg \rangle - \langle f \rangle_\rho \langle g \rangle_\rho &= \frac{N}{N-1}\left(\overline{fg} - \overline{f}\,\overline{g}\right) \\
\text{Estim. of product of expec. values} \quad \langle f \rangle_\rho \langle g \rangle_\rho &\approx \overline{f}\,\overline{g} - \frac{1}{N}\mathrm{cov}(f,g) \\
\text{Estim. of ratio of expec. values} \quad \frac{\langle f \rangle_\rho}{\langle g \rangle_\rho} &\approx \frac{\overline{f}}{\overline{g}} - \frac{1}{N}\left(\frac{\overline{f}\,\sigma_g^2}{\overline{g}^3} - \frac{\mathrm{cov}(f,g)}{\overline{g}^2}\right) \\
\mathrm{Var}\left(\overline{f}\,\overline{g}\right) &\approx \frac{1}{N}\left(\overline{g}^2\sigma_f^2 + \overline{f}^2\sigma_g^2 + 2\overline{f}\,\overline{g}\,\mathrm{cov}(f,g)\right) \\
\mathrm{Var}\left(\frac{\overline{f_\rho}}{\overline{g_\rho}}\right) &\approx \frac{1}{N}\left(\frac{\sigma_f^2}{\overline{g}^2} + \frac{\overline{f}^2\sigma_g^2}{\overline{g}^4} - 2\frac{\overline{f}\,\mathrm{cov}(f,g)}{\overline{g}^3}\right).
\end{aligned}$$

Note that the product, $\overline{f}\,\overline{g}$ is unbiased if $\mathrm{cov}(f,g) = 0$, but the ratio $\frac{\overline{f}}{\overline{g}}$ 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 covariances by sample covariances on RHS.

# Unbiased Estimators of autocorrelated variables

Independent samples:

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

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

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

where

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

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

If samples are indep., $\langle f_1 f_{1+t}\rangle_\rho = \langle f\rangle_\rho^2$ 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.

# $N_{\rm eff}$

Note that there are 2 reasons why $N_{\rm eff}$ may be smaller than $N$.

1. Serial correlations, as in Metropolis-Hastings method.
2. Weighted walkers, as in some projector MC methods (discussed later)

# Variational Monte Carlo

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

D. Ceperley, G. V. Chester and M. H. Kalos, PRB 16, 3081 (1977) (Fermions)

# 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 $\mathbf{R}_\mathrm{i} \to \mathbf{R}_\mathrm{f}$.)

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

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 $\rho$, will continue to sample $\rho$.
Want more than that: *any* initial density should evolve to $\rho$.

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

i.e., $\rho$ should be the *dominant* right eigenvector.

# Metropolis-Hastings Monte Carlo (cont)

Want that any initial density should evolve to $\rho$.

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

$\rho$ 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 $\rho$ 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 $A$

$$\frac{A(\mathbf{R}_f|\mathbf{R}_i)}{A(\mathbf{R}_i|\mathbf{R}_f)} = \frac{T(\mathbf{R}_i|\mathbf{R}_f)\ \rho(\mathbf{R}_f)}{T(\mathbf{R}_f|\mathbf{R}_i)\ \rho(\mathbf{R}_i)}\ .$$

Infinity of choices for $A$. Any function

$$F\left(\frac{T(\mathbf{R}_i|\mathbf{R}_f)\ \rho(\mathbf{R}_f)}{T(\mathbf{R}_f|\mathbf{R}_i)\ \rho(\mathbf{R}_i)}\right) = A(\mathbf{R}_f|\mathbf{R}_i)$$

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

$$A(\mathbf{R}_f|\mathbf{R}_i) = \min\left\{1,\ \frac{T(\mathbf{R}_i|\mathbf{R}_f)\ \rho(\mathbf{R}_f)}{T(\mathbf{R}_f|\mathbf{R}_i)\ \rho(\mathbf{R}_i)}\right\}.$$

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

$$A(\mathbf{R}_f|\mathbf{R}_i) = \frac{T(\mathbf{R}_i|\mathbf{R}_f)\ \rho(\mathbf{R}_f)}{T(\mathbf{R}_i|\mathbf{R}_f)\ \rho(\mathbf{R}_f) + T(\mathbf{R}_f|\mathbf{R}_i)\ \rho(\mathbf{R}_i)}.$$

Metropolis: $T(\mathbf{R}_i|\mathbf{R}_f) = T(\mathbf{R}_f|\mathbf{R}_i)$, Hastings: $T(\mathbf{R}_i|\mathbf{R}_f) \neq T(\mathbf{R}_f|\mathbf{R}_i)$

# Metropolis-Hastings Monte Carlo (cont)

## Choice of Proposal Matrix $T$

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

However, there is considerable scope for using one's ingenuity to come up with good proposal matrices, $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$, 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)

$$A(\mathbf{R}_f|\mathbf{R}_i) = \min\left\{1, \frac{T(\mathbf{R}_i|\mathbf{R}_f)\ \rho(\mathbf{R}_f)}{T(\mathbf{R}_f|\mathbf{R}_i)\ \rho(\mathbf{R}_i)}\right\}$$

Use freedom in $T$ to make $\quad \dfrac{T(\mathbf{R}_i|\mathbf{R}_f)\ \rho(\mathbf{R}_f)}{T(\mathbf{R}_f|\mathbf{R}_i)\ \rho(\mathbf{R}_i)} \approx 1.$

$T(\mathbf{R}_f|\mathbf{R}_i) \propto \rho(\mathbf{R}_f)$ optimal if $T(\mathbf{R}_f|\mathbf{R}_i)$ 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 $\quad T(\mathbf{R}_f|\mathbf{R}_i) = \dfrac{S(\mathbf{R}_f|\mathbf{R}_i)}{\int d\mathbf{R}_f\, S(\mathbf{R}_f|\mathbf{R}_i)} \approx \dfrac{S(\mathbf{R}_f|\mathbf{R}_i)}{S(\mathbf{R}_i|\mathbf{R}_i)\Omega(\mathbf{R}_i)}$

$S(\mathbf{R}_f|\mathbf{R}_i)$ is non-zero only in domain $D(\mathbf{R}_i)$ of volume $\Omega(\mathbf{R}_i)$ around $\mathbf{R}_i$).

$$\frac{A(\mathbf{R}_f,\mathbf{R}_i)}{A(\mathbf{R}_i,\mathbf{R}_f)} = \frac{T(\mathbf{R}_i|\mathbf{R}_f)}{T(\mathbf{R}_f|\mathbf{R}_i)}\frac{\rho(\mathbf{R}_f)}{\rho(\mathbf{R}_i)} \approx \frac{\Omega(\mathbf{R}_i)}{\Omega(\mathbf{R}_f)}\frac{S(\mathbf{R}_i|\mathbf{R}_i)}{S(\mathbf{R}_f|\mathbf{R}_f)}\frac{S(\mathbf{R}_i|\mathbf{R}_f)}{S(\mathbf{R}_f|\mathbf{R}_i)}\frac{\rho(\mathbf{R}_f)}{\rho(\mathbf{R}_i)}$$

from which it is apparent that the choice

$$S(\mathbf{R}_f|\mathbf{R}_i) \overset{\propto}{\approx} \sqrt{\rho(\mathbf{R}_f)/\Omega(\mathbf{R}_f)} \quad \text{yields} \quad A(\mathbf{R}_f,\mathbf{R}_i)/A(\mathbf{R}_i,\mathbf{R}_f) \approx 1.$$

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

To be more precise, if the log-derivatives of $T(\mathbf{R}_f|\mathbf{R}_i)$ equal those of $\sqrt{\rho(\mathbf{R}_f)/\Omega(\mathbf{R}_f)}$ at $\mathbf{R}_f = \mathbf{R}_i$, 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 $\Delta$ is the linear dimension of $D(\mathbf{R}_i)$.

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

Another possible choice, motivated by (DMC) is

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

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 $S(\mathbf{R}_f|\mathbf{R}_i)$ or one for which $S(\mathbf{R}_f|\mathbf{R}_i) \overset{\propto}{\sim} \rho(\mathbf{R}_f)$, but less efficient than one for which $S(\mathbf{R}_f|\mathbf{R}_i) \overset{\propto}{\sim} \sqrt{\rho(\mathbf{R}_f)/\Omega(\mathbf{R}_f)}$.

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

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

Another possible choice, motivated by (DMC) is

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

Advantage: allows Metropolis Monte Carlo and diffusion Monte Carlo programs to share almost all the code.

# Some examples

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

$$T(\mathbf{R}_f|\mathbf{R}_i) = \frac{S(\mathbf{R}_f|\mathbf{R}_i)}{\int d\mathbf{R}_f\, S(\mathbf{R}_f|\mathbf{R}_i)} \approx \frac{S(\mathbf{R}_f|\mathbf{R}_i)}{S(\mathbf{R}_i|\mathbf{R}_i)\Omega(\mathbf{R}_i)}$$

and since the acceptance is

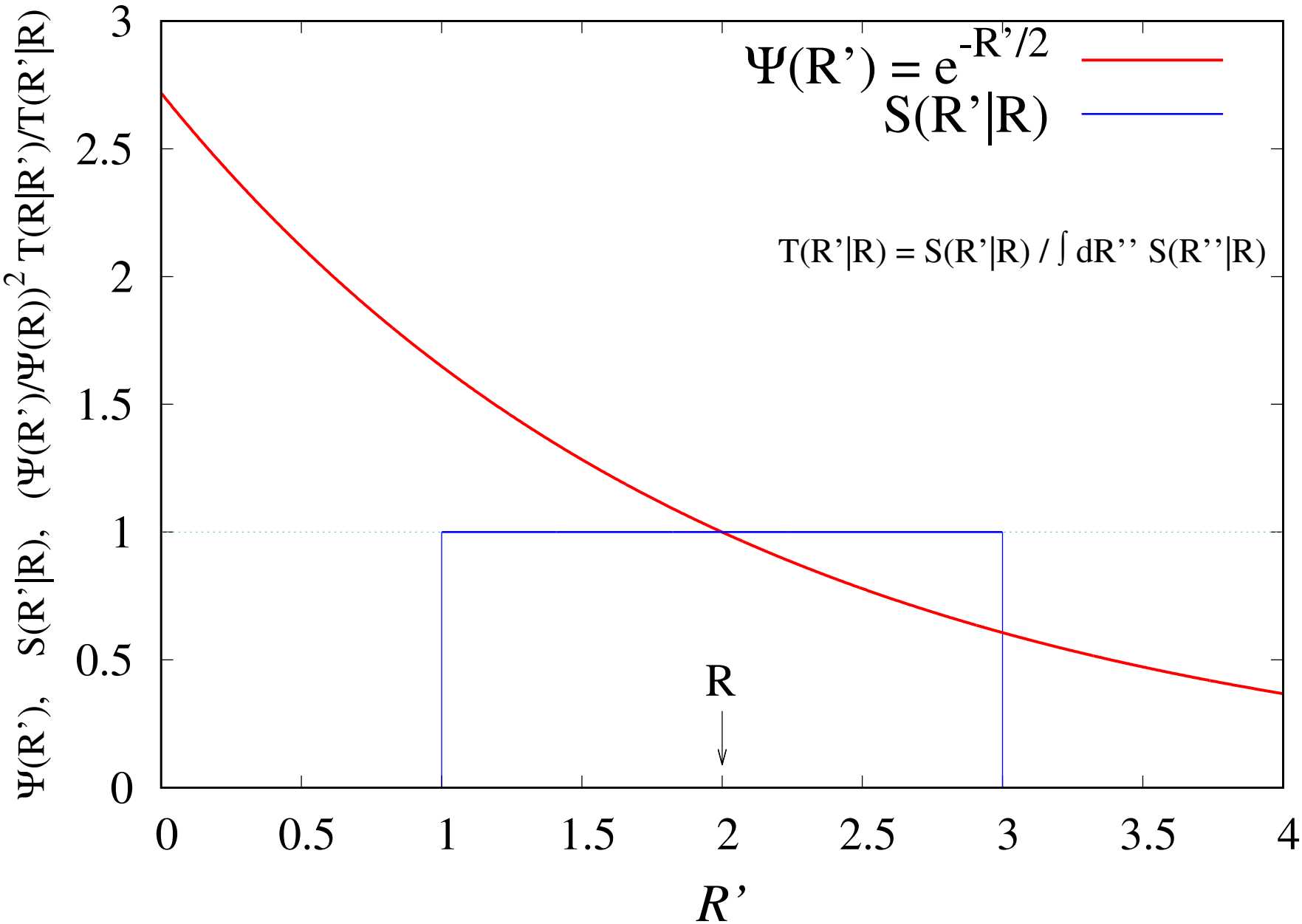
$$A(\mathbf{R}_f|\mathbf{R}_i) = \min\left\{1, \frac{|\Psi(\mathbf{R}_f)|^2\ T(\mathbf{R}_i|\mathbf{R}_f)}{|\Psi(\mathbf{R}_i)|^2\ T(\mathbf{R}_f|\mathbf{R}_i)}\right\}$$

