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 16×16 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.

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

Ψ_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 Ψ_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: $E_0 = \dfrac{\langle \Psi_0 | \hat{H} | \Psi_0 \rangle}{\langle \Psi_0 | \Psi_0 \rangle} = \dfrac{\langle \Psi_0 | \hat{H} | \Psi_{\rm T} \rangle}{\langle \Psi_0 | \Psi_{\rm T} \rangle}$

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

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

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

$$= \sum_{i}^{N_{\rm st}} \frac{e_i t_i}{\sum_{k}^{N_{\rm st}} e_k t_k} \, E_{\rm L}(i) \;=\; \frac{\left[\sum_{i}^{N_{\rm MC}} E_{\rm L}(i)\right]_{\Psi_{\rm T}\Psi_0}}{N_{\rm MC}} \;\xrightarrow{\Psi_G \neq \Psi_T}\; \frac{\left[\sum_{i}^{N_{\rm MC}} \left(\frac{t_i}{g_i}\right) E_{\rm L}(i)\right]_{\Psi_{\rm G}\Psi_0}}{\left[\sum_{k}^{N_{\rm MC}} \left(\frac{t_k}{g_k}\right)\right]_{\Psi_{\rm G}\Psi_0}}$$

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

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

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

e.g. FCIQMC is a PMC method where $\Psi_{\rm G} = \mathbf{1} \neq \Psi_{\rm T}$.

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.

† 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 π (inefficiently).

1930s First significant scientific application of MC: Enrico Fermi used it for neutron transport in fissile material.
Segre: "Fermi took great delight in astonishing his Roman colleagues with his "too-good-to-believe" predictions of experimental results."

1940s Monte Carlo named by Nicholas Metropolis and Stanislaw Ulam

1953 Algorithm for sampling any probability density
Metropolis, Rosenbluth, Rosenbluth, Teller and Teller
(generalized by Hastings in 1970)

1962,1974 First PMC calculations, Kalos, and, Kalos, Levesque, Verlet.

1965 First VMC calculations (of liquid He), Bill McMillan.

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!

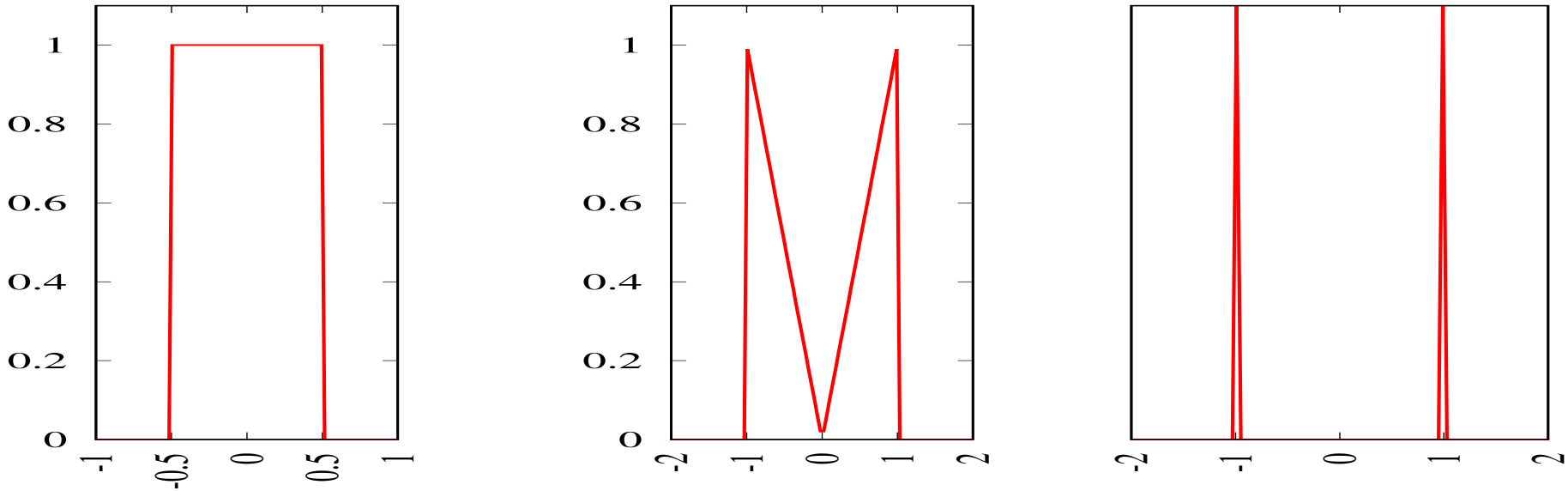
Here he is:

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, μ, and variance σ^2. The central limit theorem states that as the sample size N increases, the probability density of the sample average, $\bar{X}$, of these random variables approaches the normal distribution, $\sqrt{\frac{N}{2\pi\sigma^2}}e^{-(x-\mu)^2/(2\sigma^2/N)}$, with mean μ, and variance σ^2/N, irrespective of the original probability density function, e.g.:



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

(Weak) Law of Large Numbers

Cardano, Bernouli, Borel, Cantelli, Kolmogorov, Khinchin

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

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

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

Lorentzian

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

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

Lorentzian(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)

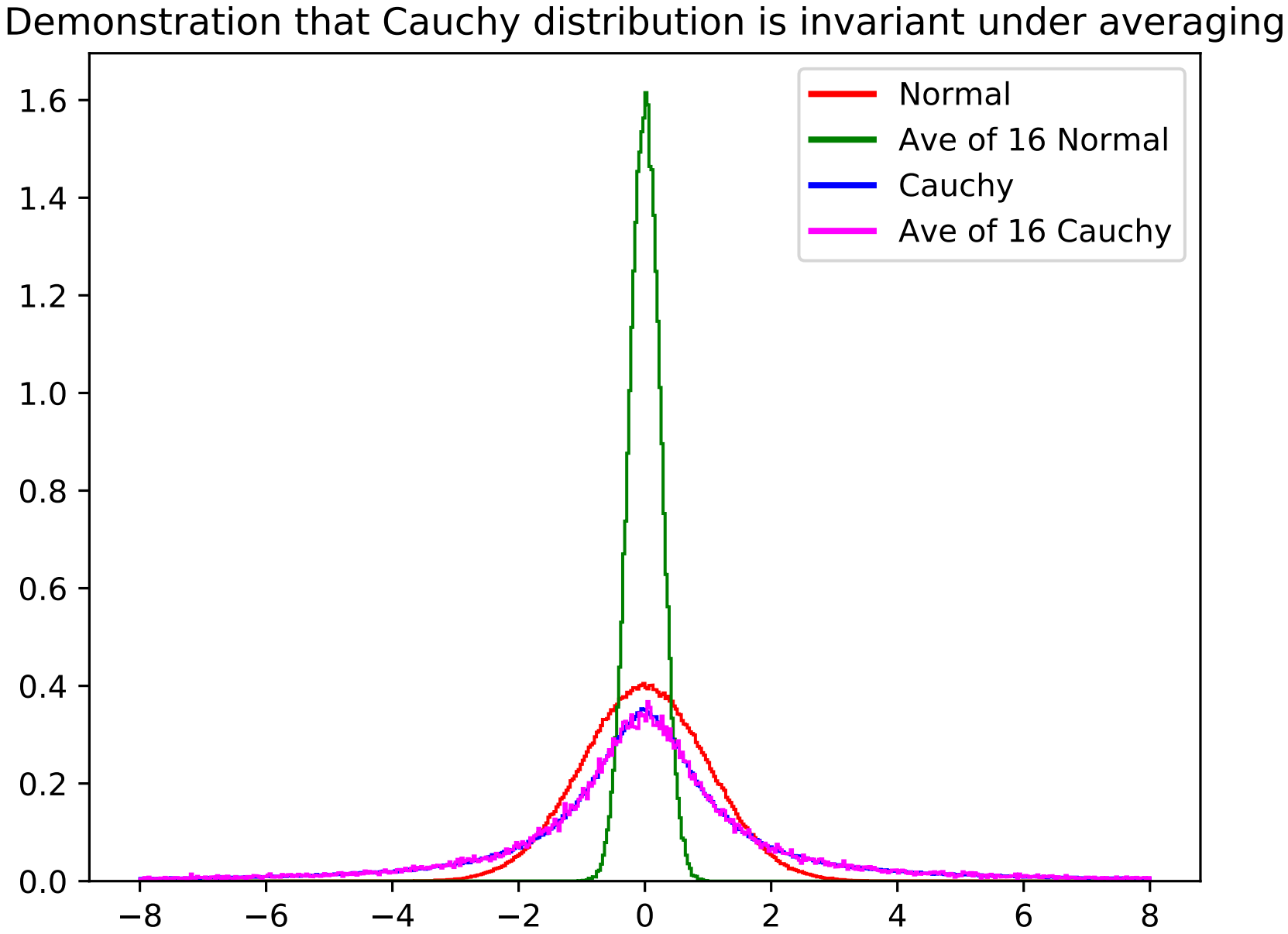
We are all brought up to believe that if we average numbers drawn from some probability density then the distribution of the averages will be narrower than the distribution of the individual values.

