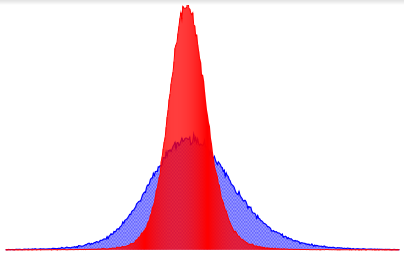


Probability and Statistics in Quantum Monte Carlo

Pablo López Ríos

University of Cambridge

6 August 2013

The need for statistical analysis

- A QMC calculation produces **millions** of data values
- We want a single number (with its **error bar**) as a result:

$$E \pm \sigma_E$$

- Serial correlation needs to be removed
- How to **manipulate** quantities with error bars

Basic statistics

- The configurations $\{\mathbf{R}_i\}_{i=1}^{i=M}$ distributed according to $|\Psi(\mathbf{R})|^2$
- The local energy $E_i = E_L(\mathbf{R}_i) = \Psi^{-1}(\mathbf{R}_i)\hat{H}\Psi(\mathbf{R}_i)$
- $E_L(\mathbf{R})$ forms a distribution with:

Mean

$$E_V = \frac{\langle\Psi|\hat{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle} \approx \bar{E} = \frac{\sum_{i=1}^{M} E_i}{M}$$

Variance

$$\sigma_{E_L}^2 = \frac{\langle\Psi|\hat{H}^2|\Psi\rangle}{\langle\Psi|\Psi\rangle} - \left[\frac{\langle\Psi|\hat{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle}\right]^2 \approx \tilde{\sigma}_{E_L}^2 = \frac{\sum_{i=1}^{M}\left(E_i - \bar{E}\right)^2}{M-1}$$

Basic statistics

- $\bar{E}$ can be determined to a given degree of certainty
  $\rightarrow$ it has an error bar $\sigma_{\bar{E}}$
  $\rightarrow$ it is a distribution with:

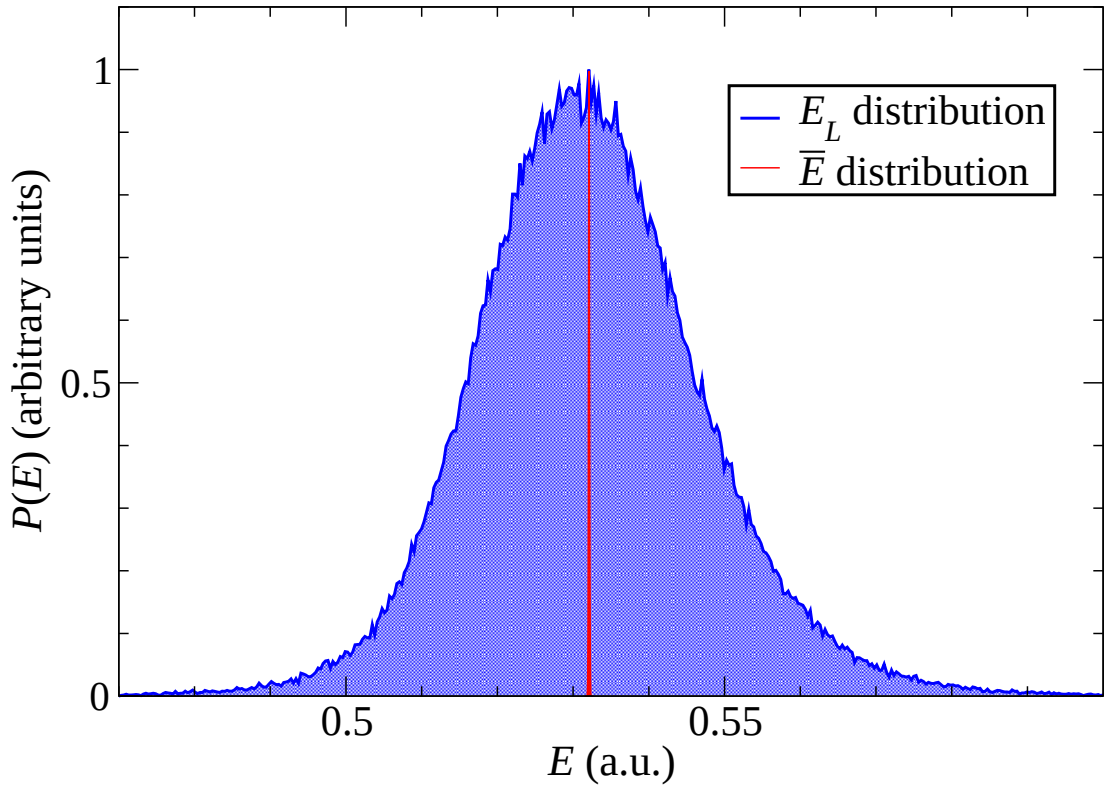
Mean

$$\bar{E} = \frac{\sum_{i=1}^{M} E_i}{M}$$

Variance

$$\sigma_{\bar{E}}^2 \approx \tilde{\sigma}_{\bar{E}}^2 = \frac{\sum_{i=1}^{M} \left(E_i - \bar{E}\right)^2}{M(M-1)}$$

Local energy and mean energy



The local energy distribution is what we sample.
The mean energy distribution is what we obtain.

Sampling of configuration space

$\{\mathbf{R}_i\}_{i=1}^{i=M}$ must be distributed according to $|\Psi(\mathbf{R})|^2$.

Sampling algorithm at i-th step

- Start at config $\mathbf{R}_i$
- **Propose** a new config $\mathbf{R}'_i$
- Compute the **wave function ratio** $q_i = \left|\frac{\Psi(\mathbf{R}'_i)}{\Psi(\mathbf{R}_i)}\right|^2$
- Generate **random number** ξ uniform in $[0,1)$
- Accept/reject step:
  - if $\xi < q_i \rightarrow$ set $\mathbf{R}_{i+1} = \mathbf{R}'_i$ (accept new config)
  - if $\xi > q_i \rightarrow$ set $\mathbf{R}_{i+1} = \mathbf{R}_i$ (reject new config)

Proposing $\mathbf{R}_i \to \mathbf{R}_i'$

- If $\mathbf{R}_i'$ proposed at random:
  $\to$ chances of landing in a **reasonable region** of configuration spaces are slim
  $\to$ q_i will be small
  $\to$ most moves are rejected
  $\to$ poor sampling
- If $\mathbf{R}_i'$ is $\mathbf{R}_i$ plus a small displacement:
  $\to$ $\mathbf{R}_i'$ similar to $\mathbf{R}_i$
  $\to$ $E_L(\mathbf{R}_i')$ similar to $E_L(\mathbf{R}_i)$
  $\to$ **Serial correlation**

Effect of serial correlation

- Consider an uncorrelated set of energies $\{E_1, E_2, E_3, \ldots, E_M\}$
- Generate a new set with artificial serial correlation:

$$\{\underbrace{E_1, \ldots, E_1}_{\tau}, \underbrace{E_2, \ldots, E_2}_{\tau}, \underbrace{E_3, \ldots, E_3}_{\tau}, \ldots, \underbrace{E_M, \ldots, E_M}_{\tau}\}$$