we want

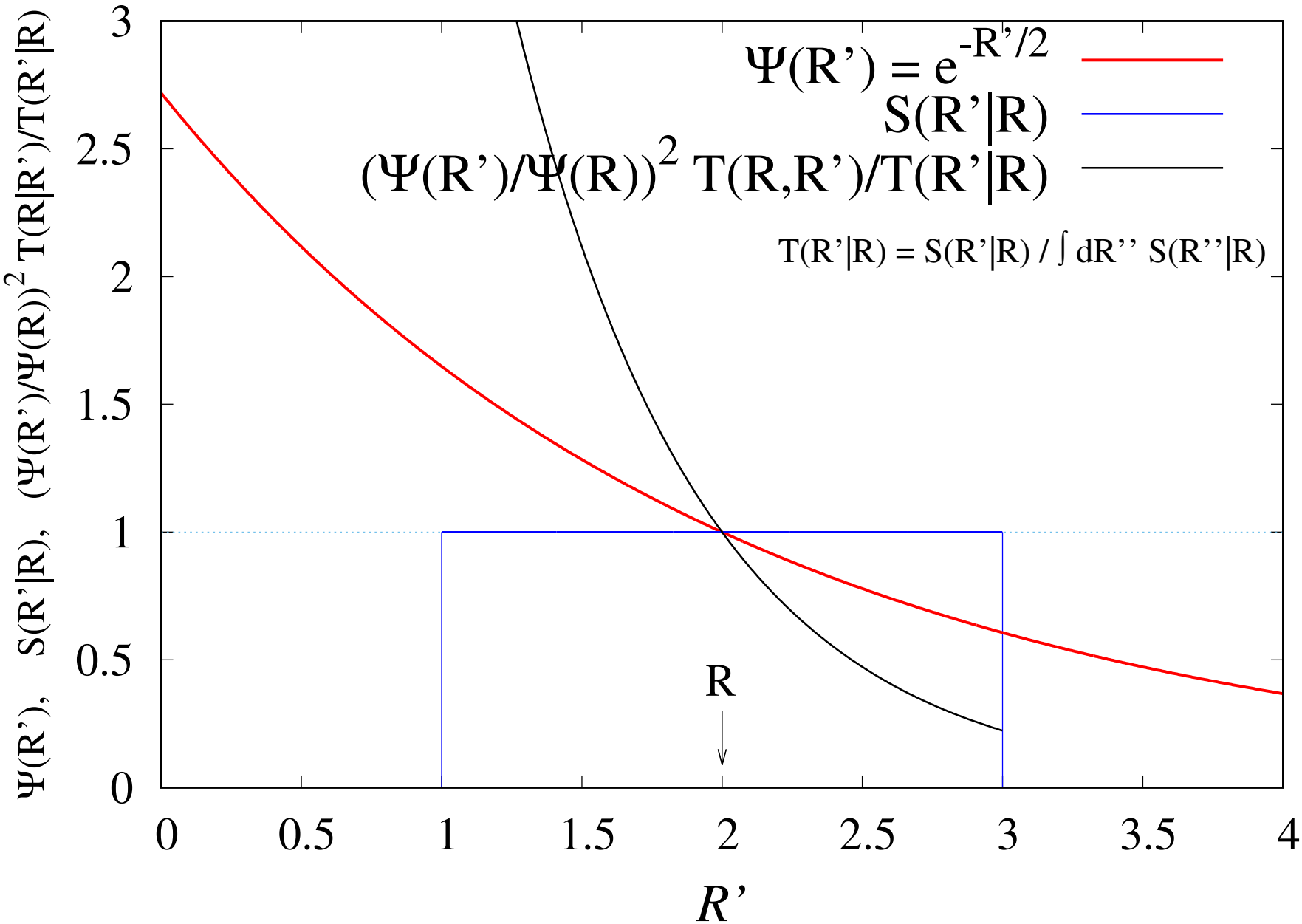
$$\frac{|\Psi(\mathbf{R}_f)|^2\ T(\mathbf{R}_i|\mathbf{R}_f)}{|\Psi(\mathbf{R}_i)|^2\ T(\mathbf{R}_f|\mathbf{R}_i)}$$

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

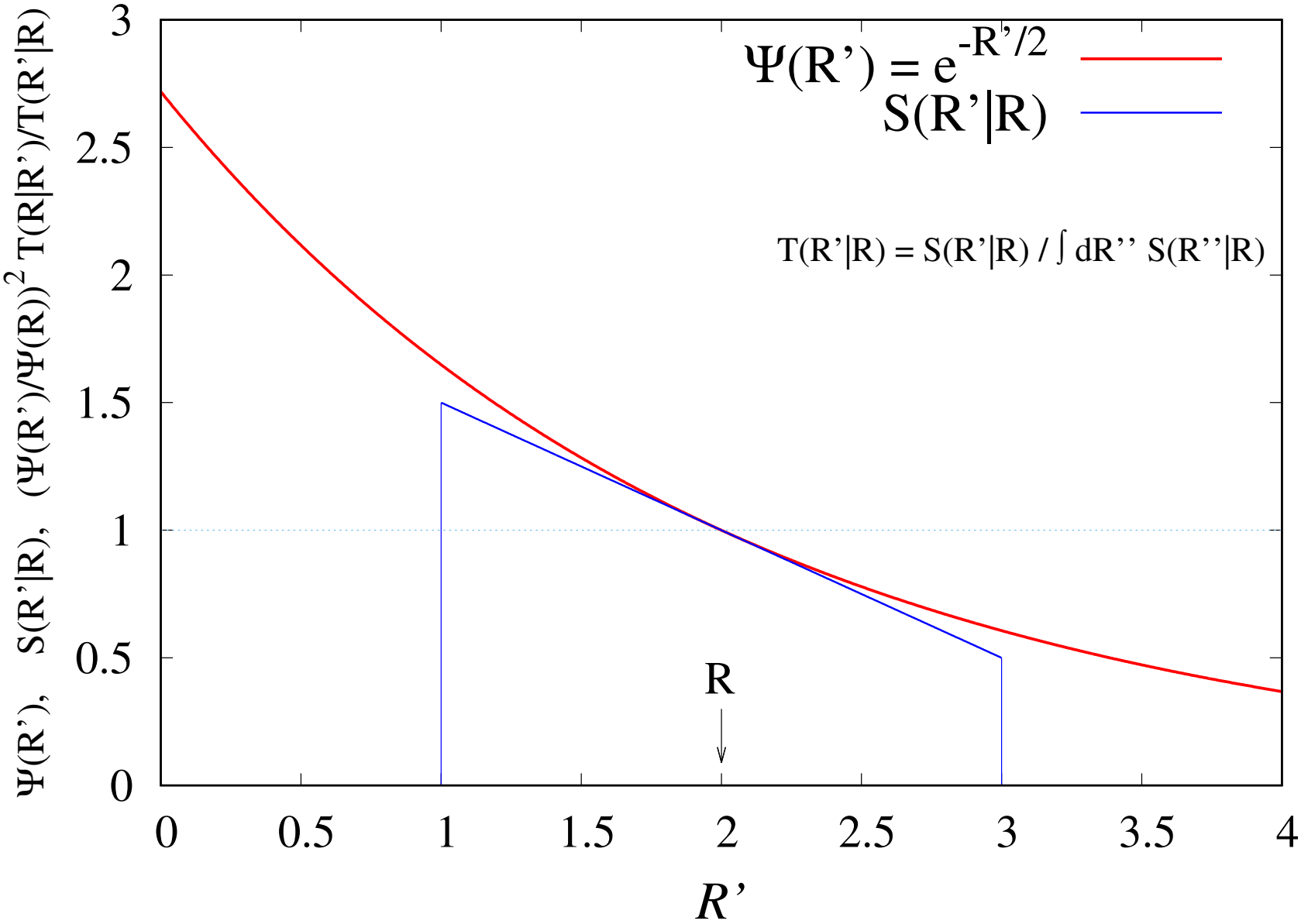
# Symmetrical $T$ in Metropolis



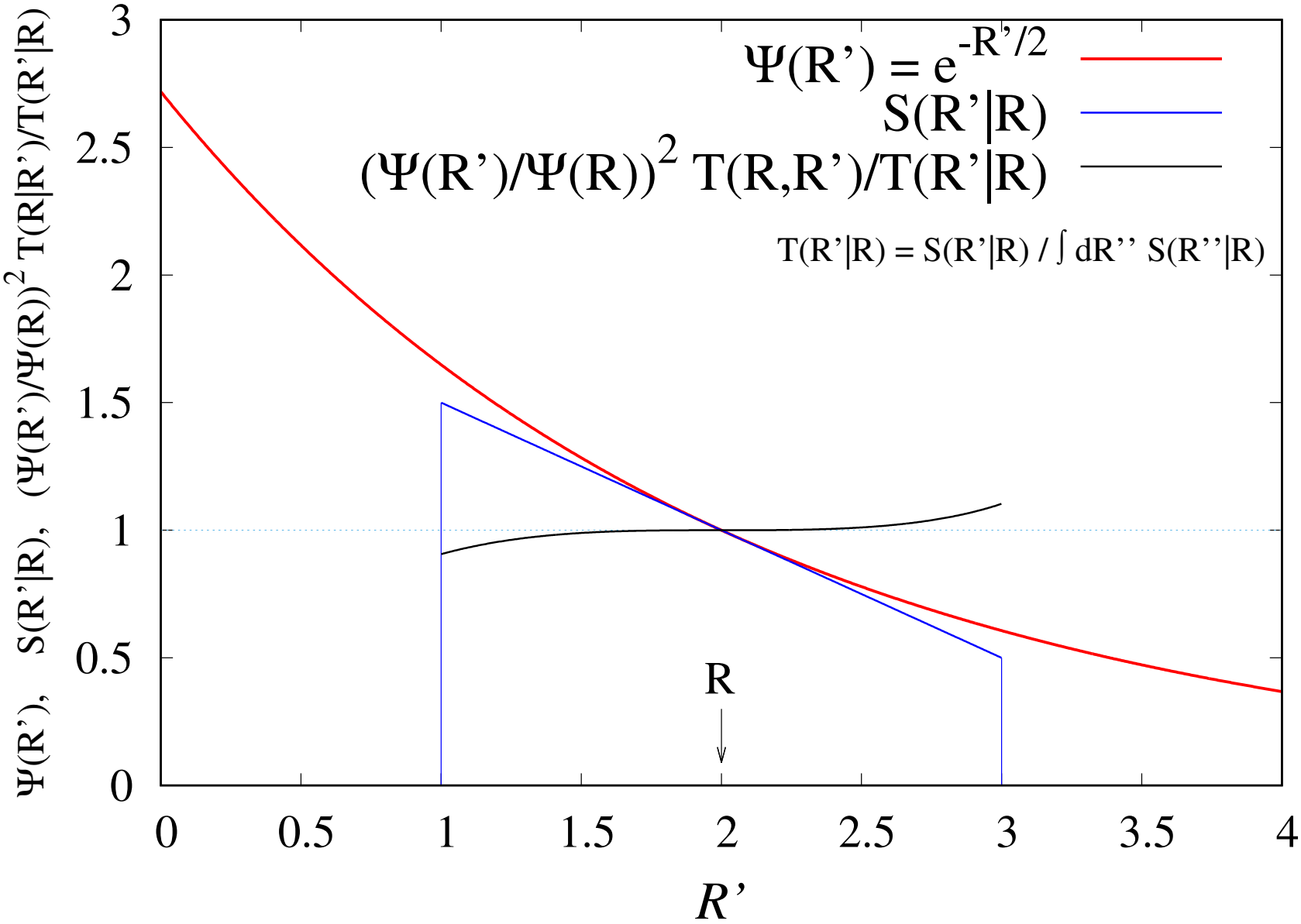
# Symmetrical $T$ in Metropolis



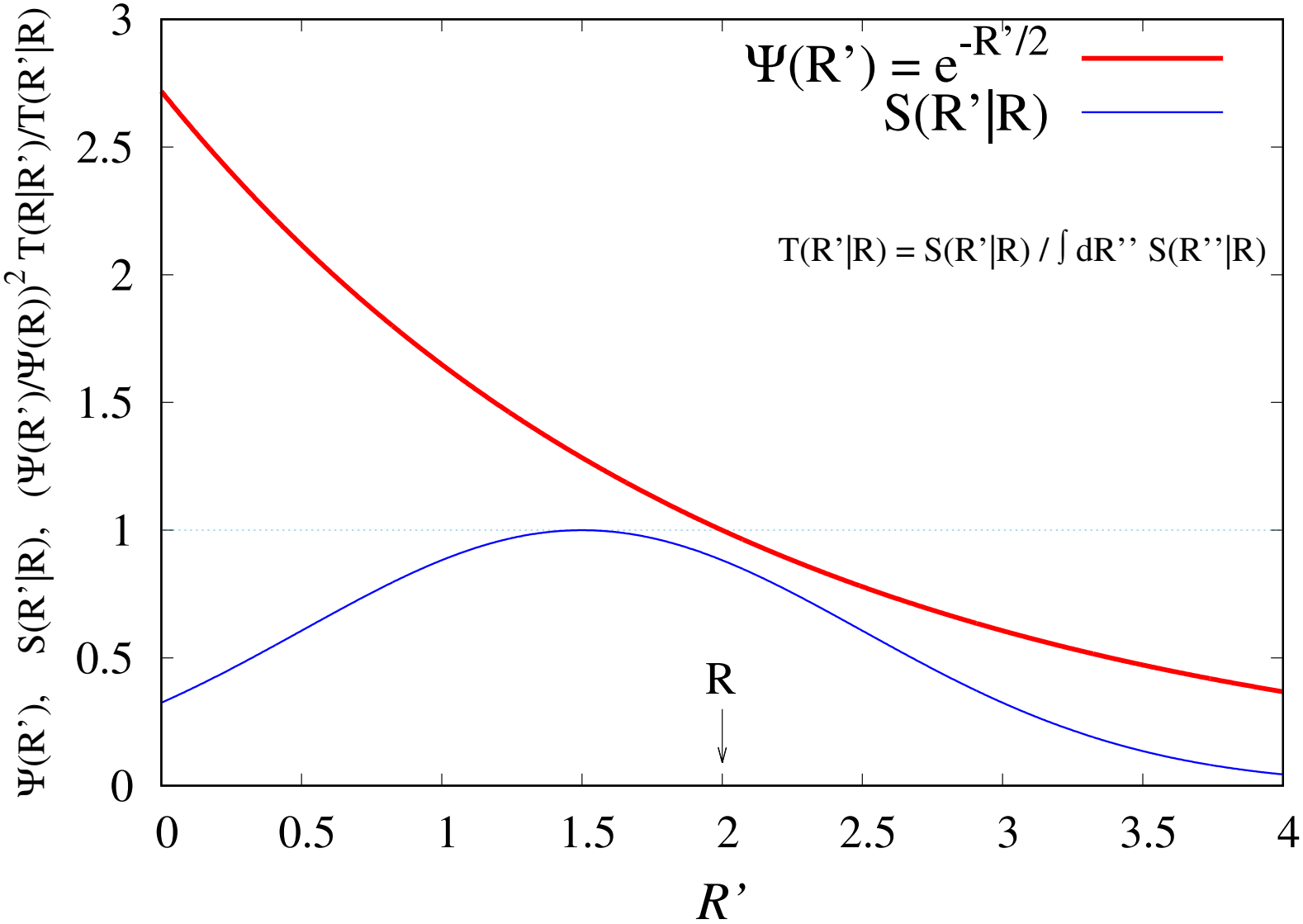
# Non-symmetrical linear $T$ in Metropolis-Hastings