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 σ^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 μ and finite variance σ^2, we have much weaker bounds given by Chebychev's inequality:

The probability of a variable lying between $\mu - n\sigma$ and $\mu + n\sigma$ is $> 1 - 1/n^2$, as compared to erf$(n/\sqrt{2})$ for a Gaussian.

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

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

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 σ increases with sample size, N and tends to have upward jumps. So the estimated error of the sample mean, $\sigma_N = \sigma/\sqrt{N}$, goes down more slowly than $\frac{1}{\sqrt{N}}$, or even does not go down at all.

Monte Carlo versus Deterministic Integration methods

Deterministic Integration Methods:

Integration Error, ϵ, using $N_{\rm int}$ integration points:

1-dim Simpson rule: $\epsilon \leq c N_{\rm int}^{-4}$, (provided derivatives up to 4^{th} exist)

d-dim Simpson rule: $\epsilon \leq c N_{\rm int}^{-4/d}$, (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

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

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

Monte Carlo integration

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

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

(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

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

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

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

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)

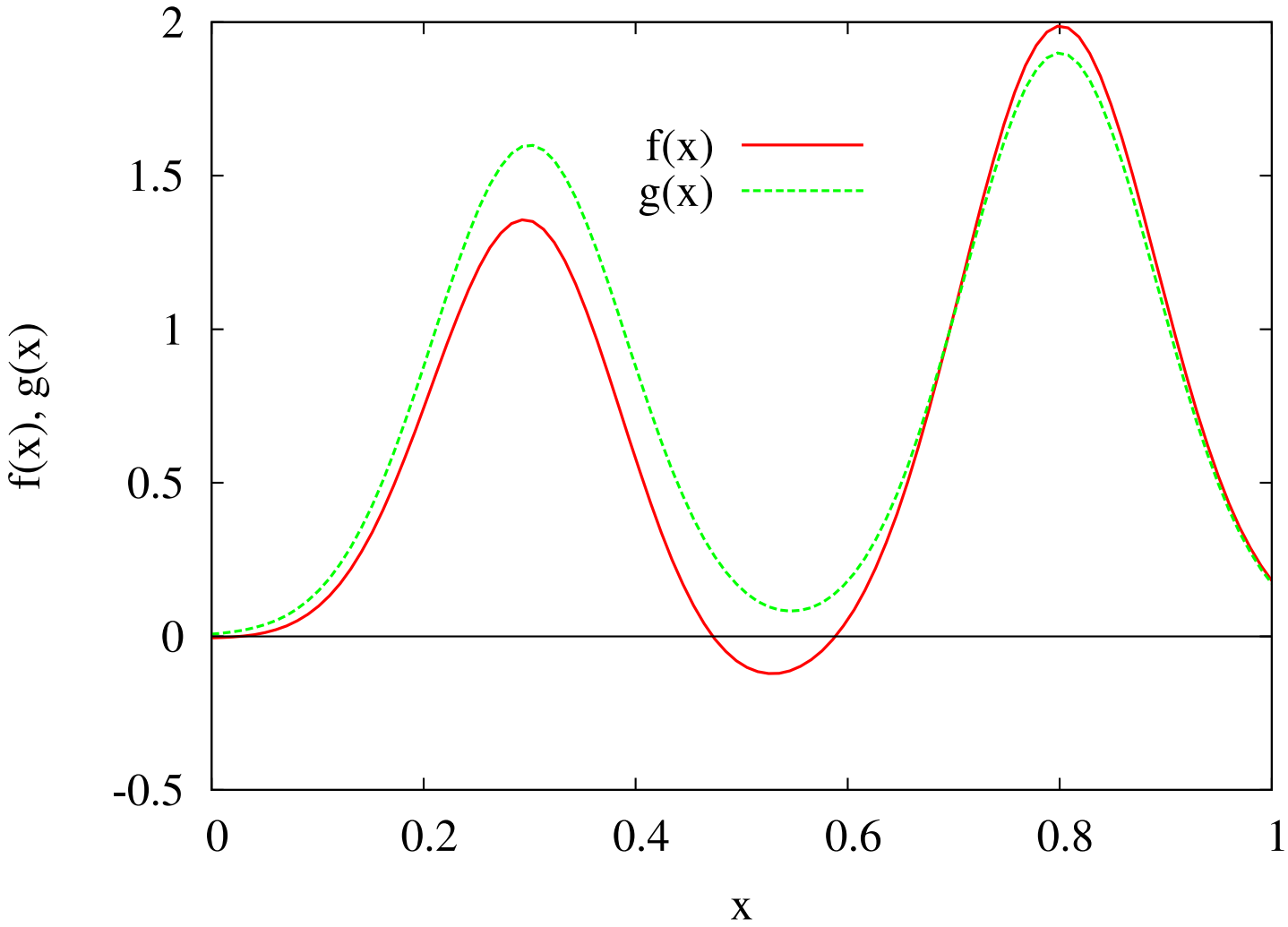
$$I = \int_V g(x)\frac{f(x)}{g(x)}dx = \overline{\left(\frac{f}{g}\right)} \pm \sqrt{\frac{\overline{\left(\frac{f}{g}\right)^2} - \overline{\left(\frac{f}{g}\right)}^2}{N-1}}$$

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!