- No new information $\rightarrow$ mean and error bar should not change
- Computed mean of new set is $\bar{E}' = \bar{E}$
- Computed error bar of new set is $\tilde{\sigma}'_{\bar{E}} = \tilde{\sigma}_{\bar{E}}/\sqrt{\tau}$
  $\rightarrow$ **error bar underestimated!**

Removing serial correlation

- In this example we can remove the serial correlation by ignoring $\tau - 1$ of every τ consecutive energies
- For real data the correlation time τ varies during the run
  $\rightarrow$ would need to ignore $\tau_{\max} - 1$ of each $\tau_{\max}$ data to be safe
  $\rightarrow$ lots of relevant data discarded
  $\rightarrow$ inefficiency
- However we may assume that

$$\tilde{\sigma}_{\bar{E}} = \sqrt{\tau}\tilde{\sigma}'_{\bar{E}}$$

  still holds, where τ is the **average** correlation time.
- This is an alternative approach to the **reblocking algorithm**

The reblocking algorithm

- Consider the following operation on data, where the item under each brace is the average of the two numbers above:

$$\underbrace{\underbrace{E_1^{(0)} \quad E_2^{(0)}}_{E_1^{(1)}} \quad \underbrace{E_3^{(0)} \quad E_4^{(0)}}_{E_2^{(1)}}}_{\cdots} \quad \underbrace{\underbrace{E_5^{(0)} \quad E_6^{(0)}}_{E_3^{(1)}} \quad \underbrace{E_7^{(0)} \quad E_8^{(0)}}_{E_4^{(1)}}}_{\cdots}$$

- If applied until $\tau_{\max}$ original data grouped together
  $\rightarrow$ **resulting (smaller) data set is not serially correlated**
- Cannot compute $\tau_{\max}$ directly

Error estimator after reblocking

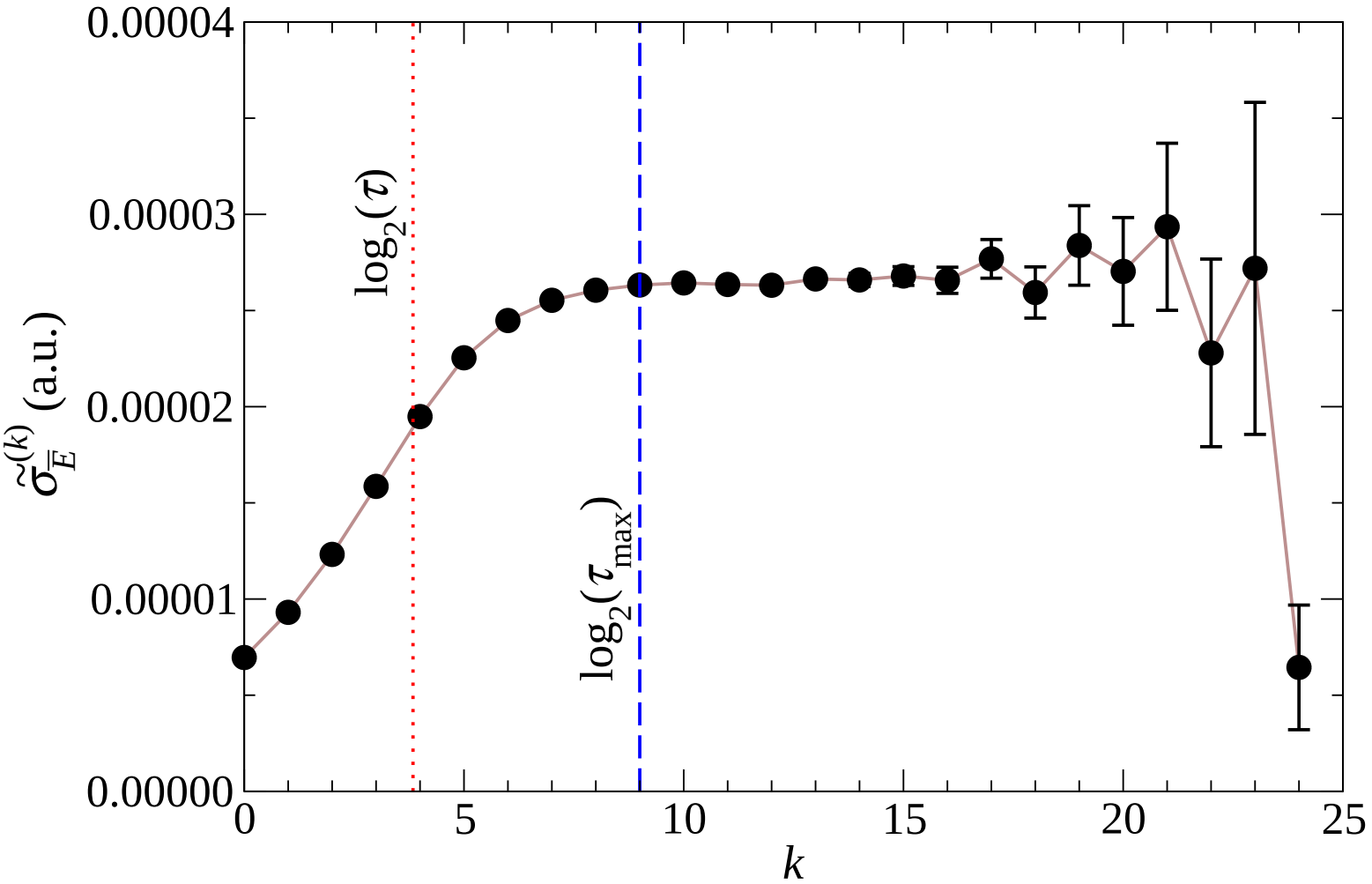
- At the k-th iteration in this procedure:

$$\tilde{\sigma}_{\bar{E}}^{(k+1)2} \approx \tilde{\sigma}_{\bar{E}}^{(k)2} + \frac{2\sum_{i=1}^{M^{(k)}/2}\left(E_{2i-1}^{(k)} - \bar{E}\right)\left(E_{2i}^{(k)} - \bar{E}\right)}{M^{(k)}\left(M^{(k)} - 2\right)}$$

- If there is no serial correlation, the last term tends to zero
- If there is serial correlation, the last term is positive
- $\tilde{\sigma}_{\bar{E}}^{(k)}$ increases to become true error bar by $k \approx \log_2(\tau_{\max})$

Plateau in $\tilde{\sigma}_{\bar{E}}^{(k)}$ signals convergence of reblocking algorithm

Reblock plot



How to run efficient VMC calculations

- Reducing serial correlation, by
  - Choosing an appropriate timestep
  - Using electron-by-electron sampling
  - Skipping the right number of steps between every two calculations of expectation values
- Reducing the intrinsic variance/expense, by
  - Using appropriate trial wave functions