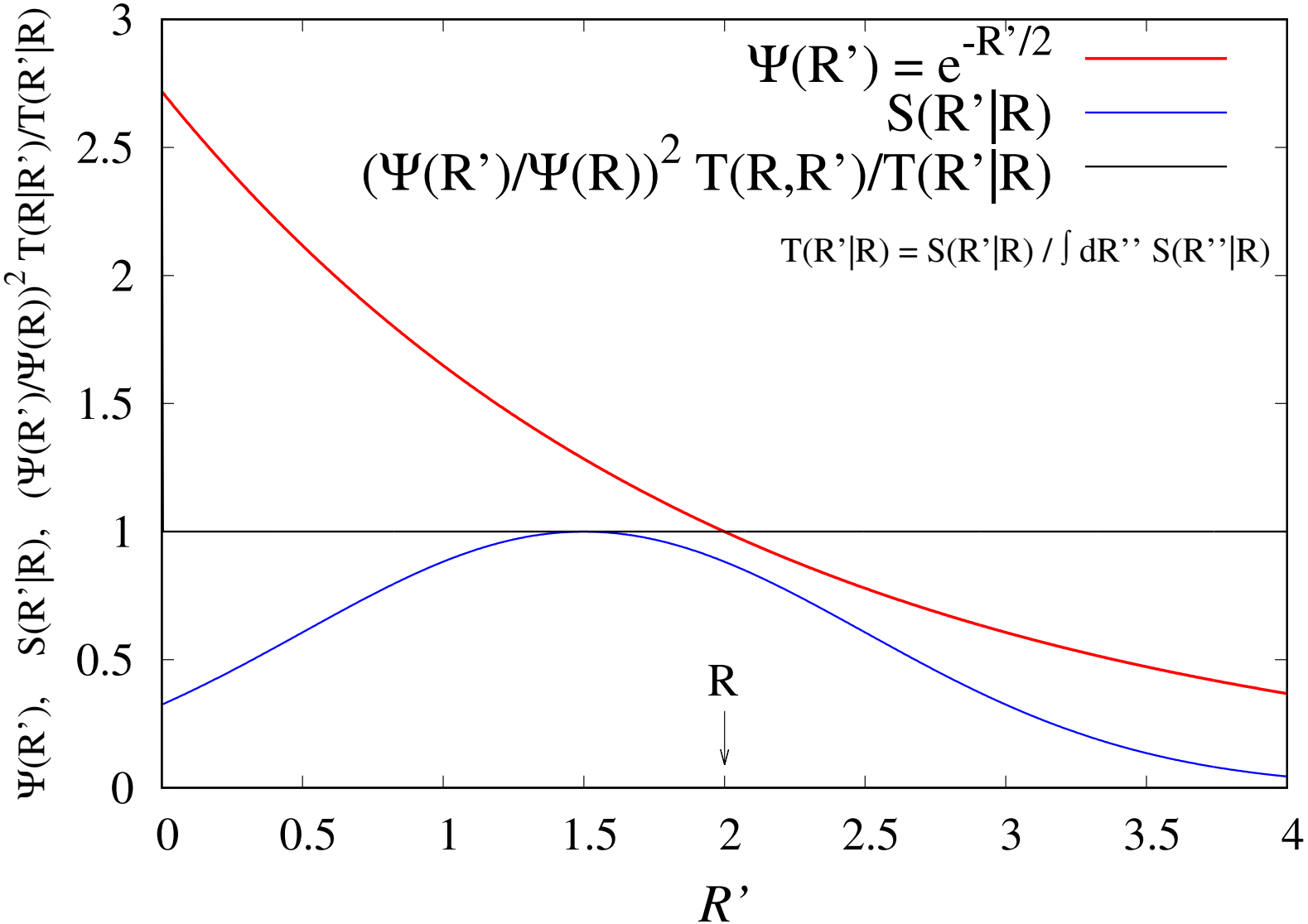
# Non-symmetrical linear $T$ in Metropolis-Hastings



# Non-symmetrical drifted Gaussian $T$ in Metropolis-Hastings

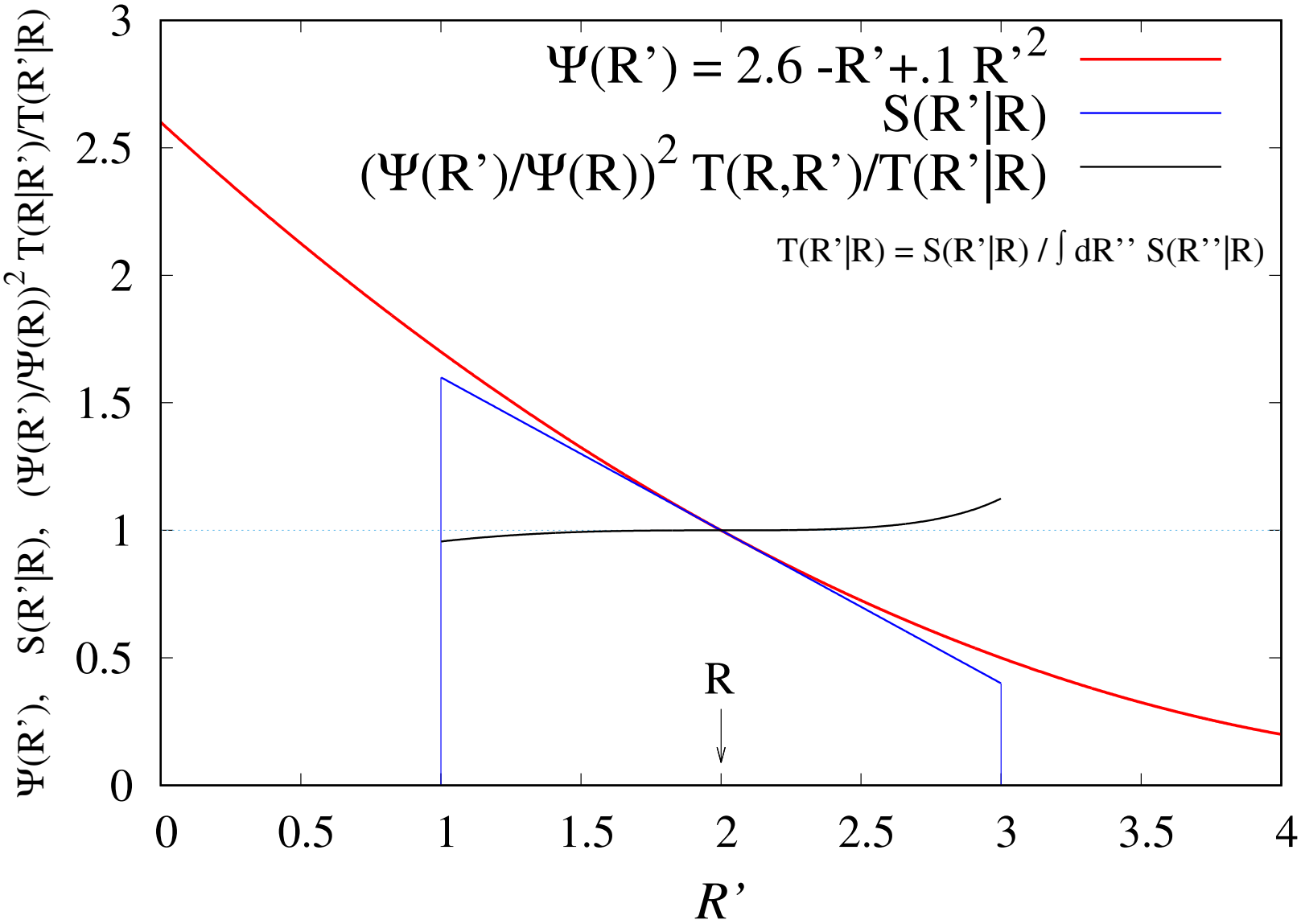


# Non-symmetrical drifted Gaussian $T$ in Metropolis-Hastings



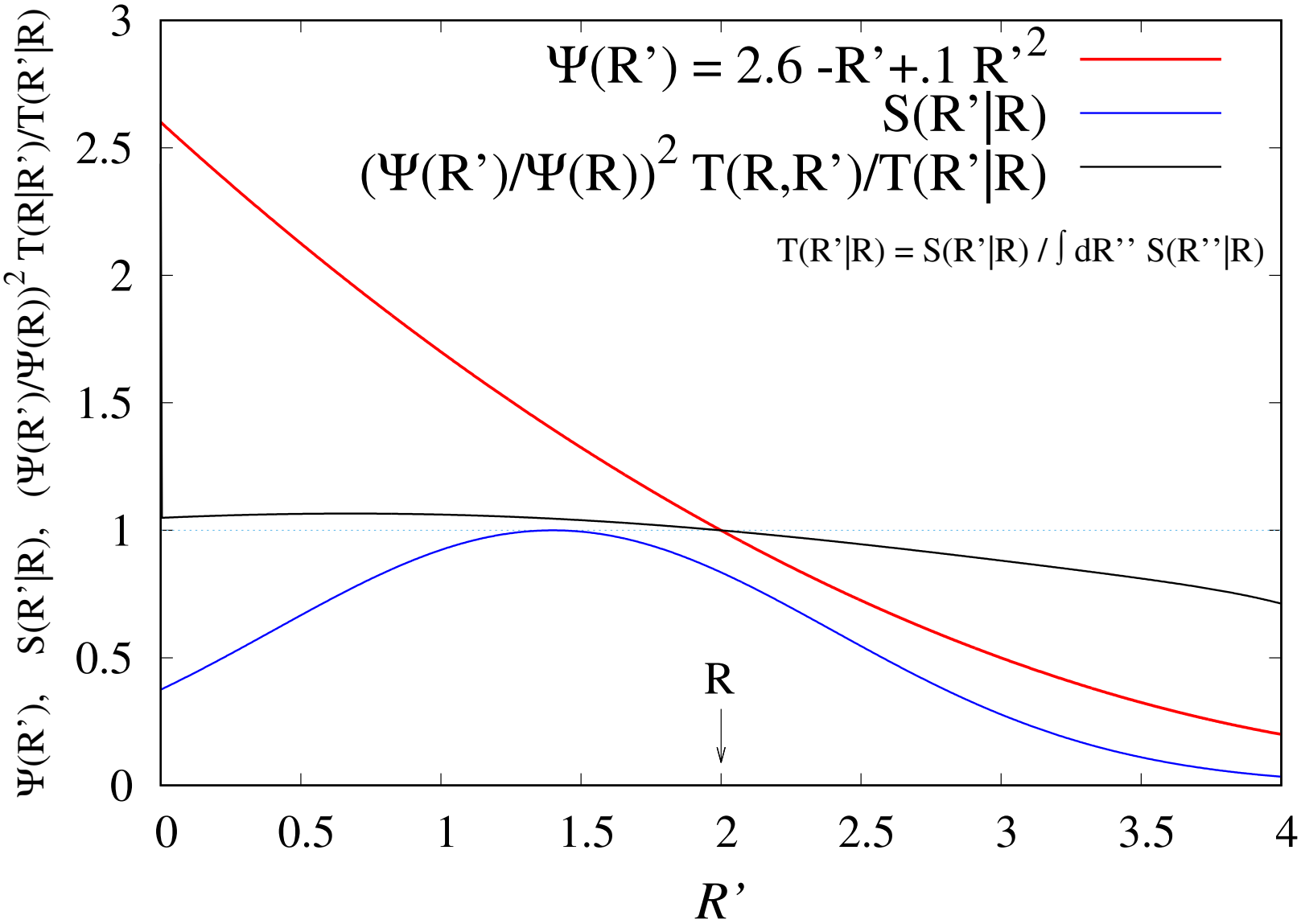

For this $\Psi(R')$, the drifted Gaussian gives perfect acceptance since **V** is constant and drift and diffusion commute! Not true for general $\Psi(R')$.

# 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?

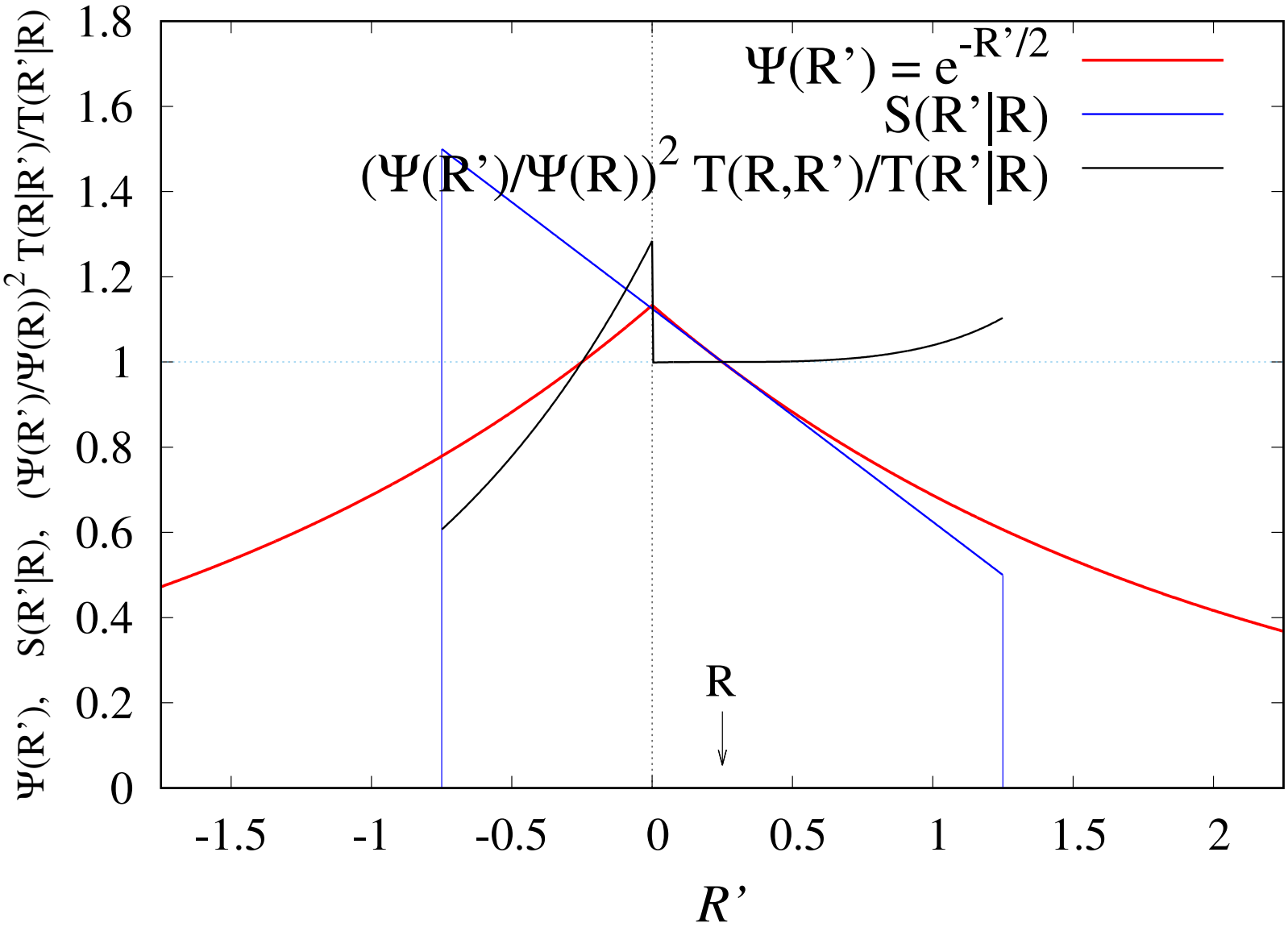
# Choice of Proposal Matrix $T$ in Metropolis-Hastings (cont)

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.