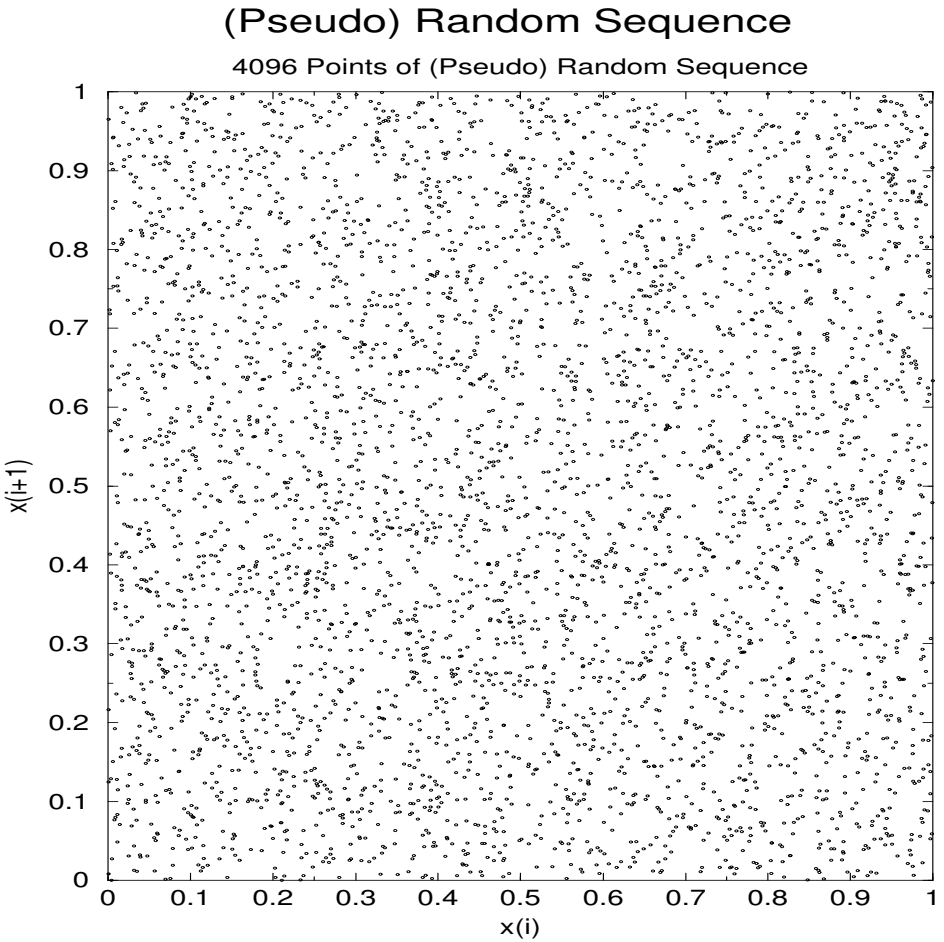
Illustration of Importance Sampling

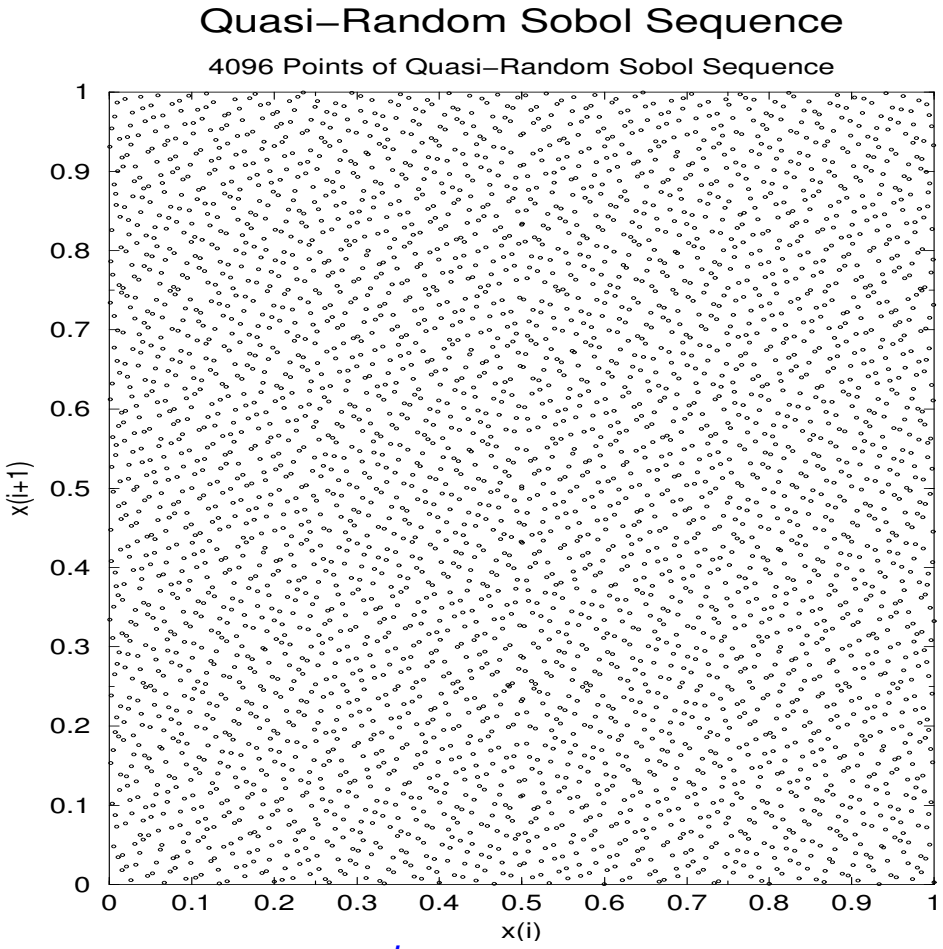
$f(x)$ is the function to be integrated. $g(x)$ is a function that is "similar" to $f(x)$ and has the required properties: a) $g(x) \geq 0$, b) $\int dx\, g(x) = 1$, and, c) we know how to sample it. $\int f(x)dx$ can be evaluated efficiently by sampling $g(x)$ and averaging $f(x)/g(x)$.



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 ρ, (discrete or continuous).

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

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

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 ξ, to get x sampled from desired probability density $f(x)$.

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

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

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

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

Examples of Transformation Method

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

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

Examples of Transformation Method

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

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

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

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

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

Examples of Transformation Method

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

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

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

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

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

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

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

Examples of Transformation Method

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

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

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

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

(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 σ^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 ± 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 $\mathcal{O}(N)$, $\mathcal{O}(\log_2(N))$, $\mathcal{O}(1)$?

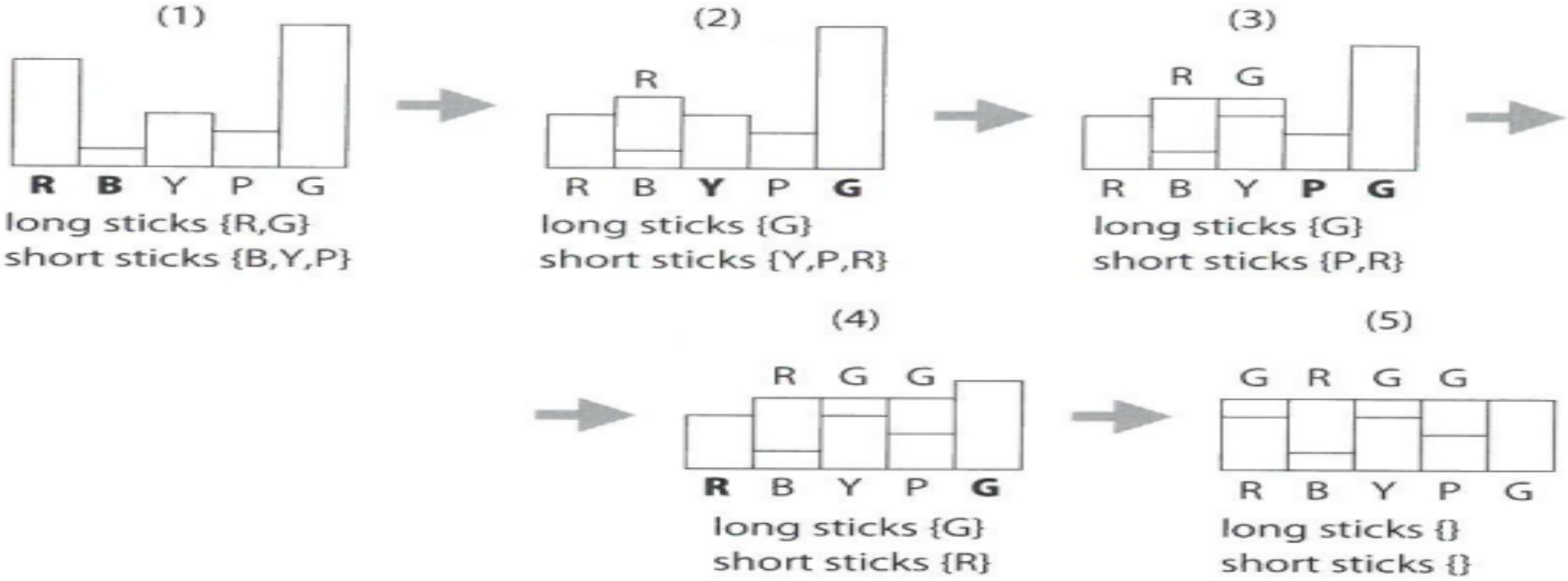
Straightforward $\mathcal{O}(\log_2(N))$ method with binary search:

1. Before starting sampling, construct array of cumulative probabilities.
2. Draw a random number, ξ, 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.

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

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

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,

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

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

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

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

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

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

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

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

Metropolis-Hastings Monte Carlo (cont)

Want that any initial density should evolve to ρ.