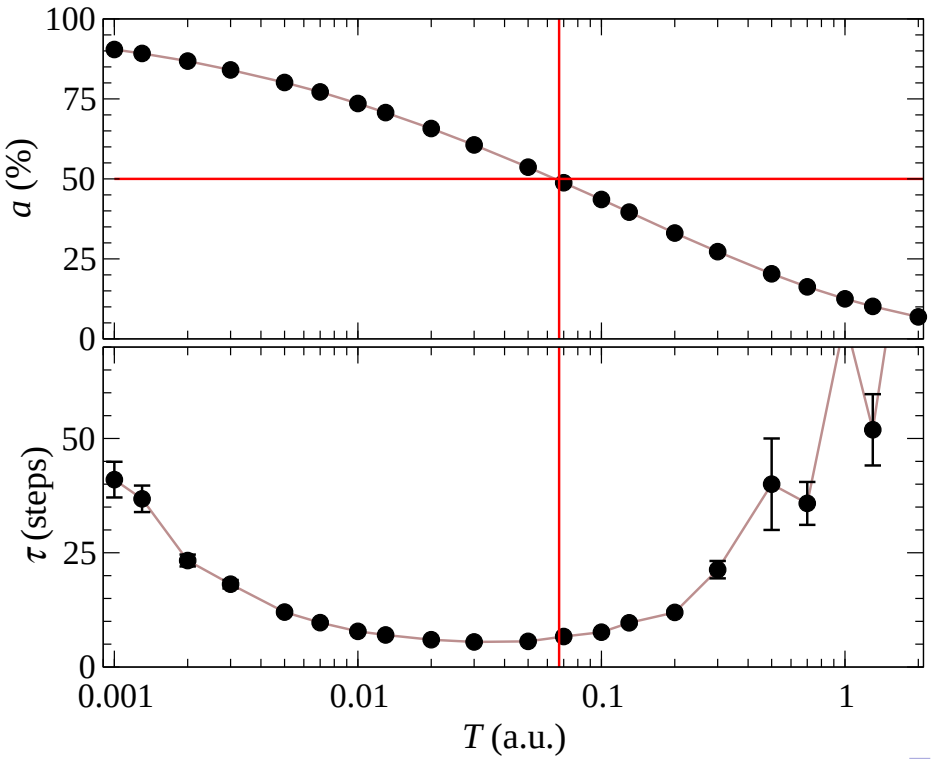
The VMC timestep

- The "timestep" T is the variance of the distribution used to generate the random displacements when proposing moves
- It is actually a squared length, but can be regarded a time if considering a diffusion process
- T does **not** enter the VMC formalism
  $\rightarrow$ can be chosen so that statistics are improved
  - T small $\rightarrow$ $\mathbf{R}_i'$ very similar to $\mathbf{R}_i$
    $\rightarrow$ serial correlation increased
  - T large $\rightarrow$ $\mathbf{R}_i'$ very dissimilar from $\mathbf{R}_i$
    $\rightarrow$ most moves are rejected
    $\rightarrow$ serial correlation increased

The 50% rule

The 50% rule

Choose T such that the acceptance ratio $a = 50\%$



Electron-by-electron sampling

- QMC sampling usually described using configuration moves,
  $\rightarrow$ Configuration-by-configuration sampling (CBCS)
- In practice: accept/reject single-electron moves individually
  $\rightarrow$ Electron-by-electron sampling (EBES)
- Set T to the same value in CBCS and EBES
  $\rightarrow a_{\mathrm{C}} = a_{\mathrm{E}}^{N} \ll a_{\mathrm{E}}$
  $\rightarrow$ EBES wins
- Set T so that $a_{\mathrm{C}} = a_{\mathrm{E}} = a$
  $\rightarrow$ probability of $\mathbf{R}_{i+1} = \mathbf{R}_i$ in the CBCS is a
  $\rightarrow$ probability of $\mathbf{R}_{i+1} = \mathbf{R}_i$ in the EBES is $a^N \ll a$
  $\rightarrow$ EBES wins

The EBES is more efficient

Choosing the right wave function

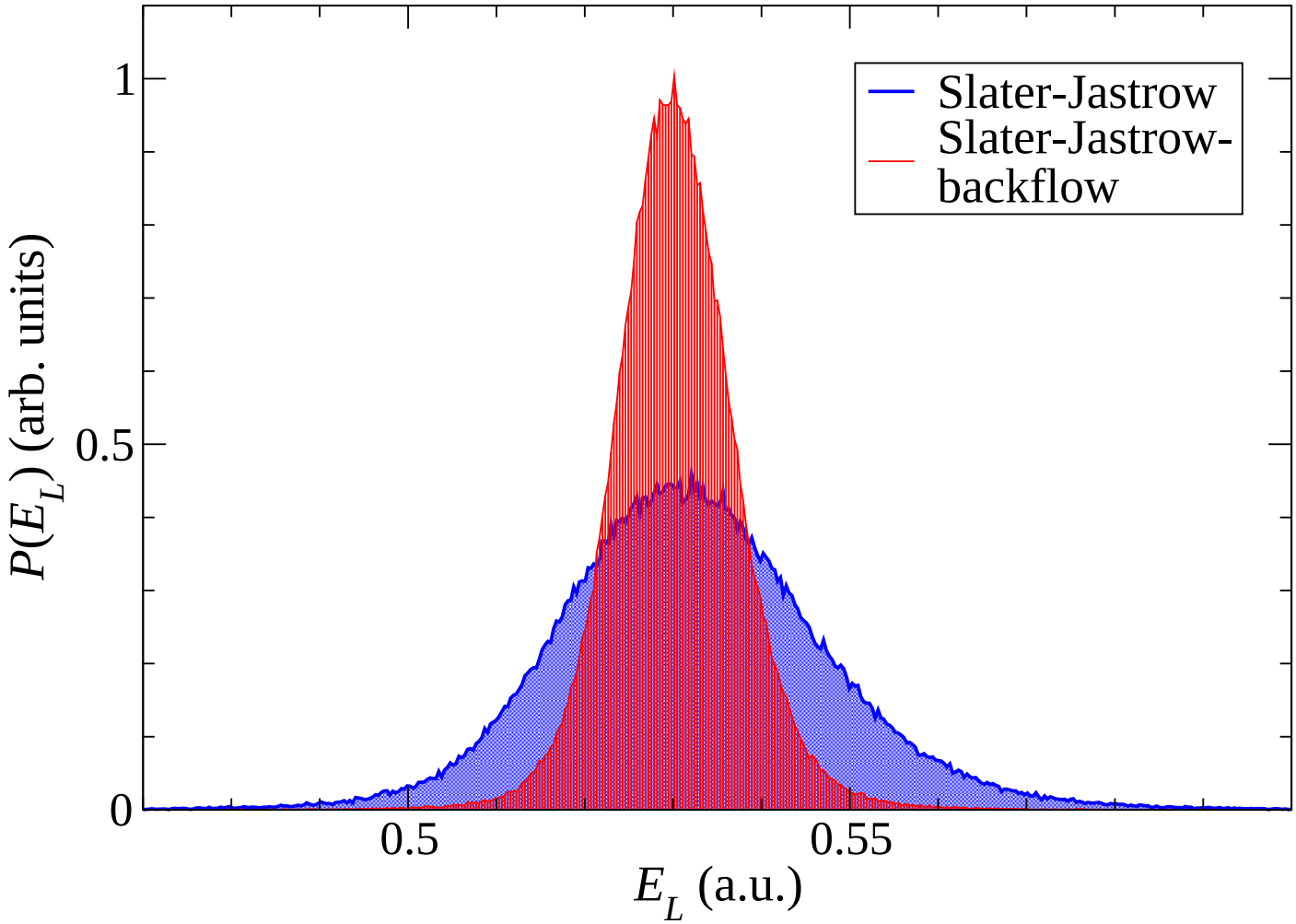
- The **better** the wave function, the **lower** the variance
  $\rightarrow$ fewer data required to achieve a given errorbar
  $\rightarrow$ lower cost
- Generally, **better** wave functions are more expensive to evaluate
  $\rightarrow$ higher cost

Important!