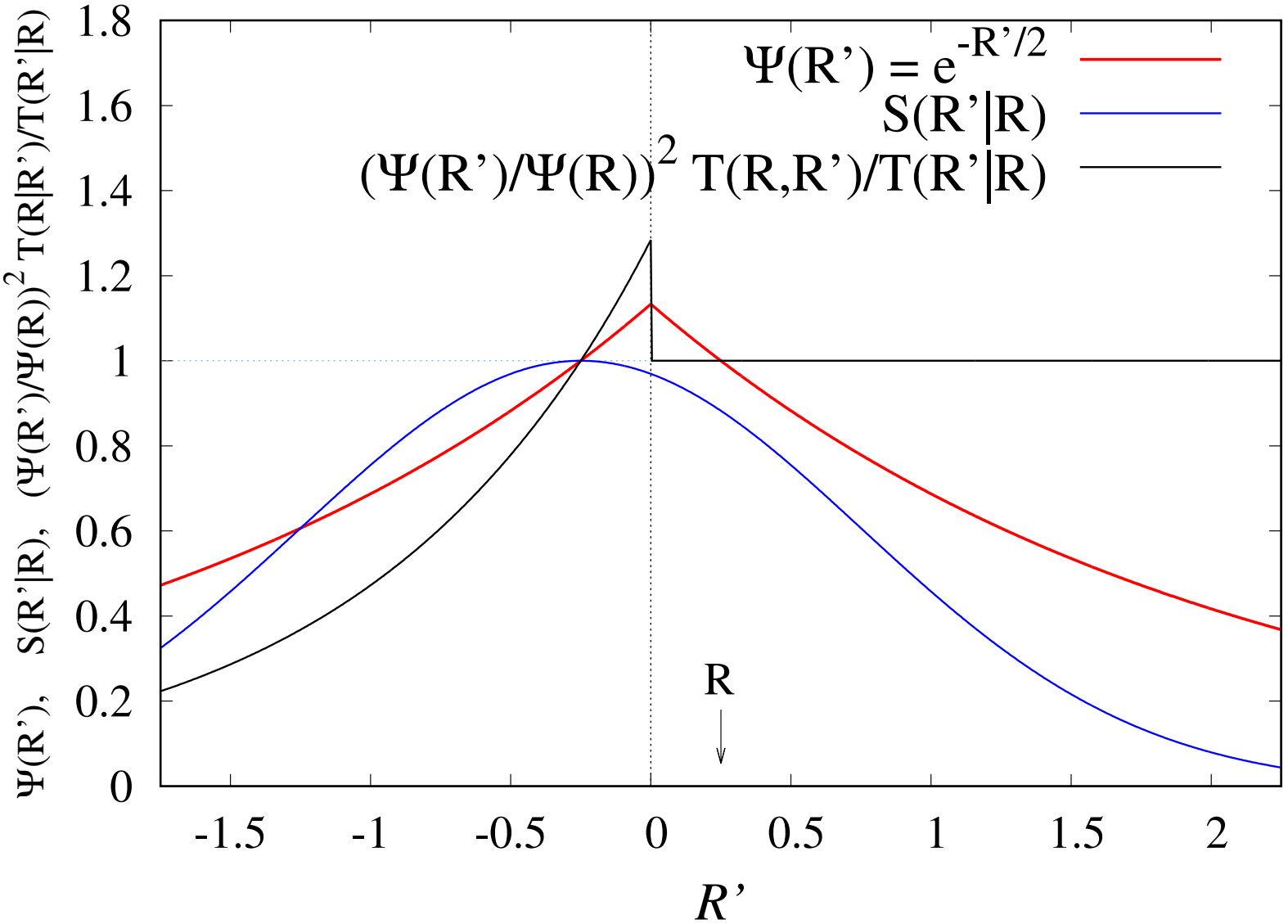
Let's see what happens when it is not.

# 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 $T_{\rm corr} \approx 1$.

# Choice of Proposal Matrix $T$ in Metropolis-Hastings (cont)
# Generalization to molecules and solids

Extension to molecules is simple. Do everything as before but relative to the nearest nucleus.

The nearest nucleus before and after the move need not be the same.

For some proposed moves the reverse move may not be possible.

In that case detailed balance demands that the proposed move be rejected.

Since these rejections can be done on purely geometrical grounds (they do not require evaluation of $\psi_{\rm T}(\mathbf{R})$ or its gradient or Laplacian) these rejections do not lead to any appreciable loss of efficiency.

Bottom line is one can get $T_{\rm corr}$ close to 1.

# Metropolis-Hastings Monte Carlo (cont)

## Some Observations about Metropolis-Hastings Method

1. To sample states with relative density $\rho$ it is not necessary to know the normalization of $\rho$. Metropolis automatically samples $\rho(\mathbf{R}_\mathrm{i})/\int d\mathbf{R}_\mathrm{f}\ \rho(\mathbf{R}_\mathrm{f})$. So, it is useful for calculating quantities of the form

$$\frac{\int d\mathbf{R}_\mathrm{i}\ e(\mathbf{R}_\mathrm{i})\ \rho(\mathbf{R}_\mathrm{i})}{\int d\mathbf{R}_\mathrm{f}\ \rho(\mathbf{R}_\mathrm{f})}$$

which is the form encountered in quantum mechanics and statistical mechanics.
(Can also be used to calculate $\int d\mathbf{R}_\mathrm{i}\ f(\mathbf{R}_\mathrm{i})$, with importance sampling provided one has a $g(\mathbf{R}_\mathrm{i})$ that mimics $f(\mathbf{R}_\mathrm{i})$ but whose integral is known. Of course if in addition $g(\mathbf{R}_\mathrm{i})$ can be sampled directly then one would not use Metropolis, so this is rarely useful.)

2. The variance of the estimate for the expectation value $\langle X \rangle$ is given by

$$\frac{1}{N/T_\mathrm{corr} - 1}\left(\frac{\sum X(\mathbf{R}_\mathrm{i})^2}{N} - \left(\frac{\sum X(\mathbf{R}_\mathrm{i})}{N}\right)^2\right).$$

That is, the effective number of configurations $N_\mathrm{eff}$ is smaller than $N$ by a factor of $T_\mathrm{corr}$, which we define to be the autocorrelation time.
($T_\mathrm{corr} = 1 + 2t_\mathrm{corr}$, where $t_\mathrm{corr} = \sum_i^\infty \rho_i$ is the integrated autocorrelation time.)

# Metropolis-Hastings Monte Carlo (cont)

## Some Observations about Metropolis-Hastings Method

3. The rate of convergence to the desired density and the autocorrelation time of estimates of observables is governed by the sub-dominant eigenvalues of $M$. In practice reduce $T_{\rm corr}$ by inventing large moves that have large acceptance probabilities.
4. Folklore: when one can choose the range of the PDF from which moves are proposed the optimal one has an average acceptance close to 0.5. Reasonable choice in absence of any information, but in fact the optimal choice may have an average acceptance that is anywhere between zero and one.
   I have found instances where the optimum is as small as 0.2 or as large as 0.9.
   A much better criterion is to maximize the rate at which the system diffuses through configuration space $\langle A(\mathbf{R}_{\rm f}|\mathbf{R}_{\rm i})(\mathbf{R}_{\rm f}-\mathbf{R}_{\rm i})^2\rangle$.
   The real measure of goodness is of course to minimize the autocorrelation time for the observables of interest.

# Metropolis-Hastings Monte Carlo (cont)
## Some Observations about Metropolis-Hastings Method

7. The *Gibbs sampler* or *heat bath* algorithm can be considered to be a special case of the (generalized) Metropolis method. $T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) \propto \rho(\mathbf{R}_\mathrm{f})$ for only a small set of accessible states in a domain $D(\mathbf{R}_\mathrm{i})$ in the neighborhood of $\mathbf{R}_\mathrm{i}$:

$$T(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i}) = \begin{cases} \rho(\mathbf{R}_\mathrm{f})/\sum\rho(\mathbf{R}_\mathrm{f}) & \text{if } \mathbf{R}_\mathrm{f} \,\epsilon\, D(\mathbf{R}_\mathrm{i}) \\ 0 & \text{otherwise} \end{cases}$$

If the sum over the accessible states from $\mathbf{R}_\mathrm{i}$ and $\mathbf{R}_\mathrm{f}$ is the same, the acceptance is unity.

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

The heat bath algorithm was frequently used for lattice models where the normalization constant can be easily computed, e.g. Potts model.

One may ask, how are all states accessed if accessible states from $\mathbf{R}_\mathrm{i}$ and $\mathbf{R}_\mathrm{f}$ are the same? $1^{st}$ it is only the sum that must be the same. $2^{nd}$ it is OK because one may be cycling through the spins.

# 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.

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

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:

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

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

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

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

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).

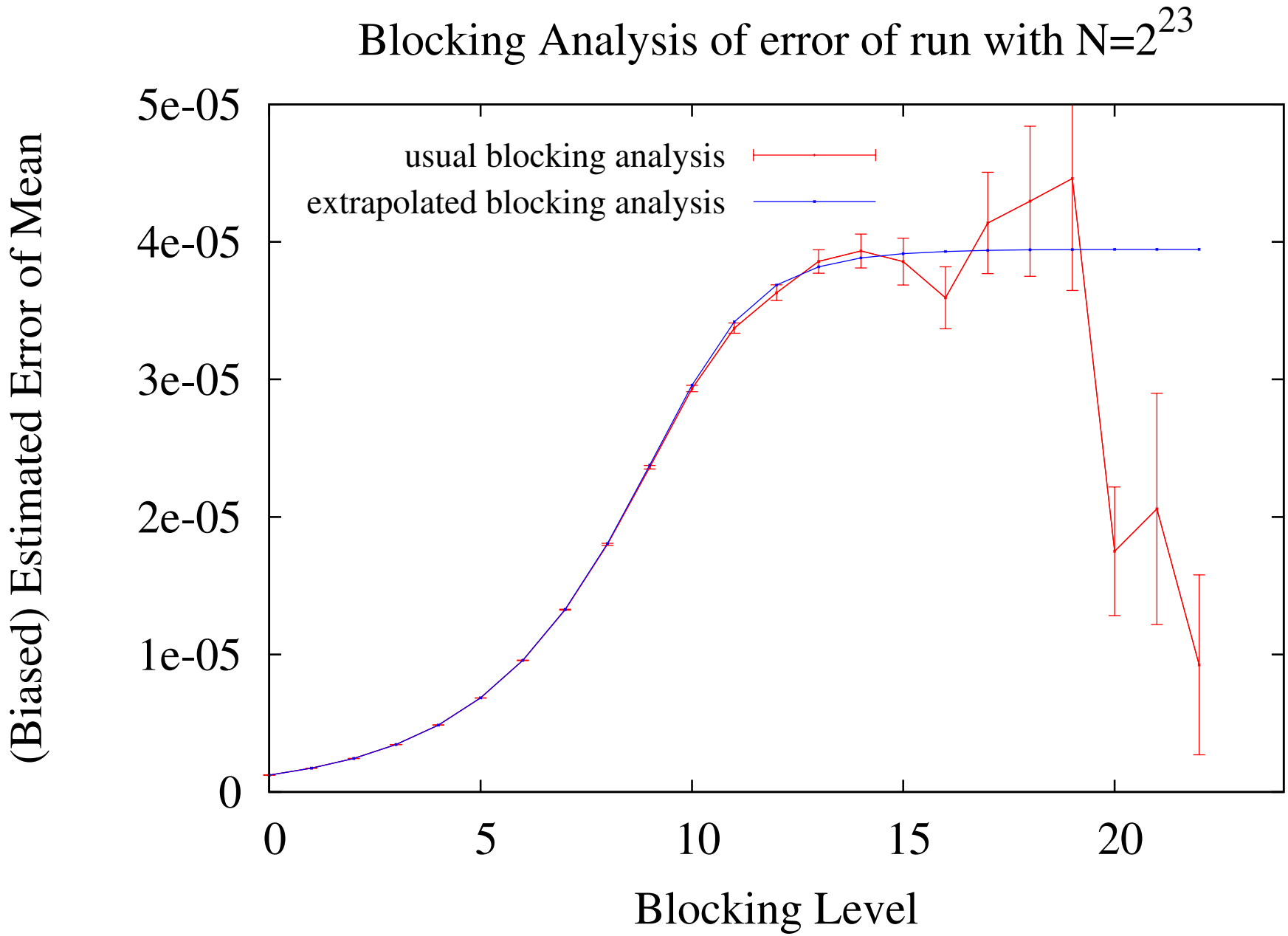
A reasonable choice of blocking level is the highest one for which the increase in the estimate for the error is larger than the increase in the estimate for the error in the error. It is possible to get a somewhat better estimate by predicting the shape of the curve and extrapolating when say $N < 1000\,T_{\rm corr}$, so that $N_s = 100\,T_{\rm corr}$ would leave us with only $N_b = 10$.

# 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 $N = 2^{23} \approx 8 \times 10^6$ 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

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

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

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

The single-particle orbitals $\phi$ are given by:

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

- 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

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

$\sim N_{\mathrm{atomtype}}$ of $\zeta_{k_\alpha}$ parms.

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

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

$d_n$, $c_{k_\alpha}$, $\zeta_{k_\alpha}$ and parms in $\mathcal{J}$ are optimized.