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

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

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

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

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

Metropolis-Hastings Monte Carlo (cont)

Construction of M

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

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

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

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

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

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

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

Metropolis-Hastings Monte Carlo (cont)

Choice of Acceptance Matrix A

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

Infinity of choices for A. Any function

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

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

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

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

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

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

Metropolis-Hastings Monte Carlo (cont)

Choice of Proposal Matrix T

So, the optimal choice for the acceptance matrix $A(\mathbf{R}_\mathrm{f}|\mathbf{R}_\mathrm{i})$ 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 Δ 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 ρ 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

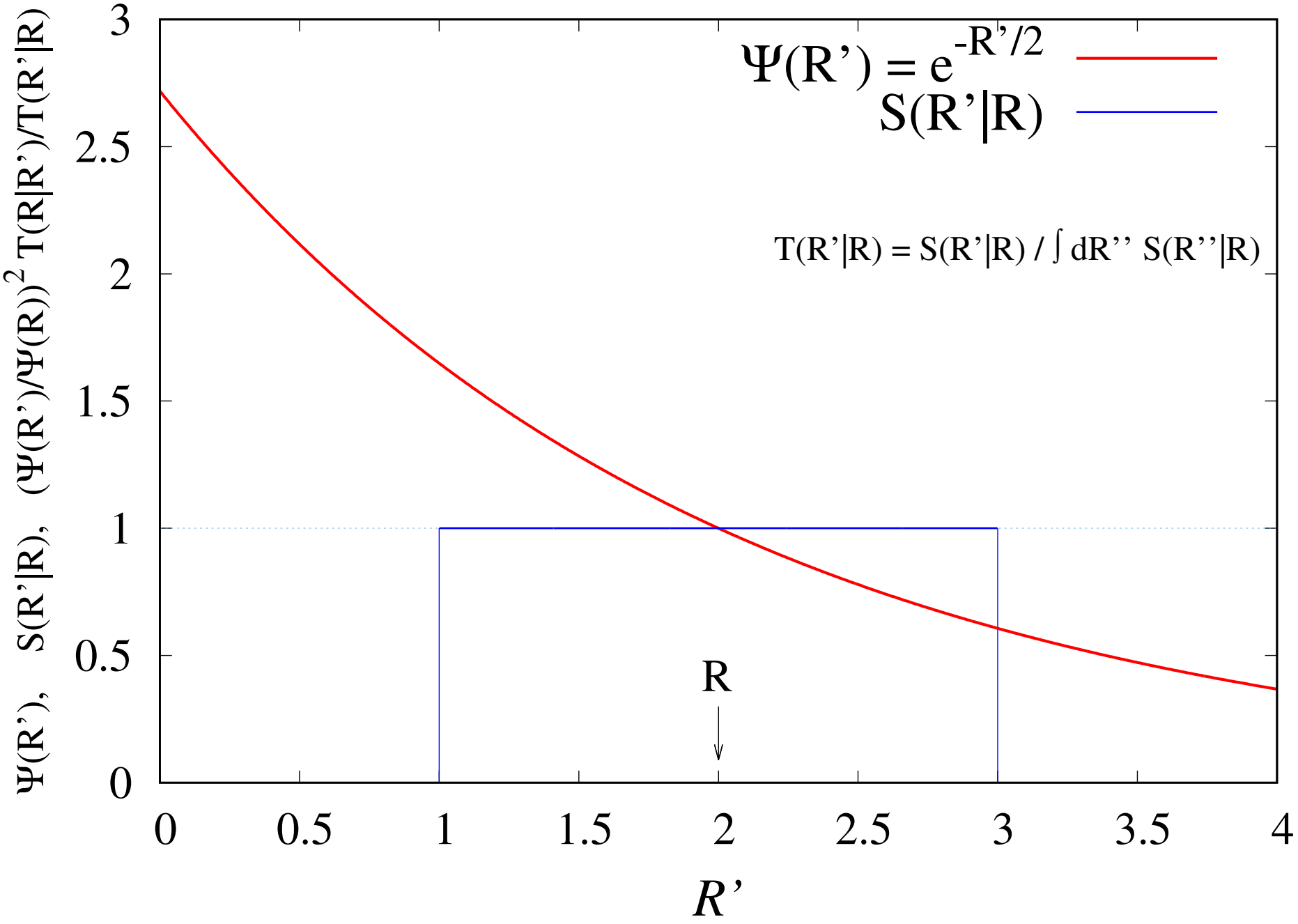
$$A(\mathbf{R}_f|\mathbf{R}_i) = \min\left\{1, \frac{|\Psi(\mathbf{R}_f)|^2\ T(\mathbf{R}_i|\mathbf{R}_f)}{|\Psi(\mathbf{R}_i)|^2\ T(\mathbf{R}_f|\mathbf{R}_i)}\right\}$$

we want

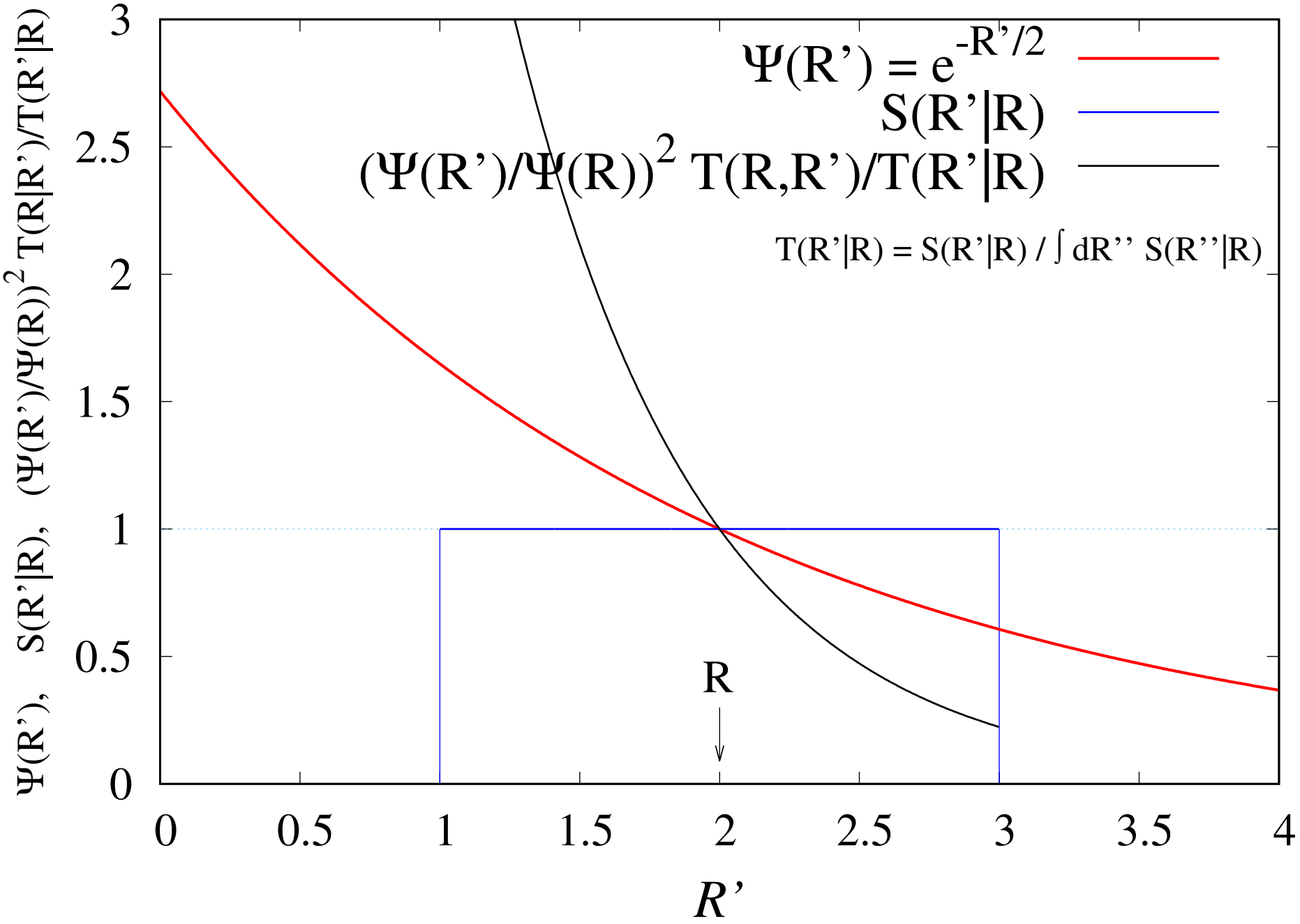
$$\frac{|\Psi(\mathbf{R}_f)|^2\ T(\mathbf{R}_i|\mathbf{R}_f)}{|\Psi(\mathbf{R}_i)|^2\ T(\mathbf{R}_f|\mathbf{R}_i)}$$