The **best** trial wave function for a problem need not be the **most complicated**

There is a risk of over-parametrization!

Wave functions and the local energy distribution



The projection method

Time-dependent Schrödinger equation

$$\hat{H}(\mathbf{R})\Phi(\mathbf{R},x) = i\frac{\partial\Phi(\mathbf{R},x)}{\partial x}$$

Imaginary time $(ix = t)$ and energy shift

$$\left(\hat{H}(\mathbf{R}) - E_T\right)\Phi(\mathbf{R},t) = -\frac{\partial\Phi(\mathbf{R},t)}{\partial t}$$

Eigenstate expansion

$$\Phi(\mathbf{R},t) = \sum_{n=0}^{\infty} c_n \Phi_n(\mathbf{R}) e^{-(E_n - E_T)t}$$

If $E_T \sim E_0$, only ground state remains as $t \to \infty$

Projection using discrete walkers

- Say we have to solve for time evolution of $f(\mathbf{R},t)$ governed by

$$\hat{H}(\mathbf{R})f(\mathbf{R},t) = -\frac{\partial f(\mathbf{R},t)}{\partial t}$$

- Given $f(\mathbf{R},t)$ at time t and Green's function (GF) $G(\mathbf{R}' \leftarrow \mathbf{R},T)$ for above problem:

$$f(\mathbf{R},t+T) = \int G(\mathbf{R}' \leftarrow \mathbf{R},T) f(\mathbf{R},t) d\mathbf{R}$$

- If $f(\mathbf{R},t)$ probability distribution $\rightarrow$ can represent discretely:

$$f(\mathbf{R},t) \approx \sum_{p=1}^{P} w_p(t) \delta[\mathbf{R} - \mathbf{R}_p(t)]$$

Projection using discrete walkers

Therefore:

$$f(\mathbf{R}, t+T) = \sum_{p=1}^{P} w_p(t) G[\mathbf{R}' \leftarrow \mathbf{R}_p(t), T] \approx$$
$$\approx \sum_{p=1}^{P} w_p(t+T) \delta[\mathbf{R} - \mathbf{R}_p(t+T)]$$

Utility of Green's function

In stochastic solution to time evolution problem, Green's function is transition probability for walkers

Diffusion Monte Carlo

- DMC consists in choosing $f(\mathbf{R},t) = \Phi(\mathbf{R},t)\Psi(\mathbf{R})$
  $\rightarrow$ Ψ is the trial wave function
  $\rightarrow$ Φ is called the DMC wave function
- If T small GF separates into **drift**, **diffusion** and **branching**
- $f(\mathbf{R},t)$ not a probability distribution unless Φ and Ψ have same sign everywhere in space

The fixed-node approximation

Φ constrained to having same nodes as Ψ
$\rightarrow$ resulting Φ is lowest energy wave function with nodal structure of Ψ
$\rightarrow$ DMC is always better than VMC

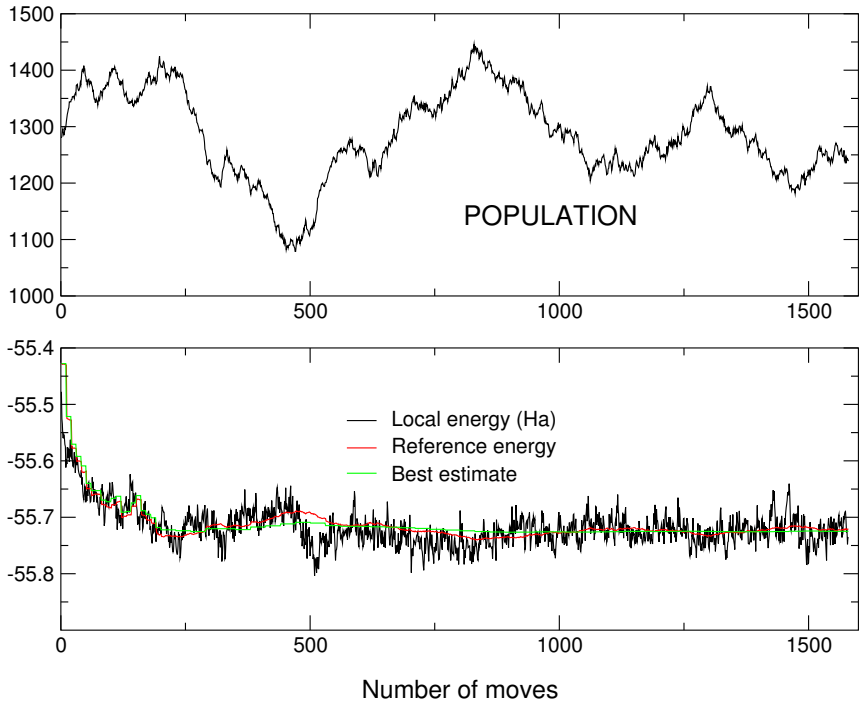
The DMC algorithm

- **Start**: P walkers $\{\mathbf{R}_{0,\alpha}\}_{\alpha=1}^{P}$ distributed as $|\Psi(\mathbf{R})|^2$ (VMC)
- DMC evolution of the walkers:
  - **Drift-diffusion**: move $\mathbf{R}_{i,\alpha} \to \mathbf{R}'_{i,\alpha}$
  - **Branching**: define weight $w_{i,\alpha}$
    $\to$ configurations breed/die as per **branching factor** $w'_{i,\alpha}/w_{i,\alpha}$
    $\to$ variable number of walkers P_i
- **Equilibrate** the walkers until we reach infinite-time limit
  $\to$ look at $E_i = \sum_{\alpha=1}^{P_i} w_{\alpha,i} E_{\alpha,i} / \sum_{\alpha=1}^{P_i} w_{\alpha,i}$
- **Accumulate** data to required accuracy

DMC mixed estimator

$$\langle A \rangle_{\mathrm{DMC}} = \lim_{t\to\infty} \langle \Psi | \hat{A} | \Phi(t) \rangle / \langle \Psi | \Phi(t) \rangle$$