Power of QMC:

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

# Cusp-conditions of Trial Wave Functions

At interparticle coalescence points, the potential diverges as

$-\frac{Z}{r_{i\alpha}}$ for the electron-nucleus potential (orbitals and e-n Jastrow)

$\frac{1}{r_{ij}}$ for the electron-electron potential (e-e Jastrow)

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 $\Psi_0$)

$\Rightarrow$ 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

Use relative coordinates close to $\mathbf{r}_{ij} = 0$ and keep only divergent terms in $\dfrac{\mathcal{H}\Psi}{\Psi}$

$$\frac{\mathcal{H}\Psi}{\Psi} = -\frac{1}{2\mu_{ij}}\frac{\nabla^2\Psi}{\Psi} + \frac{q_i q_j}{r} \sim -\frac{1}{2\mu_{ij}}\frac{\Psi''}{\Psi} - \frac{1}{\mu_{ij}}\frac{1}{r}\frac{\Psi'}{\Psi} + \frac{q_i q_j}{r}$$

$$\sim \boxed{-\frac{1}{\mu_{ij}}\frac{1}{r}\frac{\Psi'}{\Psi} + \frac{q_i q_j}{r}}$$

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

# Cusp-conditions of Trial Wave Functions

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

$$\left.\frac{\Psi'}{\Psi}\right|_{r_{ij}=0} = \mu_{ij} q_i \, q_j \quad \text{(Kato cusp condition)}$$

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

  - e-n cusp imposed on combination of the determinantal part (using Slater basis functions) and the e-n Jastrow
  - e-e cusp imposed on e-e Jastrow

For Electron-electron($\uparrow\uparrow$): $\Psi(r_{ij} = 0) = \hat{\Psi}' = 0$, $\boxed{\left.\frac{\hat{\Psi}'}{\Psi}\right|_{r=0} = 1/4}$

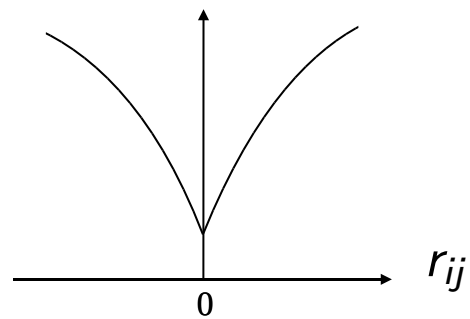
# Jastrow Factors

Example: Simple e-e Jastrow factor

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

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

Imposes cusp conditions
+
keeps electrons apart




- e-e-n Jastrow not needed for cusps, but provides considerable variational freedom in the wavefunction.

# 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:

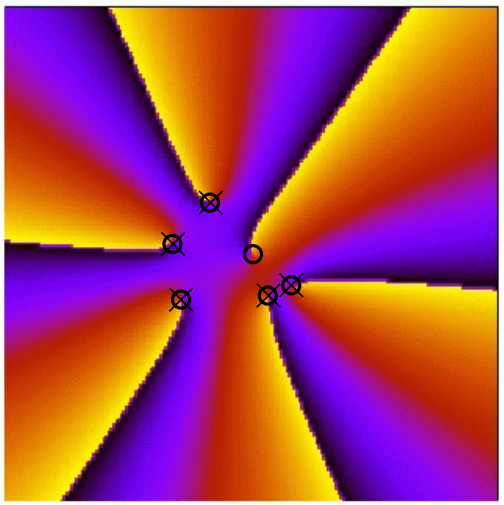
$$D = (1s^{\uparrow}, 2s^{\uparrow}, 1s^{\downarrow}, 2s^{\downarrow}) +$$
$$c\left[(1s^{\uparrow}, 2p_x^{\uparrow}, 1s^{\downarrow}, 2p_x^{\downarrow}) + (1s^{\uparrow}, 2p_y^{\uparrow}, 1s^{\downarrow}, 2p_y^{\downarrow}) + (1s^{\uparrow}, 2p_z^{\uparrow}, 1s^{\downarrow}, 2p_z^{\downarrow})\right]$$

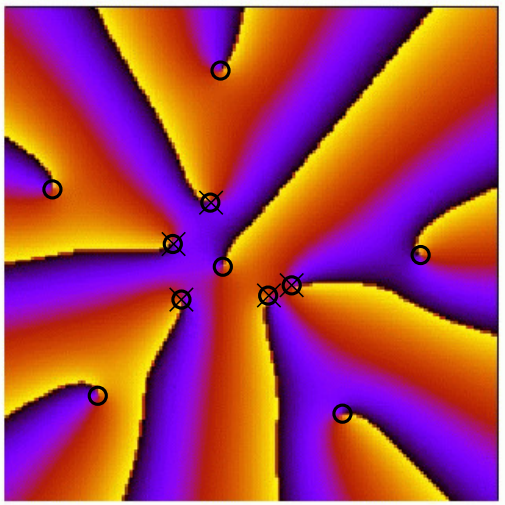
$$\begin{array}{ccc} 1s^2 2s^2 & \times \mathcal{J}(r_{ij}) & \rightarrow E_{\rm VMC}^{\rm corr} = 61\% \\ 1s^2 2s^2 & \times \mathcal{J}(r_{ij}, r_{\alpha i}, r_{\alpha j}) & \rightarrow E_{\rm VMC}^{\rm corr} = 80\% \\ 1s^2 2s^2 \oplus 1s^2 2p^2 & \times \mathcal{J}(r_{ij}, r_{\alpha i}, r_{\alpha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

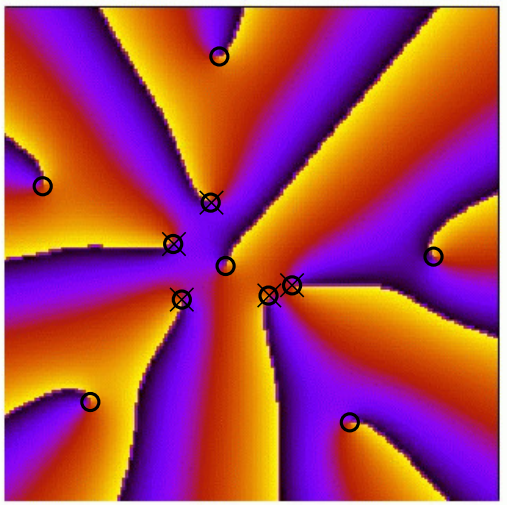
# Wavefunctions for Hole in a Filled Landau Level

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

# Computer lab on VMC and DMC (Tyler)

Get computer lab:

git clone https://github.com/CyrusUmrigar/QMC_module.git QMC_module

# Variational and Projector MC

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

$$E_0 = \underbrace{\frac{\left[\sum_i^{N_{\rm MC}} t_i \; E_{\rm L}(i)\right]_{\Psi_0}}{\left[\sum_i^{N_{\rm MC}} t_i\right]_{\Psi_0}}}_{\text{"no imp. sampl."}} = \underbrace{\frac{\left[\sum_i^{N_{\rm MC}} \frac{t_i}{g_i} E_{\rm L}(i)\right]_{\Psi_{\rm G}\Psi_0}}{\left[\sum_i^{N_{\rm MC}} \frac{t_i}{g_i}\right]_{\Psi_{\rm G}\Psi_0}}}_{\text{"imp. sampl."}}$$ (Value exact†, error depends on $\Psi_{\rm T}, \Psi_{\rm G}$.)

$$E_{\rm L}(i) = \frac{\sum_j^{N_{\rm st}} H_{ij} t_j}{t_i}$$ (Note: $E_{\rm L}(i)$ is $\infty$ if $t_i = 0$, but, $t_i E_{\rm L}(i)$ is finite.)

In both VMC and PMC compute 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 (if space continuous).

† In practice, usually necessary to make fixed-node approximation.

# 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: $\hat{P} = e^{\tau(E_T\hat{\mathbf{1}}-\hat{H})}$ ($\tau$ limited by 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 $g_i e_i = \langle \phi_i | \Psi_G \rangle \langle \phi_i | \Psi_0 \rangle$ 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 $g_i e_i$:

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

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

# Projector vs Path Integral MC Methods

Projector MC and Path Integral MC are 2 ways to interpret the same mathematical expression

$$\langle \hat{O} \rangle = \frac{\langle \Psi_0 | \hat{O} | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle} = \frac{\langle \Psi_T | \hat{P}(\beta)\, \hat{O}\, \hat{P}(\beta) | \Psi_T \rangle}{\langle \Psi_T | \hat{P}(2\beta) | \Psi_T \rangle} \qquad (\beta \text{ large})$$

In practice $\hat{P}(\beta)$ is a convolution of $L$ $\hat{P}(\tau)$ with $\beta = L\tau$

1) If $\hat{O} = \hat{H}$, then it commutes with $\hat{P}$ since $\hat{P}$ is a function of $\hat{H}$. Then

$$E_0 = \frac{\langle \Psi_T | \hat{H}\hat{P} | \Psi_T \rangle}{\langle \Psi_T | \hat{P} | \Psi_T \rangle}$$

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

$$= \underbrace{\frac{\sum_{ik}^{N_{st}} E_L(i)\; t_i\; P_{ik}\; t_k}{\sum_{ik}^{N_{st}} t_i\; P_{ik}\; t_k}} \qquad = \underbrace{\frac{\sum_{ik}^{N_{st}} E_L(i)\; \tilde{P}_{ik}\; t_k^2}{\sum_{ik}^{N_{st}} \tilde{P}_{ik}\; t_k^2}}$$

not importance sampled — importance sampled

(In the above, we assumed for simplicity $\Psi_T = \Psi_G$)

# Projector vs Path Integral MC Methods

$$E_0 = \underbrace{\frac{\sum_{ik}^{N_{\rm st}} E_{\rm L}(i)\, t_i\, P_{ik}\, t_k}{\sum_{ik}^{N_{\rm st}} t_i\, P_{ik}\, t_k}}_{\text{not importance sampled}} = \underbrace{\frac{\sum_{ik}^{N_{\rm st}} E_{\rm L}(i)\, \tilde{P}_{ik}\, t_k^2}{\sum_{ik}^{N_{\rm st}} \tilde{P}_{ik}\, t_k^2}}_{\text{importance sampled}}$$

(In the above, we assumed for simplicity $\Psi_{\rm T} = \Psi_{\rm G}$)

2 ways to turn these into MC expressions:

1. Projector MC: $\because P$ is not column stochastic, need multiplicative weights $w_{ik} = P_{ik}/T_{ik}$, e.g., $T_{ik} = \frac{P_{ik}}{\sum_i P_{ik}}$, $w_k = \sum_i P_{ik}$ (and similarly for imp. sampled).

   If no importance sampling, then do:

   1 Start run by sampling state $k$ from $t_k \equiv \langle \phi_k | \Psi_{\rm T} \rangle$.
   2 At each step sample $i$ from $P_{ik}$ and average $E_{\rm L}(i) t_i W_i$ and $t_i W_i$.

   If importance sampling, then do:

   1 Start run by sampling state $k$ from $t_k^2 \equiv \langle \phi_k | \Psi_{\rm T} \rangle^2$.
   2 At each step sample $i$ from $\tilde{P}_{ik}$ and average $E_{\rm L}(i) W_i$ and $W_i$ ($W_i = \prod_{k=2}^{i} w_{k,k-1}$) Now $i, k$ indexes sum over MC steps, not states.

# Projector vs Path Integral MC Methods

$$E_0 = \underbrace{\frac{\sum_{ik}^{N_{\rm st}} E_{\rm L}(i)\; t_i\; P_{ik}\; t_k}{\sum_{ik}^{N_{\rm st}} t_i\; P_{ik}\; t_k}}_{\text{not importance sampled}} = \underbrace{\frac{\sum_{ik}^{N_{\rm st}} E_{\rm L}(i)\; \tilde{P}_{ik}\; t_k^2}{\sum_{ik}^{N_{\rm st}} \tilde{P}_{ik}\; t_k^2}}_{\text{importance sampled}}$$

2 ways to turn these into MC expressions:

1. Projector MC: previous viewgraph
2. Reptation MC/PIGS: $\because$ we have explicit expressions for $t_k = \langle \phi_k | \Psi_{\rm T} \rangle$ and $P_{ik} = \langle \phi_i | \hat{P} | \phi_k \rangle$ so use Metropolis-Hastings to sample entire path $t_i P_{ik} t_k = \tilde{P}_{ik}\; t_k^2$. Have to store and evolve the entire path. "Walker" is now $3NL$-dimensional.

   By integrating over all other beads, we can show (provided $\beta$ is large enough) that:

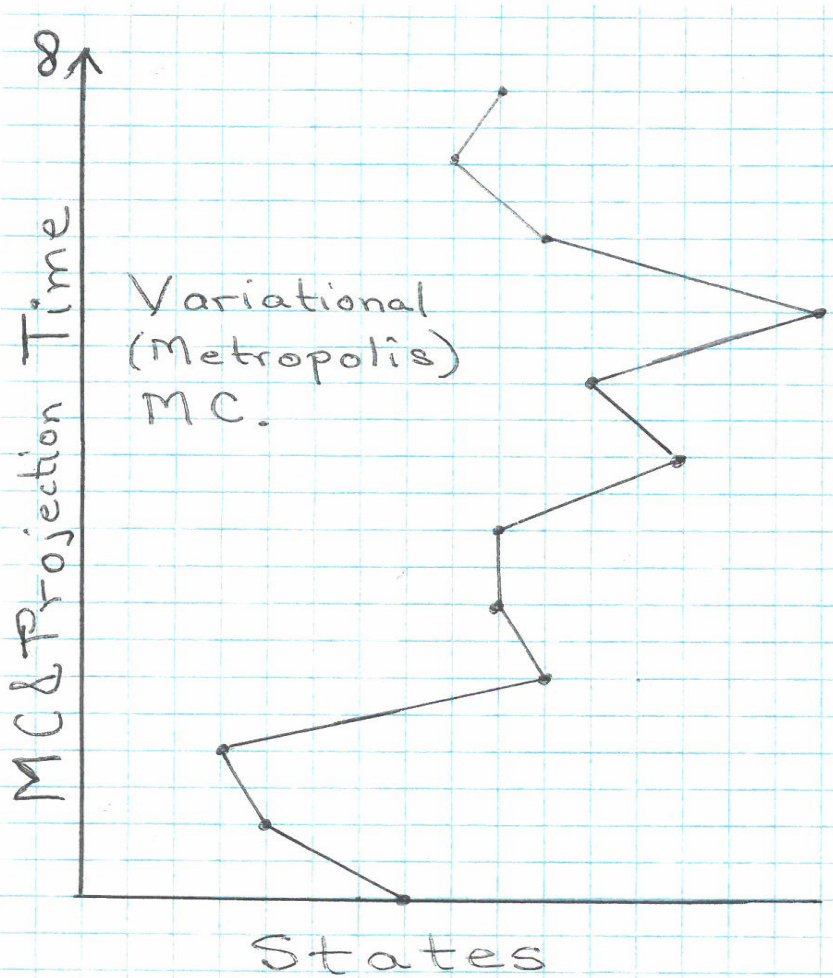
   1 Beads at the ends sample $\Psi_{\rm T}\Psi_0$, so we can compute $E_0$ by accumulating $E_{\rm L}$ there.