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

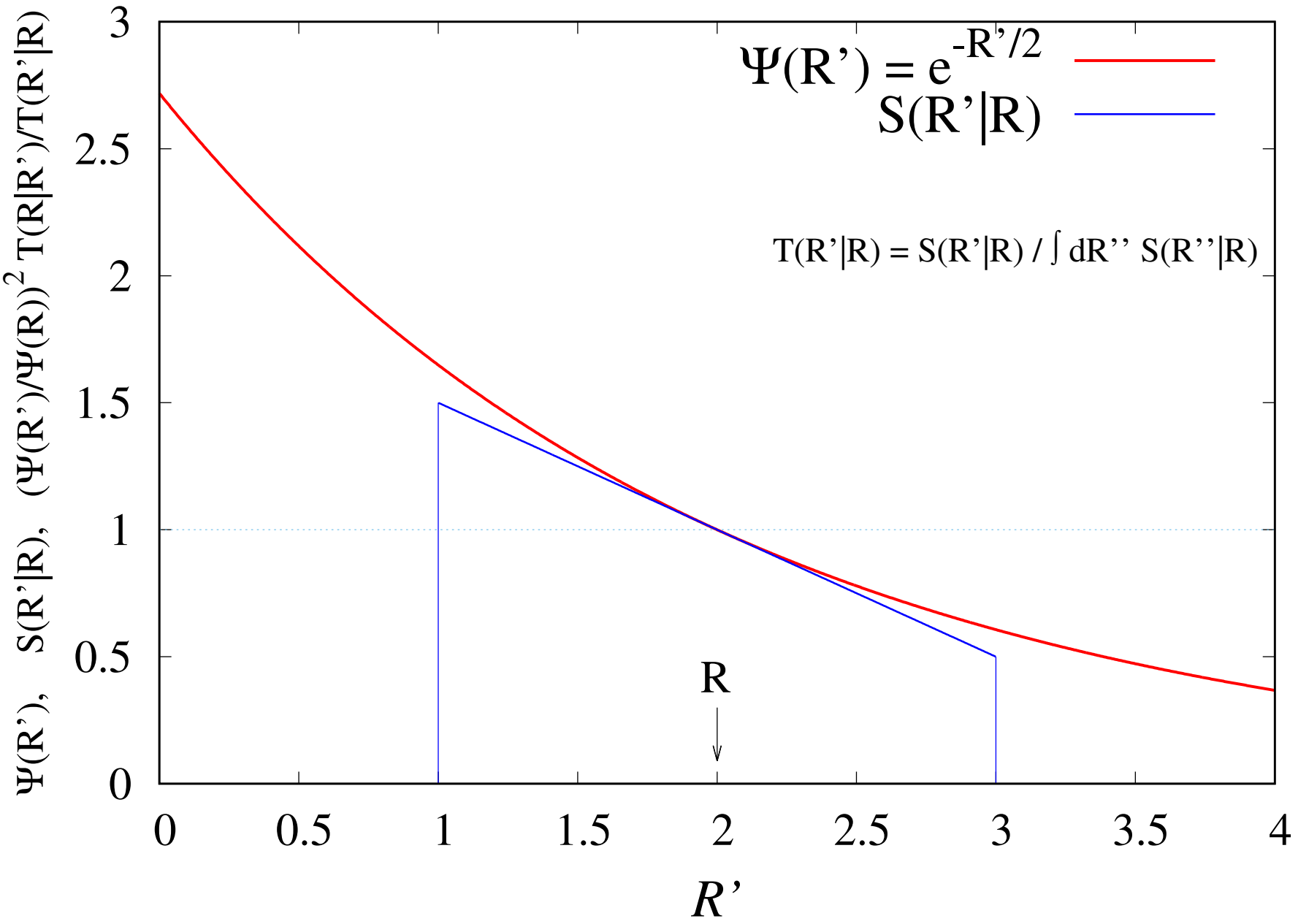
Symmetrical T in Metropolis



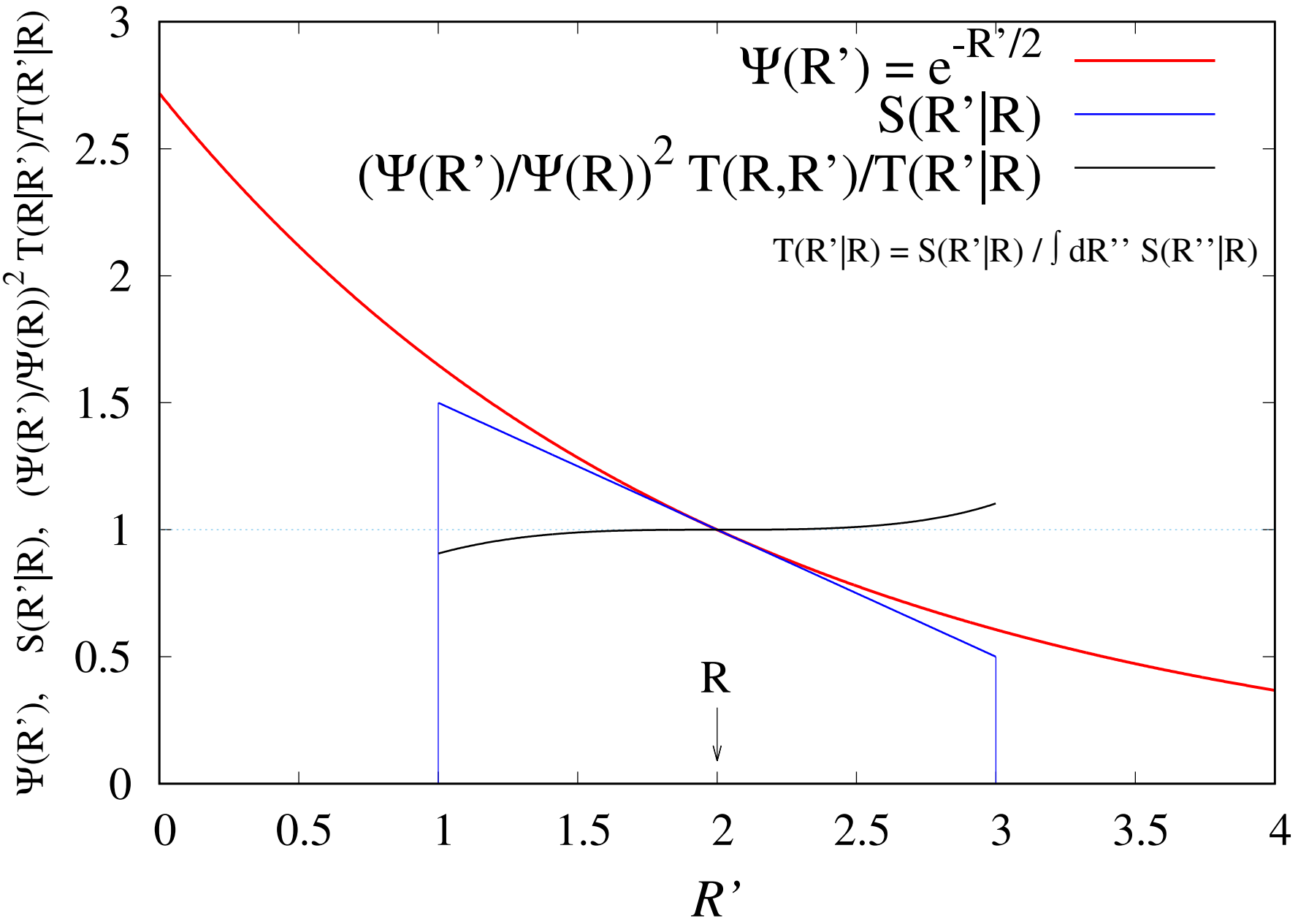
Symmetrical T in Metropolis