Calculation of the energy in DMC

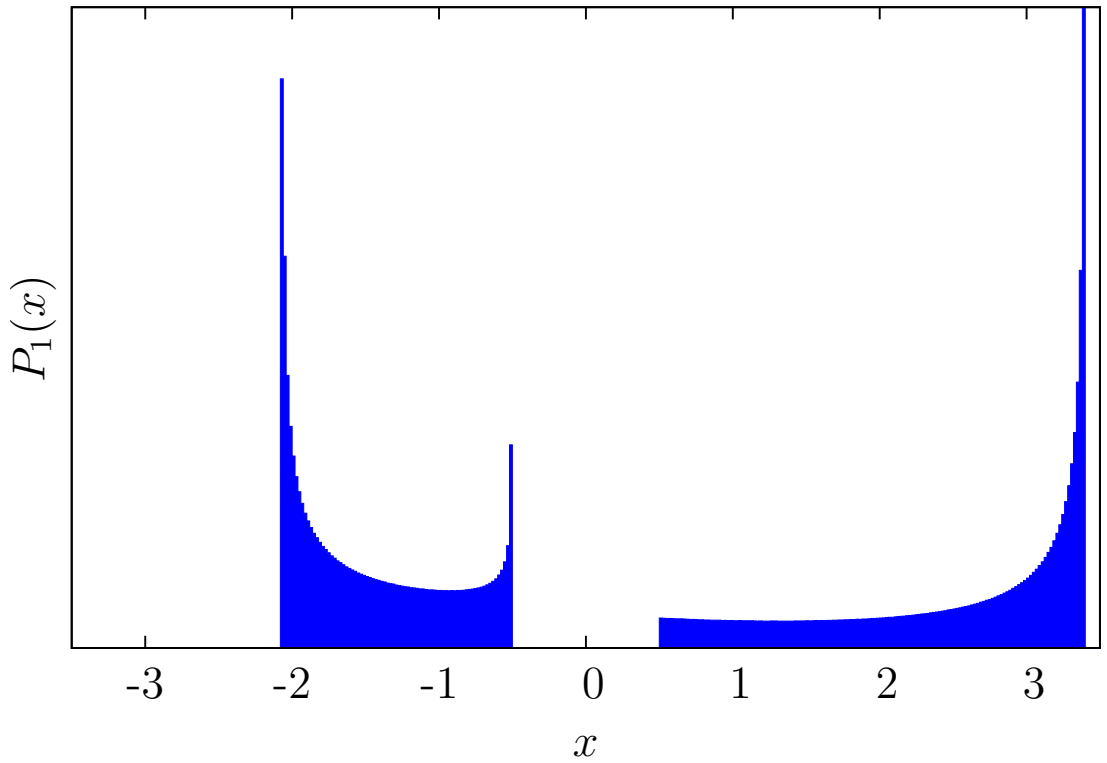


$$E_D \approx \bar{E} = \frac{\sum_{i=1}^{M} W_i E_i}{\sum_{i=1}^{M} W_i} \quad ; \quad \sigma_{\bar{E}}^2 \approx \tilde{\sigma}_{\bar{E}}^2 = \frac{\sum_{i=1}^{M} W_i \left(E_i - \bar{E}\right)^2}{M\left(\sum_{i=1}^{M} W_i - \frac{\sum_{i=1}^{M} W_i^2}{\sum_{i=1}^{M} W_i}\right)}$$

Sources of error in DMC

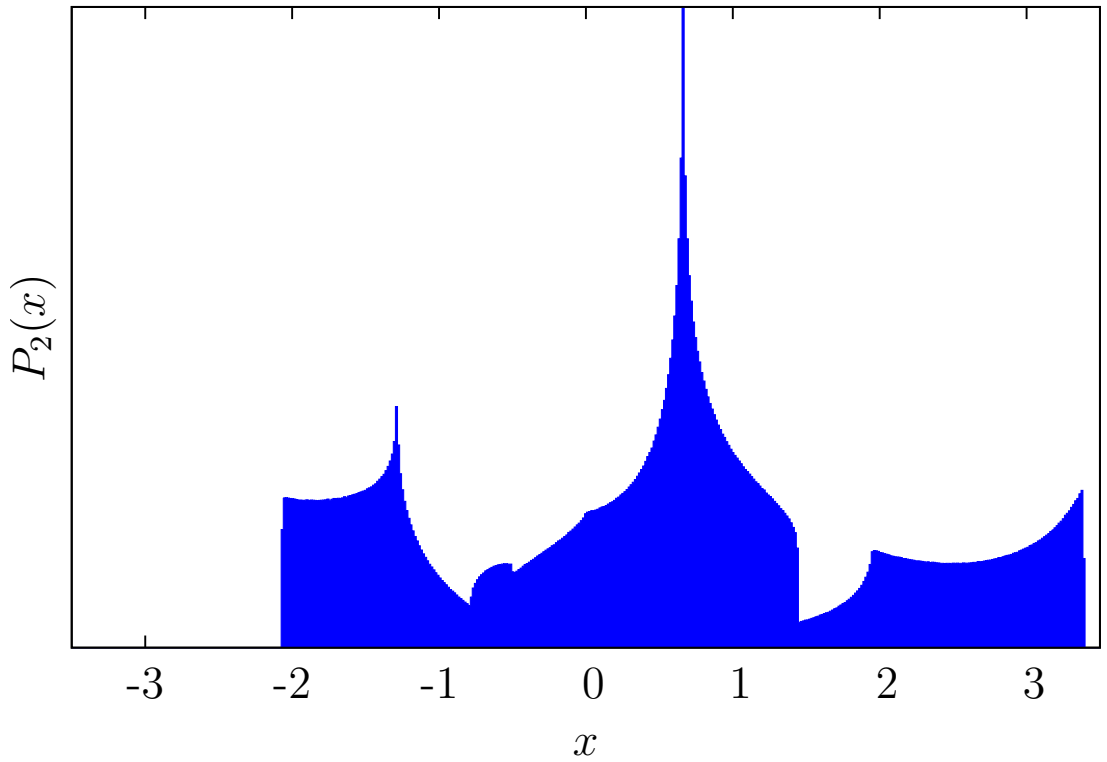
- **Timestep**: we have assumed that T is small
  $\rightarrow$ must extrapolate to zero timestep to obtain a reliable result
  $\rightarrow$ cannot use timestep to improve statistics
- **Population**: Φ is represented by set of configurations
  $\rightarrow$ must use sufficient configurations to represent it accurately
  $\rightarrow$ possible to extrapolate to infinite population
- **Fixed-node error**: only limitation of DMC
  $\rightarrow$ E_D is still variational (very important!)
  $\rightarrow$ can be reduced by using Ψ with better nodes
- **Locality approximation**: from pseudopotentials
  $\rightarrow$ E_D non-variational
  $\rightarrow$ goes away with good Ψ