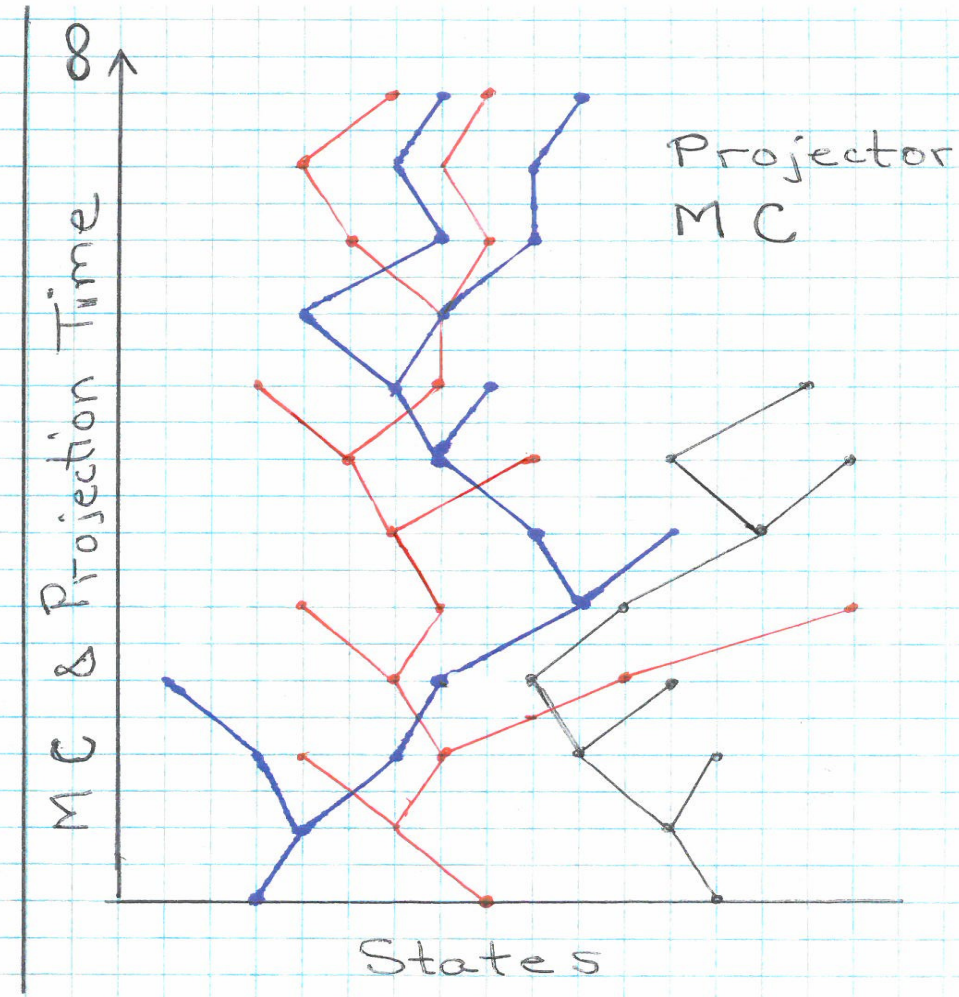
   2 Beads near the center sample $|\Psi_0|^2$, so can efficiently compute expec. values of observables diagonal in space of walk there.

$$\langle \hat{O} \rangle = \frac{\langle \Psi_{\rm T} | \hat{P}(\beta)\, \hat{O}\, \hat{P}(\beta) | \Psi_{\rm T} \rangle}{\langle \Psi_{\rm T} | \hat{P}(2\beta) | \Psi_{\rm T} \rangle} = \frac{\sum_{jik}^{N_{\rm st}} t_j P_{ji} O_{ii} P_{ik} t_k}{\sum_{jik}^{N_{\rm st}} t_j P_{ji} P_{ik} t_k}$$

In reality, each $P$ is a convolution of short-time $P$'s

# Schematic of VMC and PMC



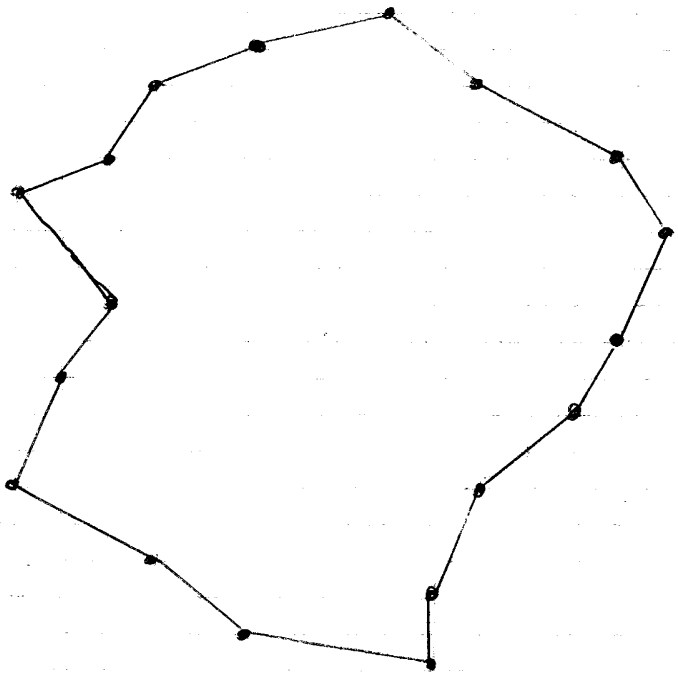



# Schematic of PIMC and Reptation MC

## Path-Integral M.C.

Average path length $\propto$ projection time/inverse $T$

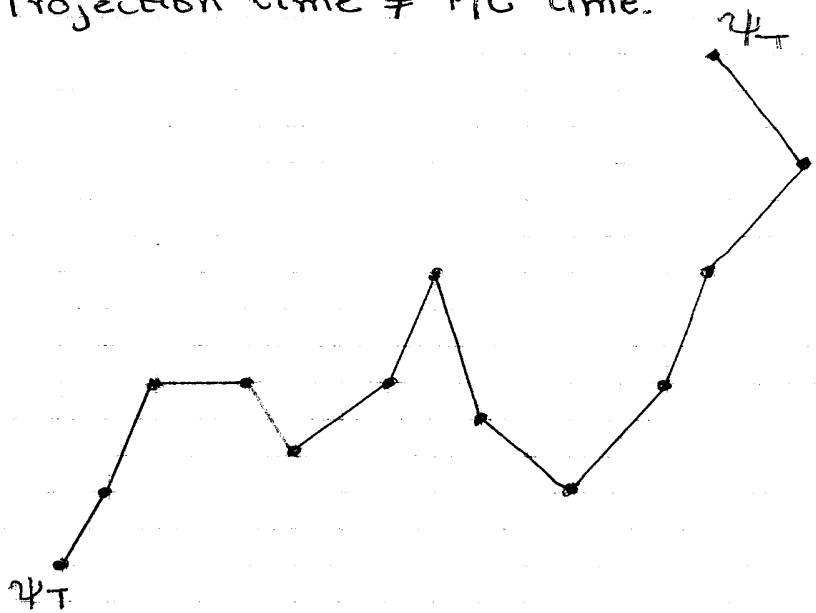
Projection time $\neq$ M.C. time



## Reptation M.C.

Average path length $\propto$ projection time

Projection time $\neq$ 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.
- Samples $\Psi_{\rm G}\Psi_0$, so requires "forward walking" to compute expectation values of observables that do not commute with $\hat{H}$.

## Reptation MC (RMC), Path-integral ground state (PIGS)

- No weights needed, but have to store and evolve the entire path and can be expensive to get to zero temperature (need long reptile).
- Samples $\Psi_{\rm G}\Psi_0$, at ends of reptile, so compute $\langle E \rangle$ by averaging $E_{\rm L}$ there.
- Samples $|\Psi_0|^2$ in the middle of the reptile, so compute expectation values of operators that do not commute with $\hat{H}$ but are diagonal in the space of the walk there.

## Path integral MC (PIMC)

- Used for finite-temperatures; very expensive for $T \to 0$ since it does not take advantage of $\Psi_{\rm T}$. No weights needed, but have to store and evolve the entire closed path.

# Taxonomy of Projector Monte Carlo Methods

The amplitudes of $\Psi_0$ in the chosen basis are obtained by using a "Projector", $\hat{P}$, that is a function of the Hamiltonian, $\hat{H}$, and has $\Psi_0$ 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. $\tau$) | $\mathbf{r}$ | $1^{st}$ |
| LRDMC (Sorella, Casula) | $e^{\tau(E_T\hat{\mathbf{1}}-\hat{H})}$ (samp. $\tau$) | $\mathbf{r}_i$ | $1^{st}$ |
| 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 $\phi_i^{\mathrm{orthog}}$, 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**

# Diffusion Monte Carlo

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

The imaginary-time evolved state $|\Psi(t)\rangle$ is

$$|\Psi(t)\rangle = e^{-\hat{H}t}|\Psi(0)\rangle = \sum_i e^{-\hat{H}t}|\Psi_i\rangle\langle\Psi_i|\Psi(0)\rangle = \sum_i e^{-E_i t}|\Psi_i\rangle\langle\Psi_i|\Psi(0)\rangle$$

$$= \sum_i c_i(t)\,|\Psi_i\rangle,$$

$c_i(t) = c_i(0)e^{-E_i t}$, $c_i(0) = \langle\Psi_i|\Psi(0)\rangle$. Expon. decay of excited states

$$E_{\rm mix} = \lim_{t\to\infty} \frac{\langle\Psi_{\rm T}|\hat{H}e^{-\hat{H}t}|\Psi_{\rm T}\rangle}{\langle\Psi_{\rm T}|e^{-\hat{H}t}|\Psi_{\rm T}\rangle} = \frac{\langle\Psi_{\rm T}|\hat{H}|\Psi_0\rangle}{\langle\Psi_{\rm T}|\Psi_0\rangle} = E_0\frac{\langle\Psi_{\rm T}|\Psi_0\rangle}{\langle\Psi_{\rm T}|\Psi_0\rangle} = E_0$$

The mixed estimator for the energy is exact (for Bosonic ground states) but expectation values of operators that do not commute with the energy are not.

# 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., $t = M\tau$.

$$G(\mathbf{R}', \mathbf{R}, t) = \langle \mathbf{R}' | \left( e^{-t\hat{H}/M} \right)^M | \mathbf{R} \rangle$$

$$= \int d\mathbf{R}_1 \cdots d\mathbf{R}_{M-1} \, \langle \mathbf{R}' | e^{-\tau\hat{H}} | \mathbf{R}_{M-1} \rangle \, \langle \mathbf{R}_{M-1} | e^{-\tau\hat{H}} | \mathbf{R}_{M-2} \rangle \, \cdots \, \langle \mathbf{R}_1 | e^{-\tau\hat{H}} | \mathbf{R} \rangle$$

$$e^{-\tau\hat{H}} \approx e^{-\tau\hat{T}} e^{-\tau\hat{V}} + \mathcal{O}(\tau^2) \qquad \text{Trotter breakup}$$

$$e^{-\tau\hat{H}} \approx e^{-\tau\hat{V}/2} e^{-\tau\hat{T}} e^{-\tau\hat{V}/2} + \mathcal{O}(\tau^3)$$

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

$$G(\mathbf{R}',\mathbf{R},\tau) \;=\; \langle \mathbf{R}'|e^{\tau(E_T-\hat{H})}|\mathbf{R}\rangle \approx \frac{1}{(2\pi\tau)^{3N/2}} e^{\left[-\frac{(\mathbf{R}'-\mathbf{R})^2}{2\tau}+\left\{E_T-\frac{(\mathcal{V}(\mathbf{R}')+\mathcal{V}(\mathbf{R}))}{2}\right\}\tau\right]}$$

# Diffusion Monte Carlo – Short-time Green's function

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

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

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

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

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 Schrödinger equation

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

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

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

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

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

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

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

# 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 $\Psi_{\rm T}$<br>$E_{\rm L} = E_0$ for $\Psi_0$ | $V \sim \frac{1}{R_\perp}$<br>for both $\Psi_{\rm T}$ and $\Psi_0$ |
| e-n and e-e coincidences | $E_{\rm L} \sim \frac{1}{x}$ if cusps not imposed<br>$E_{\rm L}$ finite if cusps are imposed<br>$E_{\rm L} = E_0$ for $\Psi_0$ | $V$ has a discontinuity<br>for both $\Psi_{\rm T}$ and $\Psi_0$ |

Modify Green's function, by approximately integrating $\mathbf{V}$ over $\tau$, making linear approximation for $\Psi_{\rm T}$ near node, 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 $\tau$ is:

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

# Combining with Metropolis to reduce time-step error

Reynolds, Ceperley, Alder, Lester, JCP 1982

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

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

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

We can sample $|\Psi_{\rm T}|^2$ 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 $\tau$ to $\tau_{\rm eff}$.
If accept/reject is done after moving all electrons, reweight only accepted moves.
If accept/reject is done after each 1-electron move, then $\tau_{\rm eff} = \tau\frac{R^2_{\rm accep}}{R^2_{\rm prop}}$.

# 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.

# 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 $1^1$S ground state. What are its nodes?

It has none!

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

$r_1 = r_2$ ($\mathbf{r}_1 = \mathbf{r}_2$ is co-dimension 2 from $r_1 = r_2$.)

Proof: Suppose $r_1 = r_2$.

If we rotate by $180^\circ$ about line joining nucleus to the midpoint of the 2 electrons, $\Psi \rightarrow \Psi$.

If we exchange the electrons, $\Psi \rightarrow -\Psi$.

So, $\Psi = 0$ when $r_1 = r_2$.