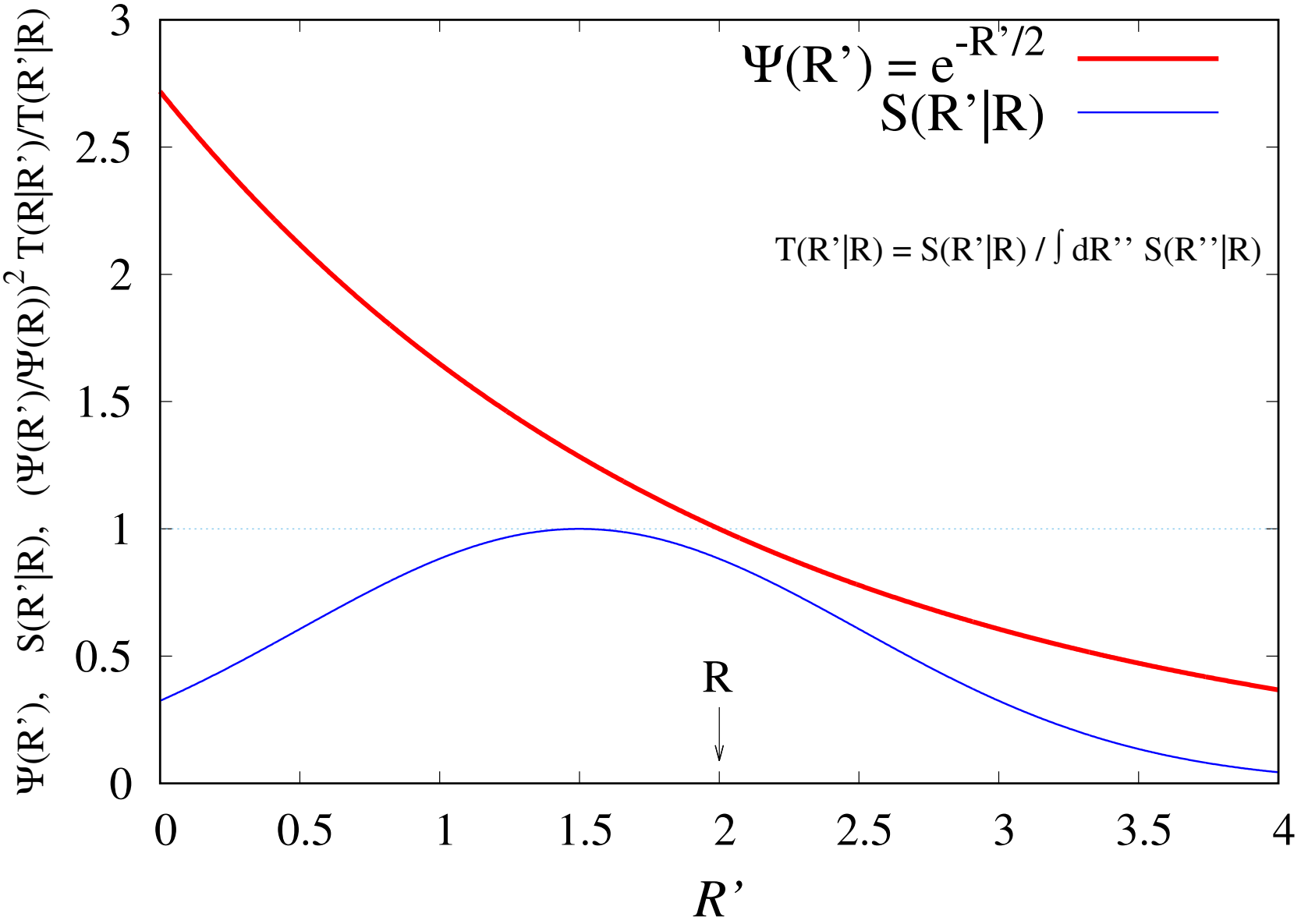
Non-symmetrical linear T in Metropolis-Hastings



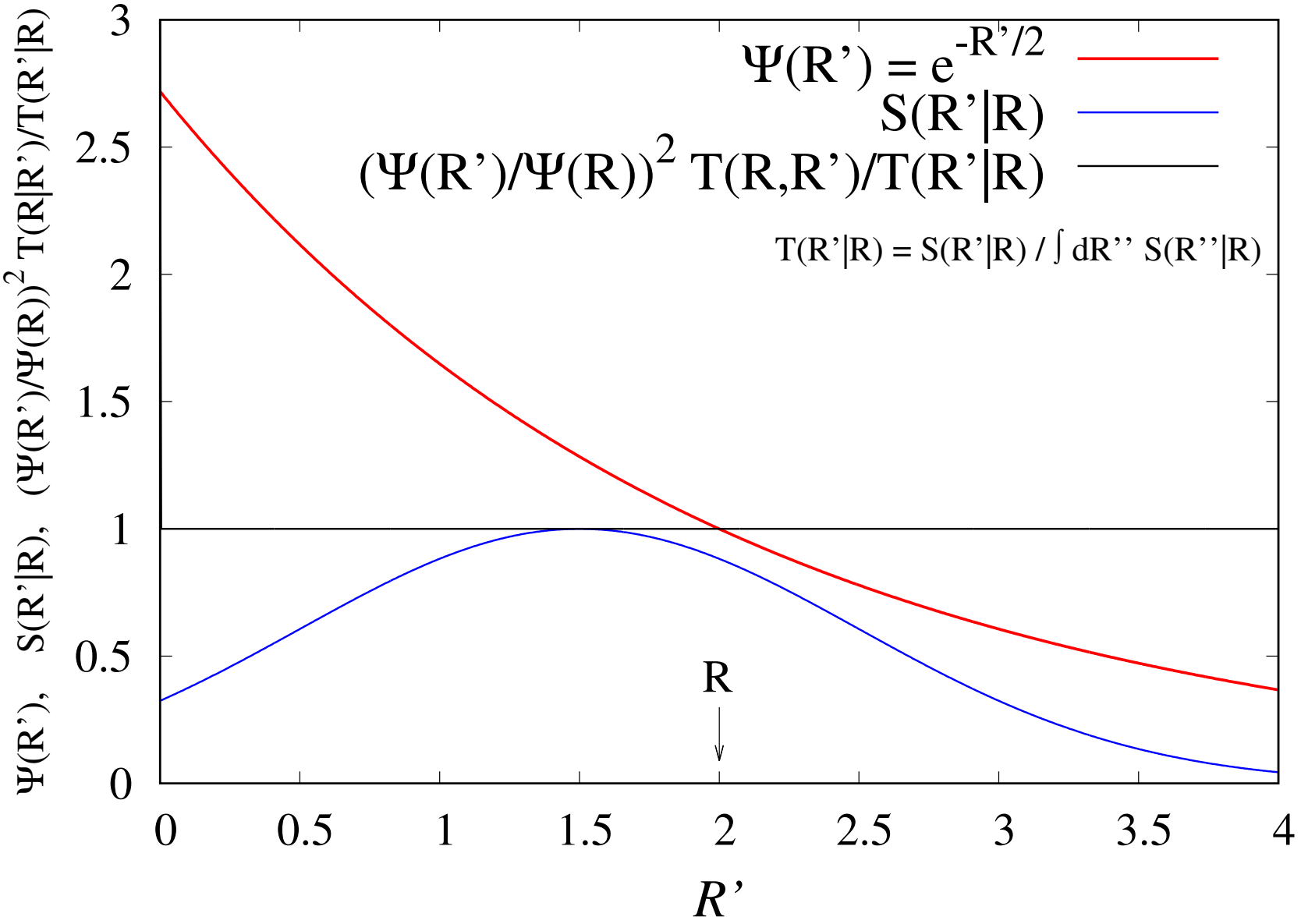
Non-symmetrical linear T in Metropolis-Hastings