Distribution of total energy estimate



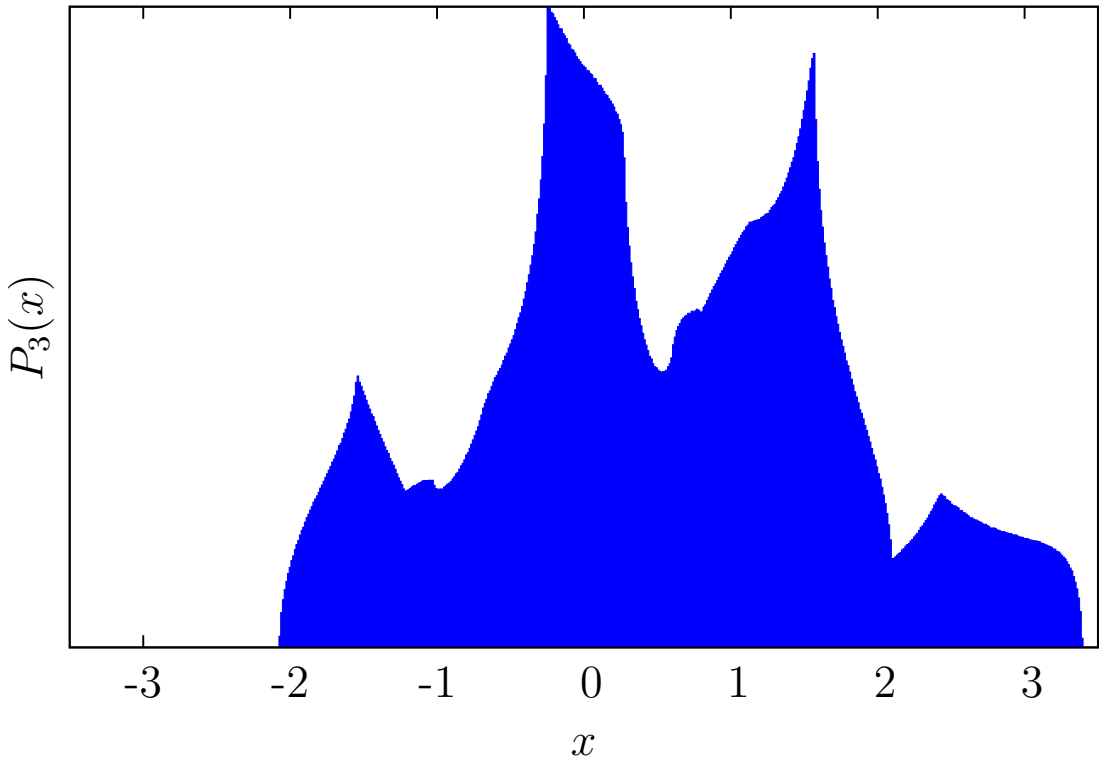
- Average of 1 random variable
- $P_1(x)$ is PDF of $x = E_L(1)$

Distribution of total energy estimate



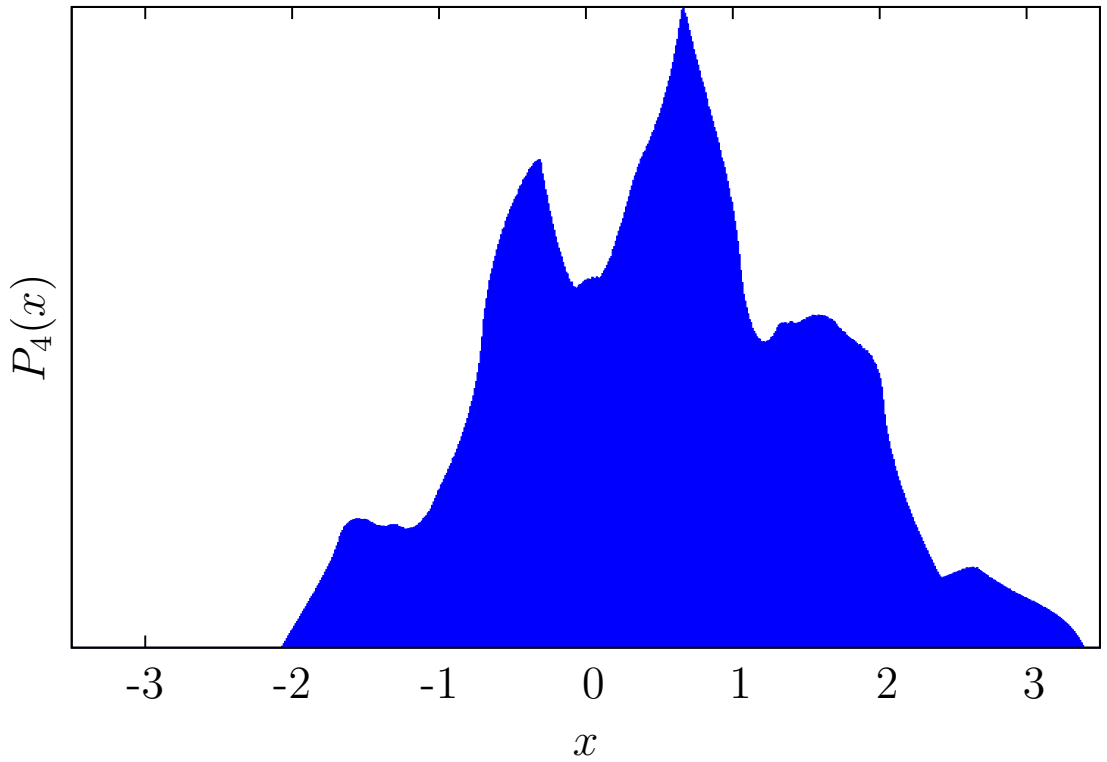
- Average of 2 random variables
- $P_2(x)$ is PDF of $x = \frac{1}{2}(E_L(1) + E_L(2))$

Distribution of total energy estimate



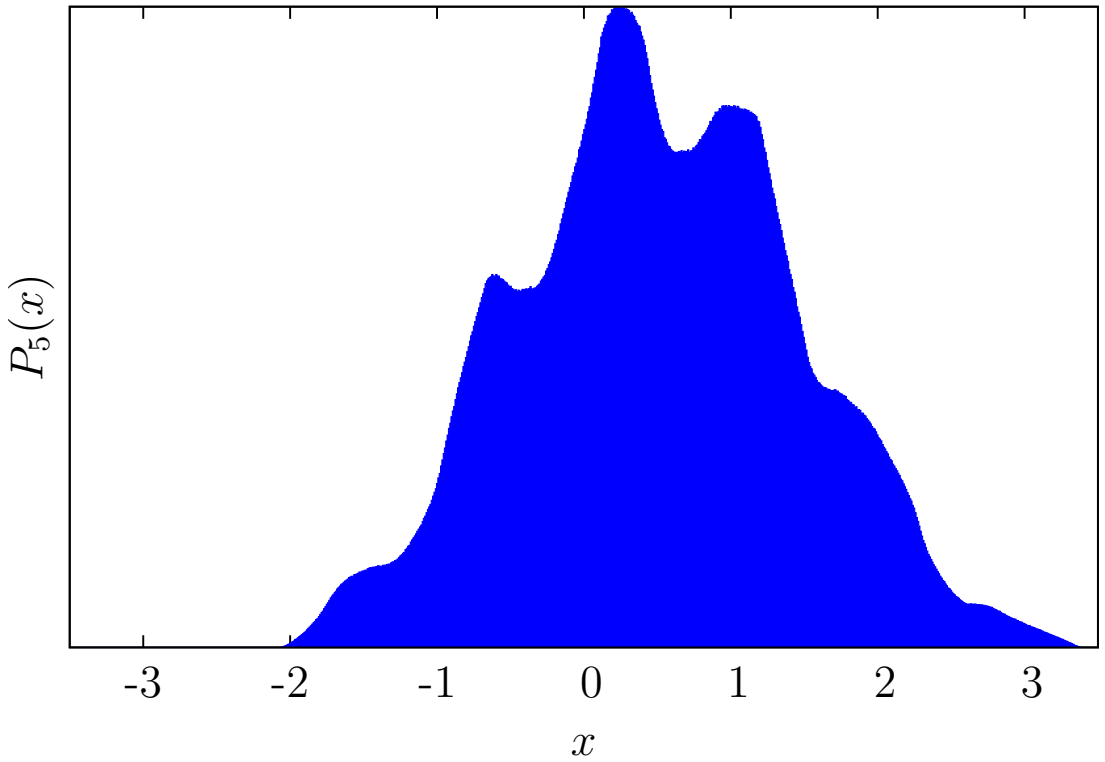
- Average of 3 random variables
- $P_3(x)$ is PDF of $x = \frac{1}{3}(E_L(1) + E_L(2) + E_L(3))$