# 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?

# 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

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

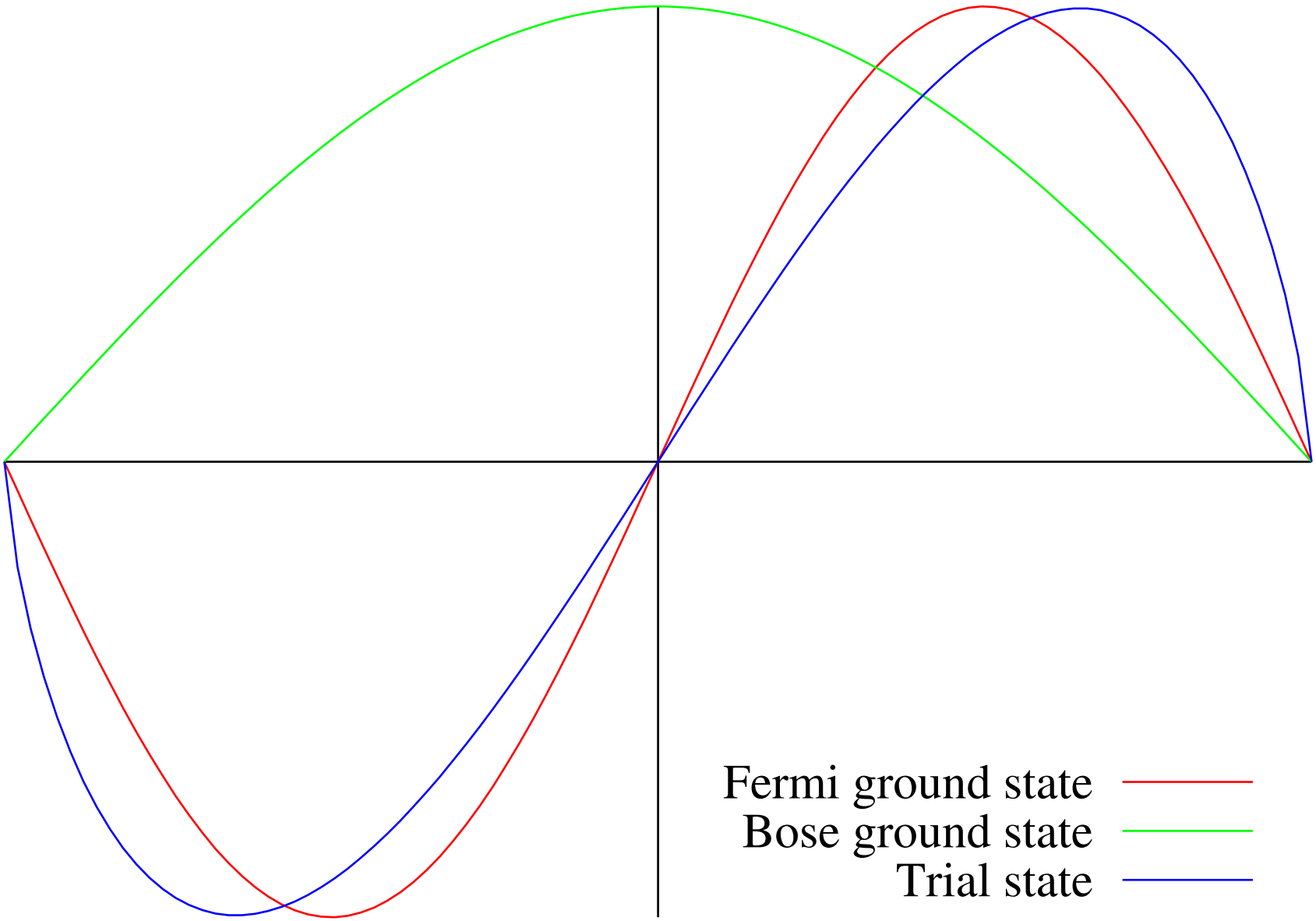
$\langle \mathbf{R}|\hat{P}(\tau)|\mathbf{R}'\rangle \approx \frac{e^{\frac{-\left(\mathbf{R}-\mathbf{R}'\right)^2}{2\tau}+\left(E_T-\frac{\mathcal{V}(\mathbf{R})+\mathcal{V}(\mathbf{R}')}{2}\right)\tau}}{(2\pi\tau)^{3N/2}}$ is nonnegative.

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.

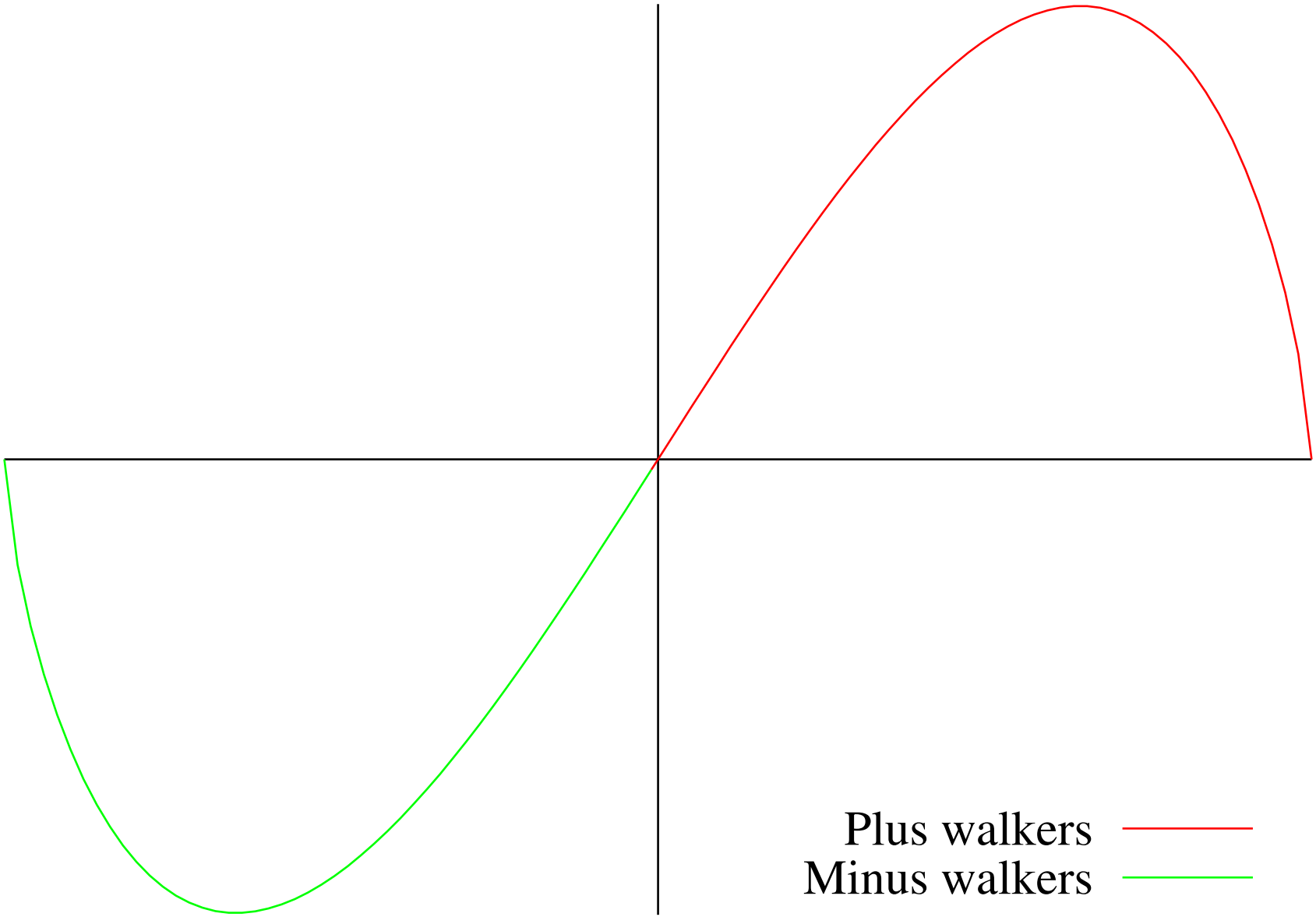
Fixed-node approximation

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 $\Psi_{\mathrm{T}}(\mathbf{R})$.
This gives an upper bound to the energy and becomes exact in the limit that $\Psi_{\mathrm{T}}$ has the same nodes as $\Psi_0$.

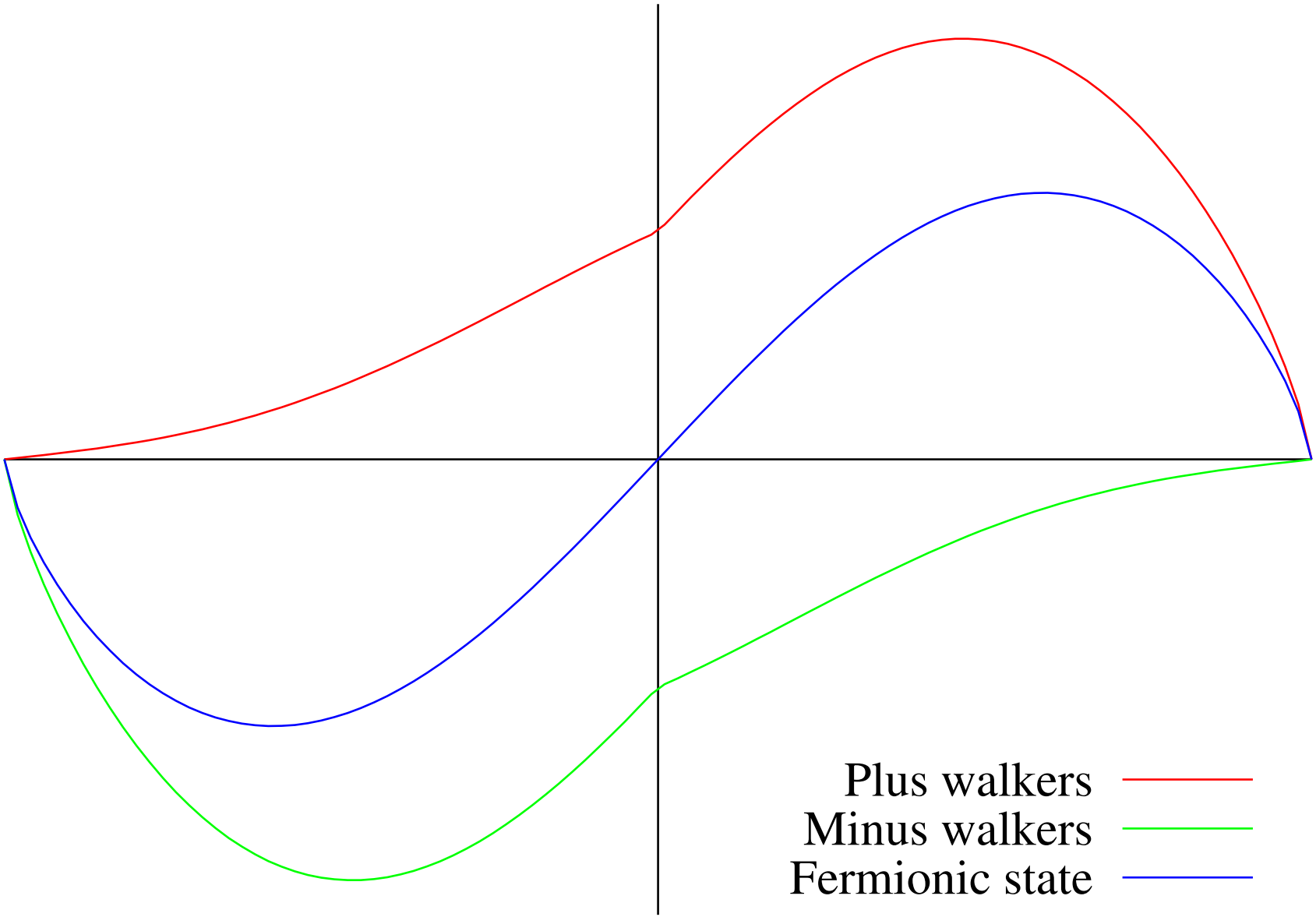
# Sign Problem in $1^{st}$ Quantization and R space



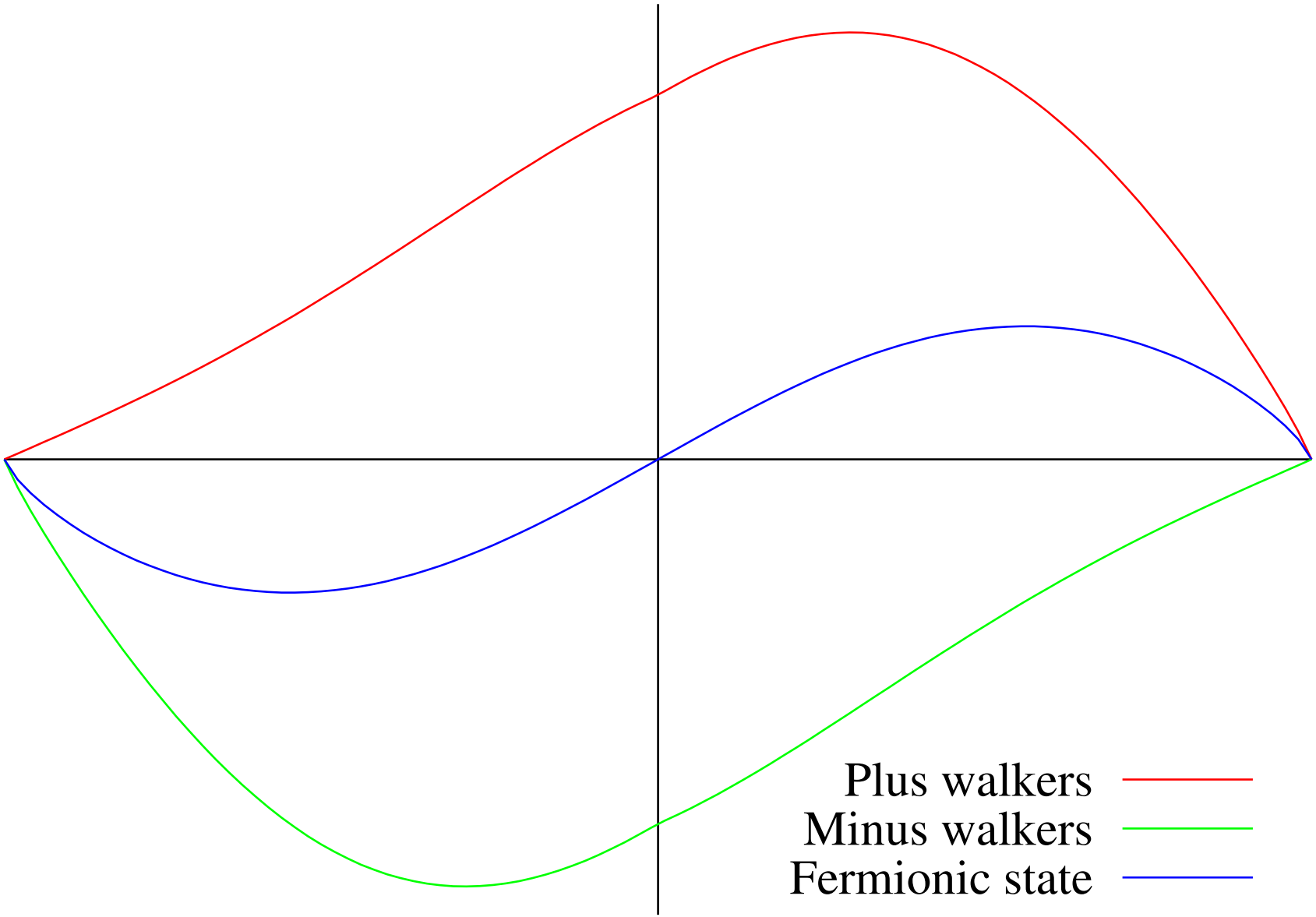
# Sign Problem in $1^{st}$ Quantization and R space