Non-symmetrical drifted Gaussian T in Metropolis-Hastings

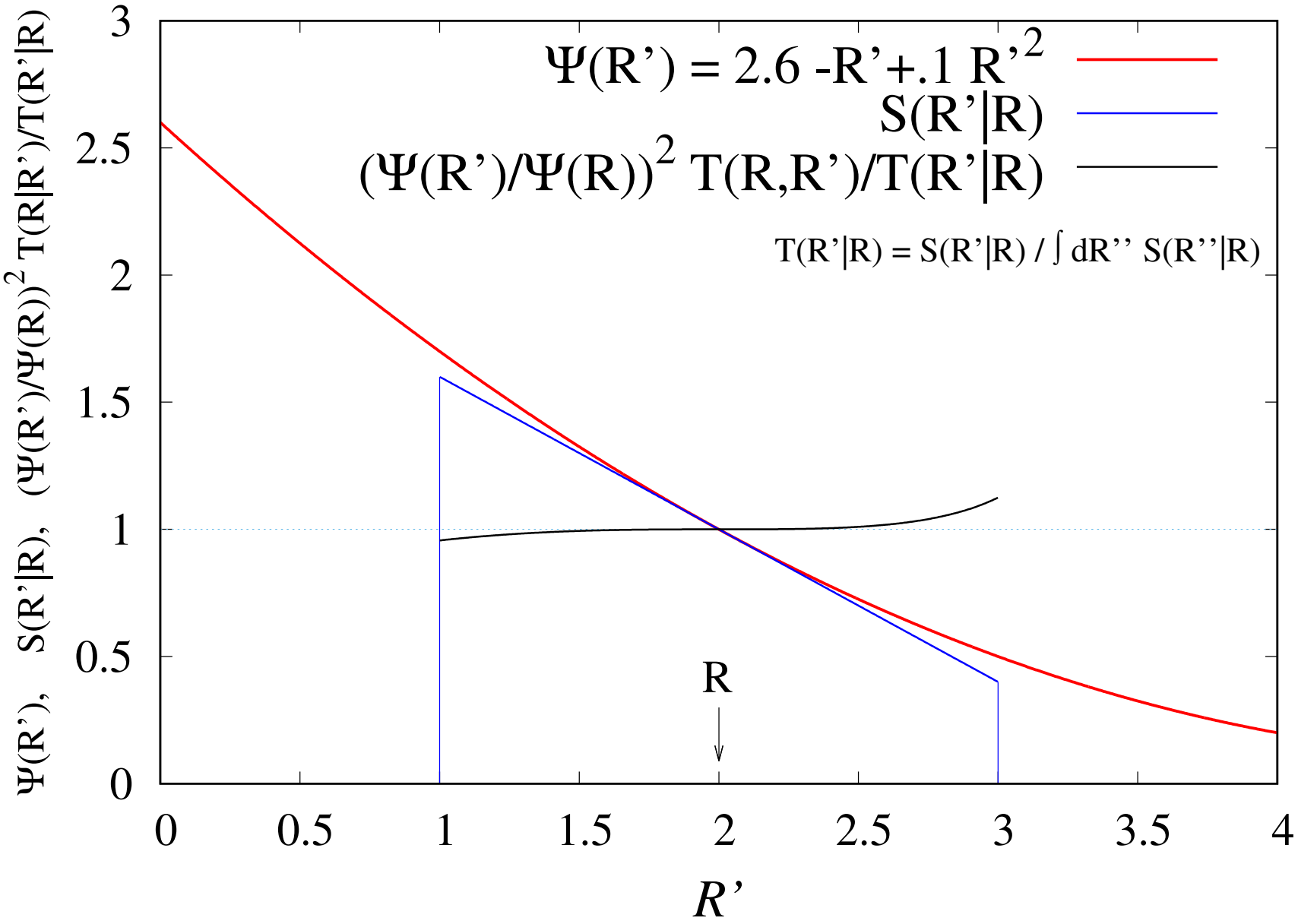


Non-symmetrical drifted Gaussian T in Metropolis-Hastings



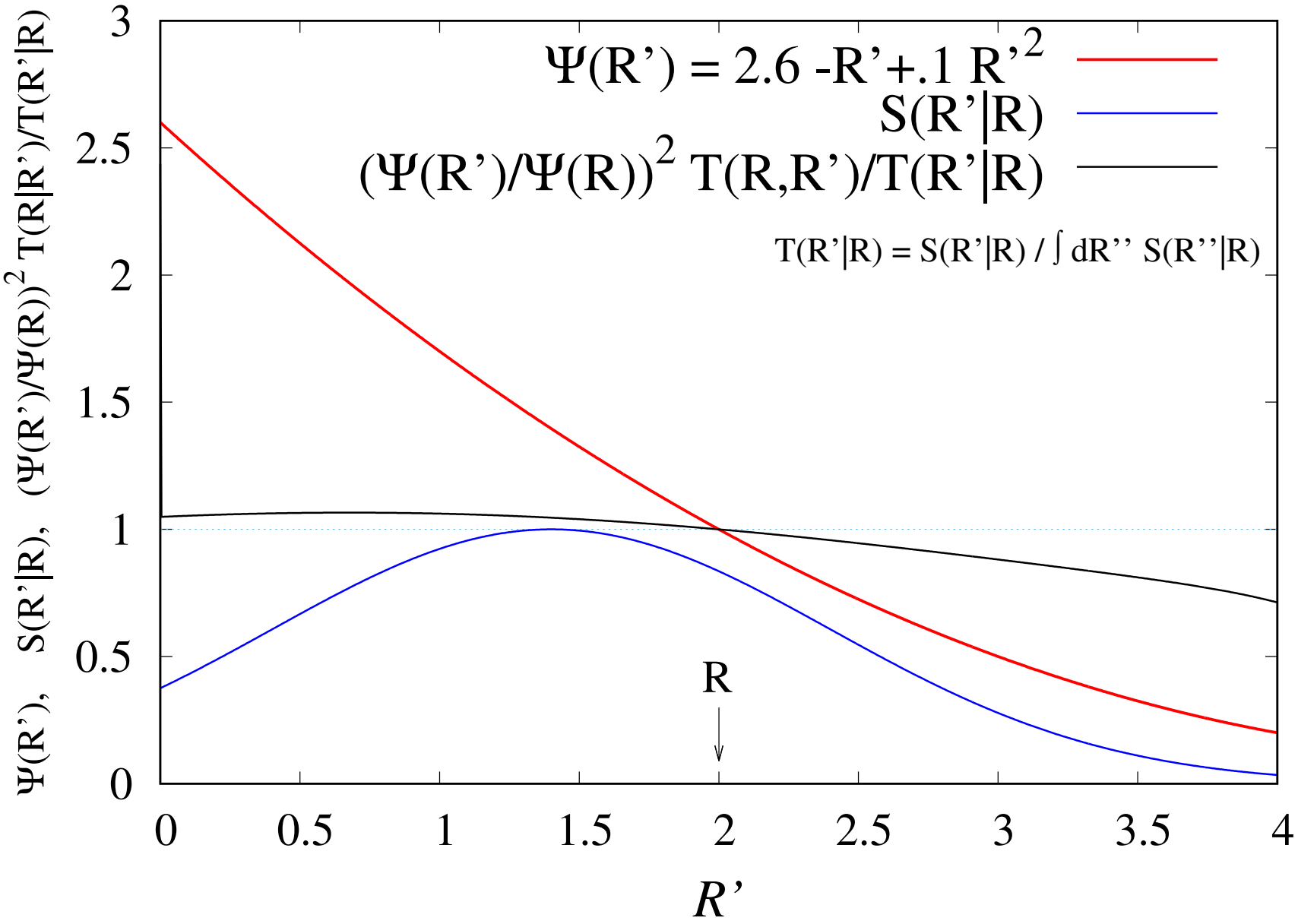
For this $\Psi(R')$, the drifted Gaussian gives perfect acceptance since **V** is constant and drift and diffusion commute! Not true for general $\Psi(R')$.