Distribution of total energy estimate



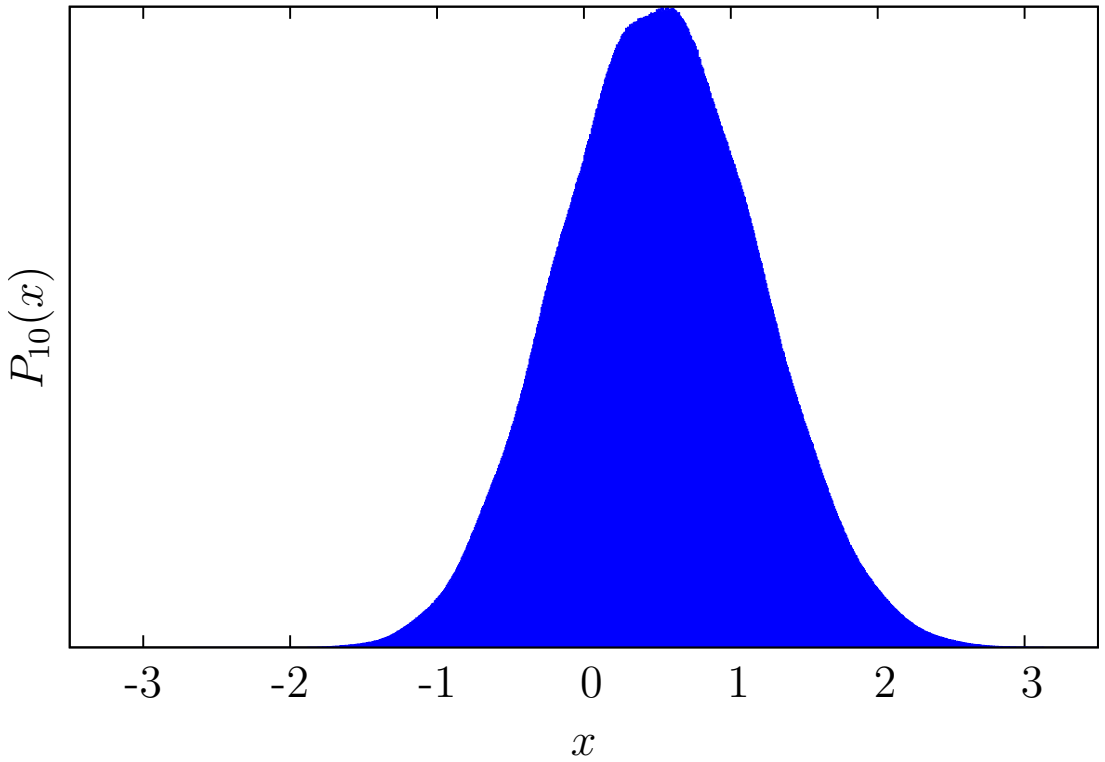
- Average of 4 random variables
- $P_4(x)$ is PDF of $x = \frac{1}{4}(E_L(1) + E_L(2) + E_L(3) + E_L(4))$

Distribution of total energy estimate



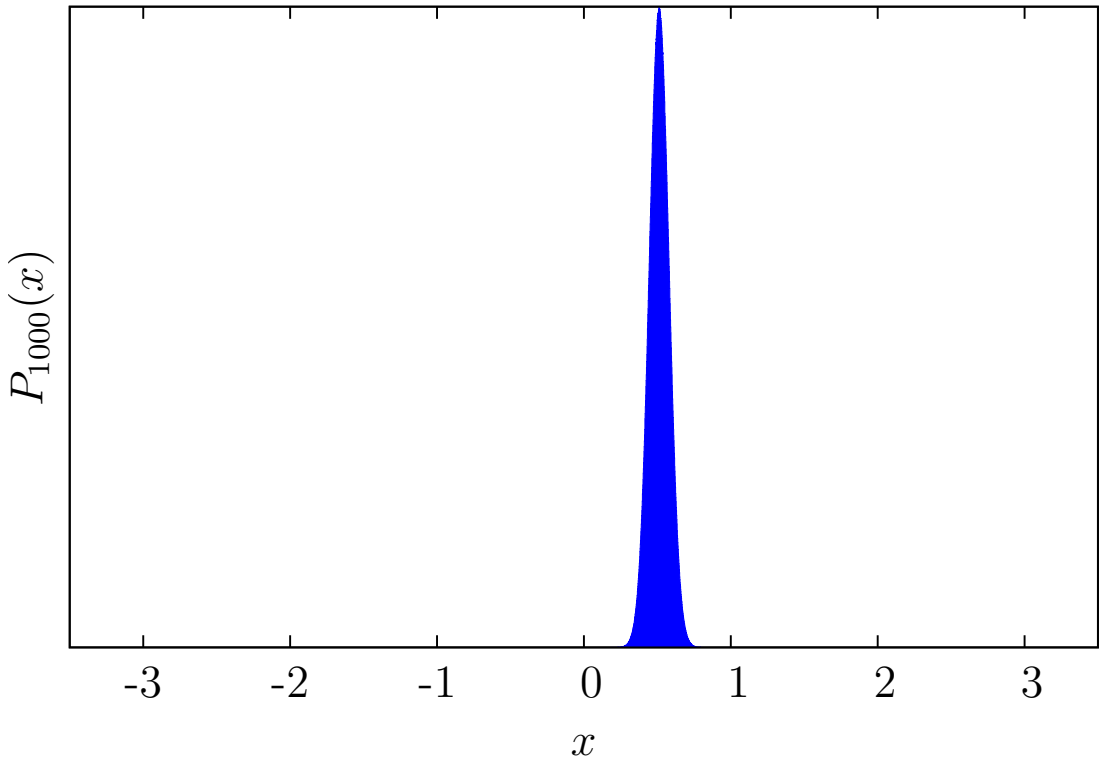
- Average of 5 random variables
- $P_5(x)$ is PDF of $x = \frac{1}{5}(E_L(1) + E_L(2) + E_L(3) + E_L(4) + E_L(5))$

Distribution of total energy estimate



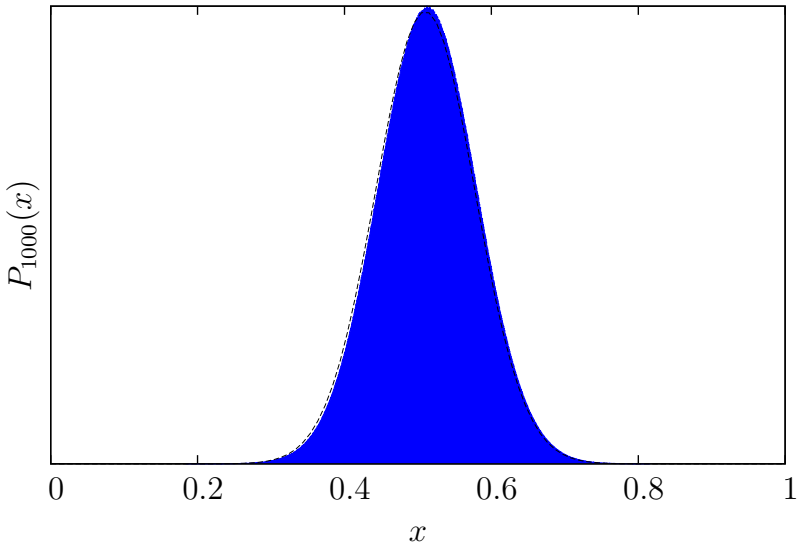
- Average of 10 random variables
- $P_{10}(x)$ is PDF of $x = \frac{1}{N}\sum_{n=1}^{N} E_L(n)$