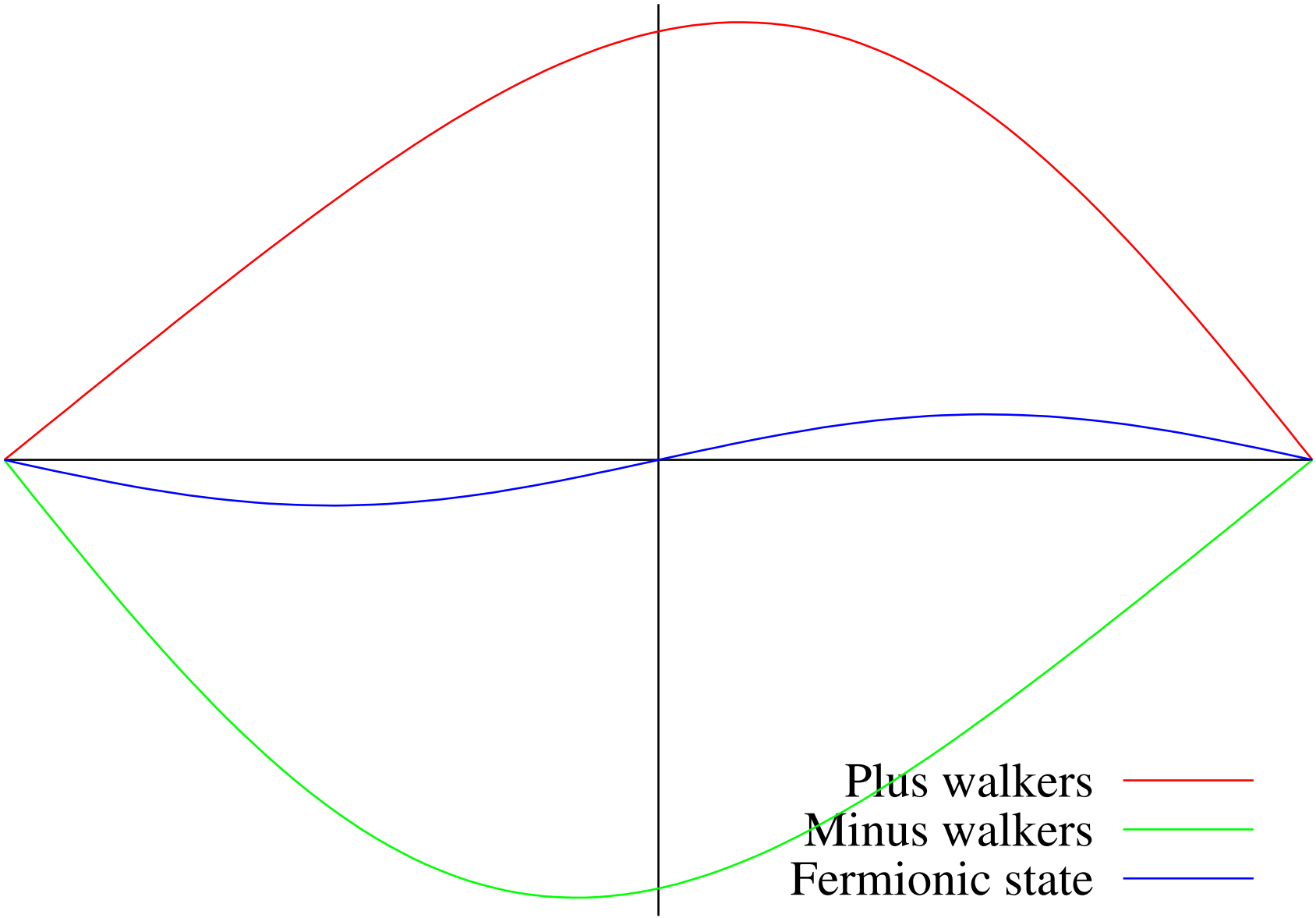
# Sign Problem in $1^{st}$ Quantization and R space



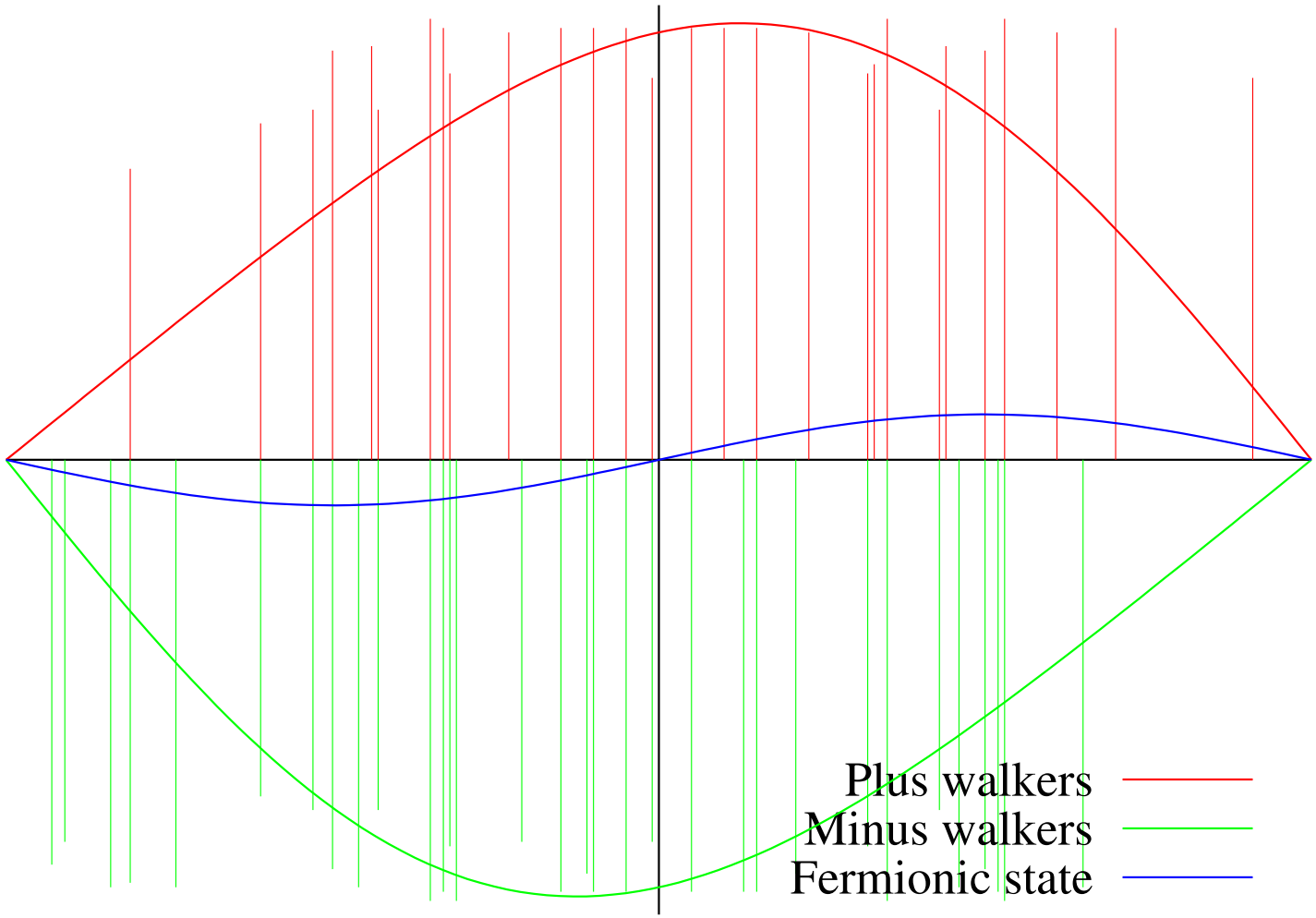
# Sign Problem in $1^{st}$ Quantization and R space



# 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, $\Psi$ 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.

# Reweighting, Branching, Population Control

1. Since columns of $G(\mathbf{R}', \mathbf{R}, t)$ are not normalized to unity, the "walkers" must get reweighted.
2. 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, we have a population of walkers, and "branching" to ensure that all walkers have approximately the same weight.
3. If left uncontrolled, the size of this population will fluctuate a lot. Hence we exercise "population control". One way to do this is to adjust $E_{\rm T}$ at each step according to:

$$E_{\rm T} \;=\; E_{\rm est} + \frac{\log(W_{\rm target}/W_{\rm gen})}{N_{\rm gen}} \qquad (1)$$

   where $E_{\rm est}$ is the best current estimator for the energy, $W_{\rm target}$ is the target weight of a generation, $W_{\rm gen}$ is the weight of the current generation, and $N_{\rm gen}$ controls the strength of the population control.
4. Population control leads to a "population control error" that is inversely proportional to the target population size. So, one way around it is to use a sufficiently large population. The better way is to correct for population control error – it is straightforward to do and requires no additional computer time.

# 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:

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

# 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.

Unlike before $E_{\rm T}$ is now kept fixed.)

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

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

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

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., $p = e^{-1/T_{\rm p}}$.

# 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

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

$$\text{and} \quad \tilde{P}_i(t,p) \;=\; \cdots f(t-2)^{p^3}\; f(t-1)^{p^2}\; f(t)^{p^1} \;=\; \prod_{n=0}^{t-1} f(t-n)^{p^{n+1}}$$

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

$$P_i(t,p) = P_i(t-1,p)^p\; f(t), \qquad \tilde{P}_i(t,p) = \left(\tilde{P}_i(t-1,p)\; f(t)\right)^p$$

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

$$E_{\rm gr} \;=\; E_{\rm T} + \frac{1}{\tau}\left(1 - \frac{\sum_{t=1}^{N_{\rm gen}-1} P_i(t+1,p)\; \sum_i^{N_w(t+1)} w_i(t+1)}{\sum_{t=1}^{N_{\rm gen}-1} \tilde{P}_i(t,p)\; \sum_i^{N_w(t)} w_i(t)}\right), \qquad \text{for the linear projector}$$

$$E_{\rm gr} \;=\; E_{\rm T} - \frac{1}{\tau}\log\left(\frac{\sum_{t=1}^{N_{\rm gen}-1} P_i(t+1,p)\; \sum_i^{N_w(t+1)} w_i(t+1)}{\sum_{t=1}^{N_{\rm gen}-1} \tilde{P}_i(t,p)\; \sum_i^{N_w(t)} w_i(t)}\right), \qquad \text{for the exponential projector}$$

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

If $[A, H] \neq 0$, we want to compute the pure (not mixed) expectation value

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

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

$$\tilde{T}(\mathbf{R}^{'},\mathbf{R}) \;=\; G_{\rm dif}(\mathbf{R}^{'},\mathbf{R}^{''})\; G_{\rm drift}(\mathbf{R}^{''},\mathbf{R}) \;=\; \frac{1}{(2\pi\tau)^{3N/2}} e^{-\frac{(\mathbf{R}^{'}-\mathbf{R}-\mathbf{V}\tau)^2}{2\tau}}$$

$$w(\mathbf{R}^{'},\mathbf{R}) \;=\; e^{\left\{ E_{\rm T}-\frac{(E_{\rm L}(\mathbf{R}^{'})+E_{\rm L}(\mathbf{R}))}{2} \right\}\tau}$$

For discrete state space and sparse $H$, define

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

$$w(\mathbf{R}^{'},\mathbf{R}) \;=\; w(\mathbf{R}) \;=\; \sum_{\mathbf{R}^{''}} \tilde{G}(\mathbf{R}^{''},\mathbf{R})$$

# 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 $i$, $A_{\mathrm{L},i} = \sum_j g_j A j i / g_i$ 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

1) Extrapolated estimator

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

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

Forward or Future Walking

$$\langle A \rangle = \frac{\langle \Psi_T | G^p(\tau) \hat{A} G^{p'}(\tau) | \Psi_T \rangle}{\langle \Psi_T | G^{p+p'}(\tau) | \Psi_T \rangle}$$ need projectors on both sides of $\hat{A}$

$$= \frac{\sum_{\mathbf{R}_{p+p'} \cdots \mathbf{R}_0} A(\mathbf{R}_{p'}) \left( \prod_{i=0}^{p+p'-1} \tilde{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) (\Psi_T(\mathbf{R}_0))^2}{\sum_{\mathbf{R}_{p+p'} \cdots \mathbf{R}_0} \left( \prod_{i=0}^{p+p'-1} \tilde{G}(\mathbf{R}_{i+1}, \mathbf{R}_i) \right) (\Psi_T(\mathbf{R}_0))^2}$$

$$= \frac{\sum_{t=T_{\rm eq}+1}^{T_{\rm eq}+T} A(\mathbf{R}_t) W_{t+p}}{\sum_{t=T_{\rm eq}+1}^{T_{\rm eq}+T} W_{t+p}}$$

$W_{t+p} = \prod_{i=0}^{p+p'-1} w(\mathbf{R}_{t+p-i}, \mathbf{R}_{t+p-i-1})$ (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 $\Psi_T$ has not been fully projected onto $\Psi_0$, 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!)

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

2) Forward/Future walking Liu, Kalos, and Chester, PRA (1974)

When using importance sampled propagator, walker at $\mathbf{R}$ has expected distribution a sufficiently long time $t$ later that excited states have died out: $f(\mathbf{R}^{'},t) = c_0 e^{(E_T - E_0)t} \Psi_T(\mathbf{R}^{'}) \Psi_0(\mathbf{R}^{'})$.

$$\delta(\mathbf{R}^{'} - \mathbf{R}) = \sum_i \Psi_i(\mathbf{R}^{'}) \Psi_i(\mathbf{R}) = \Psi_T(\mathbf{R}) \sum_i \Psi_i(\mathbf{R}^{'}) \frac{\Psi_i(\mathbf{R})}{\Psi_T(\mathbf{R}))} = \Psi_T(\mathbf{R}) \sum_i c_i \Psi_i(\mathbf{R}^{'})$$

$$\Longrightarrow \; c_0 = \frac{\Psi_0(\mathbf{R})}{\Psi_T(\mathbf{R})}$$

So, total weight of descendents is

$$\int d\mathbf{R}^{'} \; f(\mathbf{R}^{'},t) = \int d\mathbf{R}^{'} \; \frac{\Psi_0(\mathbf{R})}{\Psi_T(\mathbf{R})} e^{(E_T - E_0)t} \Psi_T(\mathbf{R}^{'}) \Psi_0(\mathbf{R}^{'}) = \frac{\Psi_0(\mathbf{R})}{\Psi_T(\mathbf{R})} e^{(E_T - E_0)t} \langle \Psi_T | \Psi_0 \rangle$$

So, to compute pure estimators, instead of using the current generation weight, use the sum of weight from all descendents $n$ generations later. The factors of $e^{(E_T - E_0)t} \langle \Psi_T | \Psi_0 \rangle$ in numerator and denominator of expectation values cancel, and, the factor $\frac{\Psi_0(\mathbf{R})}{\Psi_T(\mathbf{R})}$ converts the mixed estimator to the pure estimator.

Usual tradeoff: If $n$ is small, there is some residual bias since $\Psi_T$ has not been fully projected onto $\Psi_0$, whereas, if $n$ is large the fluctuations of the descendent weights increases the statistical noise.

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

Forward or Future Walking

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

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

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

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.