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?

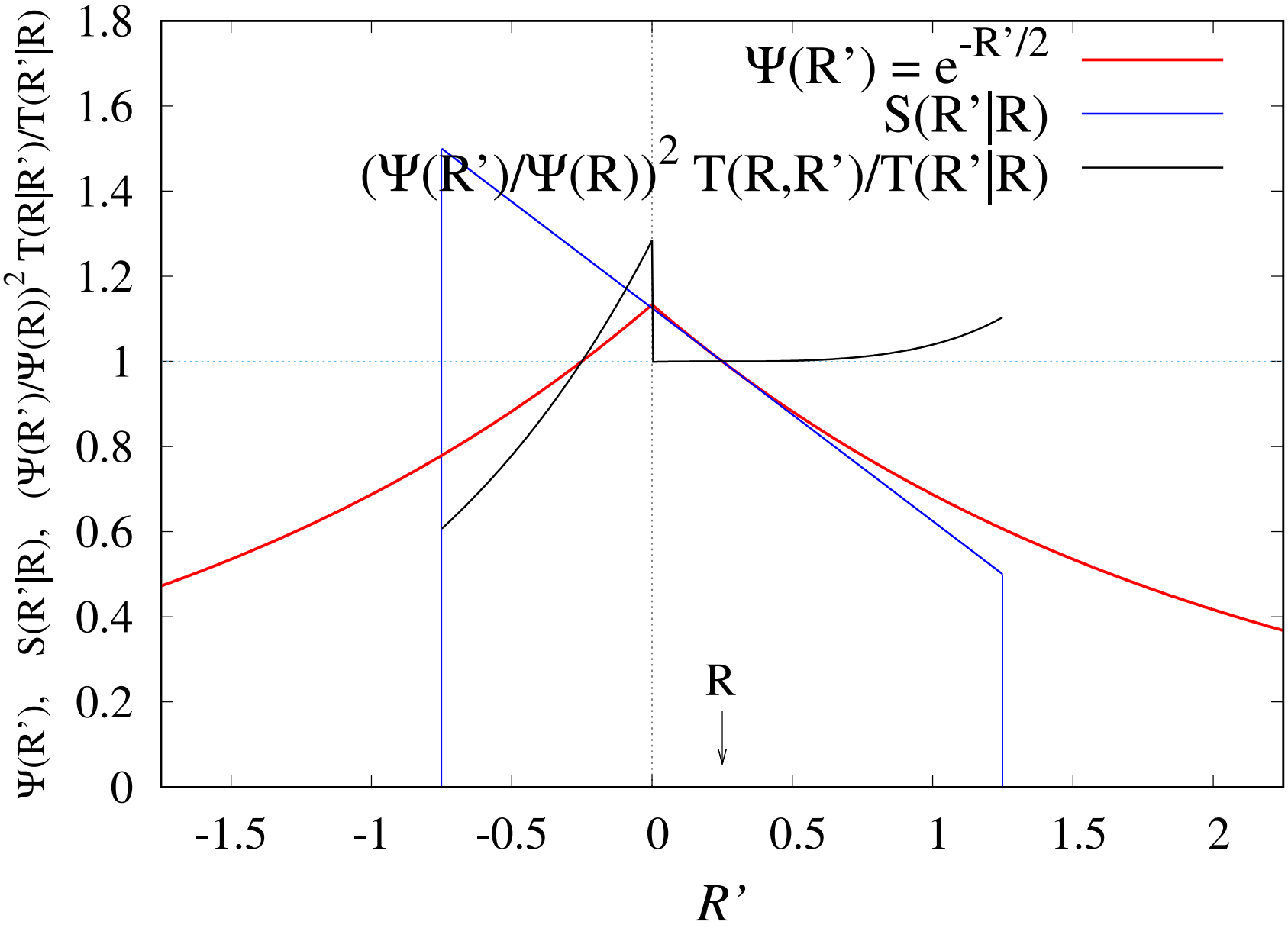
Choice of Proposal Matrix T in Metropolis-Hastings (cont)

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.

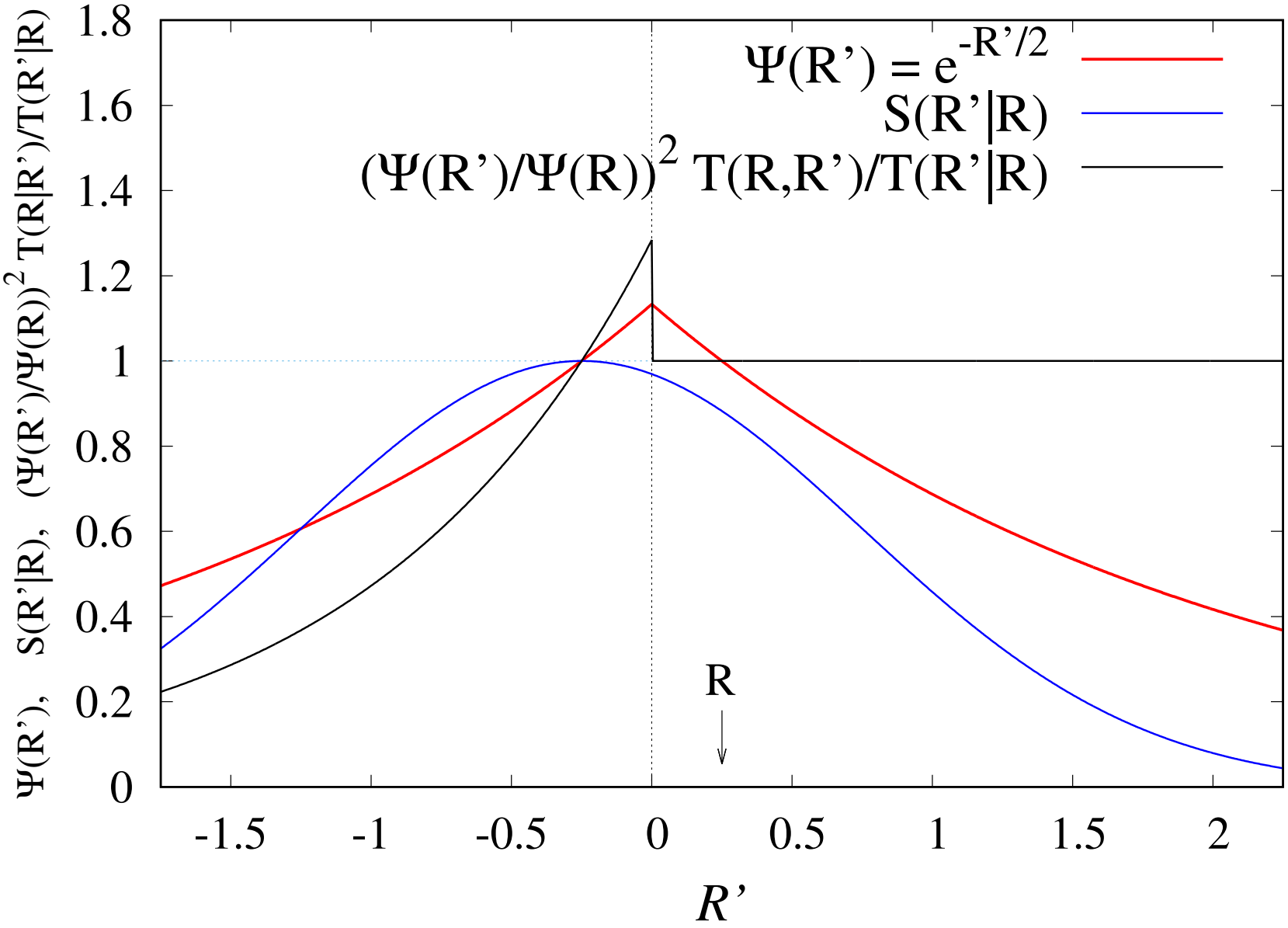
Let's see what happens when it is not.

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.