Distribution of total energy estimate



- Average of 100 random variables
- $P_{1000}(x)$ is PDF of $x = \frac{1}{N}\sum_{n=1}^{N} E_L(n)$

Central Limit Theorem



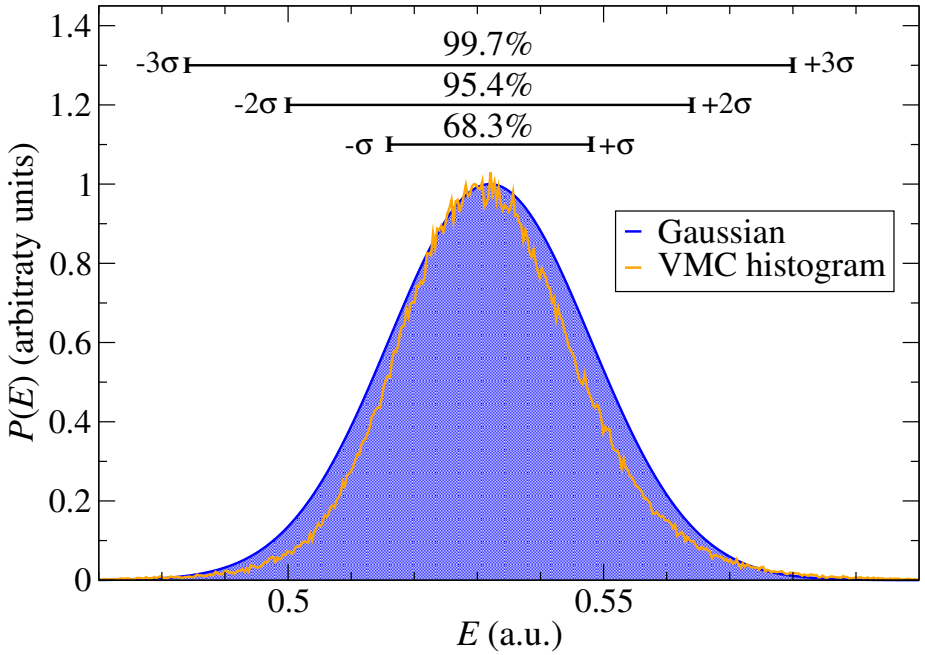
- Average of N random variables $\rightarrow$ *Normal distribution*
- Defined by 2 numbers, the mean and standard deviation
- Centred at mean, width of $\sigma \propto 1/\sqrt{N}$
- Probability is all close to the mean

The normal distribution

- The normal distribution is $D(E;\bar{E},\sigma) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left[-\frac{(E-\bar{E})^2}{2\sigma^2}\right]$
- The probability of the E being in an interval (A,B) is
  - $P(A<E<B) = f\left(\frac{B-\bar{E}}{\sigma}\right) - f\left(\frac{A-\bar{E}}{\sigma}\right)$
  - $f(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-y^2/2\right) dy$
- One-sigma interval $(\bar{E}-\sigma, \bar{E}+\sigma) \rightarrow 68.3\% \rightarrow$ unreliable
- Two-sigma interval $(\bar{E}-2\sigma, \bar{E}+2\sigma) \rightarrow 95.4\% \rightarrow$ reliable
- Three-sigma interval $(\bar{E}-3\sigma, \bar{E}+3\sigma) \rightarrow 99.7\% \rightarrow$ very reliable

The normal distribution

Comparison of a Gaussian and the local energy distribution



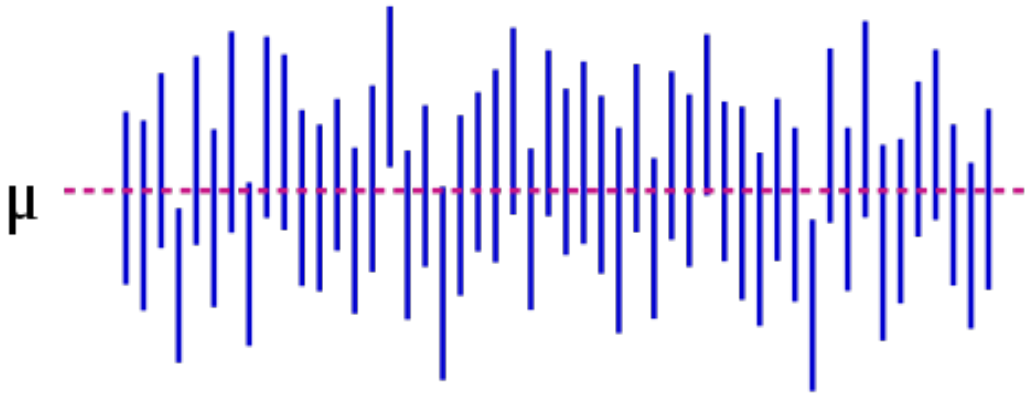
From Central Limit Theorem:

The mean energy is exactly normal

How to compare quantities with errorbars

- Want to find distribution of difference, denoted $\left(\bar{E}_- \pm \sigma_-\right) = \left(\bar{E}_1 \pm \sigma_1\right) - \left(\bar{E}_2 \pm \sigma_2\right)$
- Results in
  - $\bar{E}_- = \bar{E}_1 - \bar{E}_2$
  - $\sigma_-^2 = \sigma_1^2 + \sigma_2^2$
- Example:
  - Ψ_1 gives $E_1 = -14.66728(2)$ a.u.
  - Ψ_2 gives $E_2 = -14.66733(7)$ a.u.
  - Comparison: $E_- = 0.00005(7)$ a.u. $\rightarrow$ 76% chance of $E_2 < E_1$ $\rightarrow$ unreliable!
  - If $E_2 = -14.66733(2)$ a.u. instead $\rightarrow E_- = 0.00005(3)$ a.u. $\rightarrow$ 95% chance of $E_2 < E_1$ $\rightarrow$ reliable

What do error bars mean?



If we run the same calculation many times with different random numbers, in 95% of them the two-sigma interval will include exact value

$\rightarrow$ this is the definition of 'confidence interval'

Is random error an 'extra' error?

- Computers cannot do integration exactly
- All methods do integration approximately
- **Finite basis sets** $\rightarrow$ basis set error unkown but controlled
- **Quadrature on grid** $\rightarrow$ quadrature error unkown but controlled
- **Monte Carlo** $\rightarrow$ random error is known and controlled

QMC has a different **type** of integration error

Summary

- **Reblocking algorithm** applied using the REBLOCK utility
- **Average correlation time** τ given in VMC runs and REBLOCK utility
- **VMC timestep** automatically optimized to give $a = 50\%$ (do not apply on HEG)
- **EBEA** is the default in both VMC and DMC
- **DMC statistics** monitored using GRAPHIT utility
- **Timestep extrapolation** carried out using the EXTRAPOLATE_TAU utility