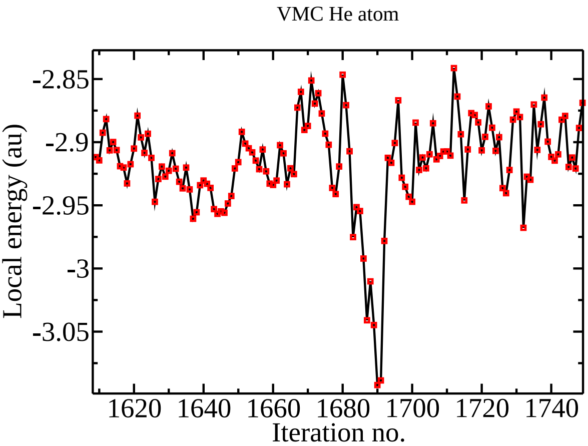
Random errors and outliers in QMC

Robert Lee
rml38@cam.ac.uk
TCM Group
Cavendish Laboratory
University of Cambridge

Outline

- QMC and serial correlation
- Correlation lengths
- Reblocking
- Statistical errors
- The effect of uncertainty in the correlation length

Serial correlation



Both VMC and DMC data typically show some degree of serial correlation.

Correlation lengths

If we perform a VMC calculation, taking n steps and obtaining an estimate σ_0^2 of the variance, then the standard error is

$$\begin{aligned}\Delta_{\mathrm{naive}} &= \frac{\sigma_0}{\sqrt{n}} \\ \Delta_{\mathrm{correct}} &= \frac{\sigma_0}{\sqrt{n/n_{\mathrm{corr}}}} \,,\end{aligned}$$

where n_{corr} is the correlation length.

It is usually the case that we only have an estimate of n_{corr}, probably from the data itself.

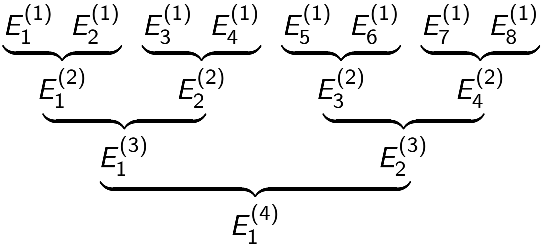
Correlation lengths

Estimate the correlation length using

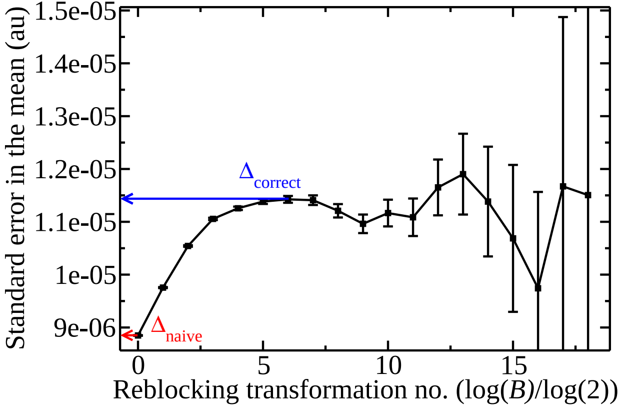
$$n_{\mathrm{corr}}(L) = 1 + 2\sum_{k=1}^{L} \left\langle \, (A_j - \langle A \rangle) \, (A_{j+k} - \langle A \rangle) \, \right\rangle_j \, ,$$

where the sum is truncated as soon as the inequality $L < 3n_{\mathrm{corr}}(L)$ is violated.

Reblocking



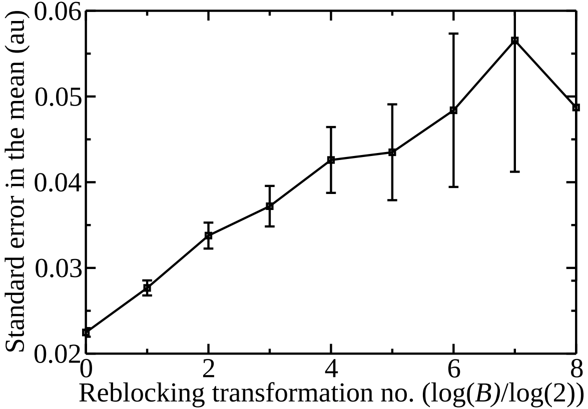
$E_1^{(1)}$ $E_2^{(1)}$ $E_3^{(1)}$ $E_4^{(1)}$ $E_5^{(1)}$ $E_6^{(1)}$ $E_7^{(1)}$ $E_8^{(1)}$
$E_1^{(2)}$ $E_2^{(2)}$ $E_3^{(2)}$ $E_4^{(2)}$
$E_1^{(3)}$ $E_2^{(3)}$
$E_1^{(4)}$



Standard error in the mean (au)
1.5e-05
1.4e-05
1.3e-05
1.2e-05
1.1e-05
1e-05
9e-06
Δ_{correct}
Δ_{naive}
0
5
10
15
Reblocking transformation no. (log(B)/log(2))

Reblocking

Particularly for short/expensive calculations, both methods can lead to considerable uncertainty in the error bar



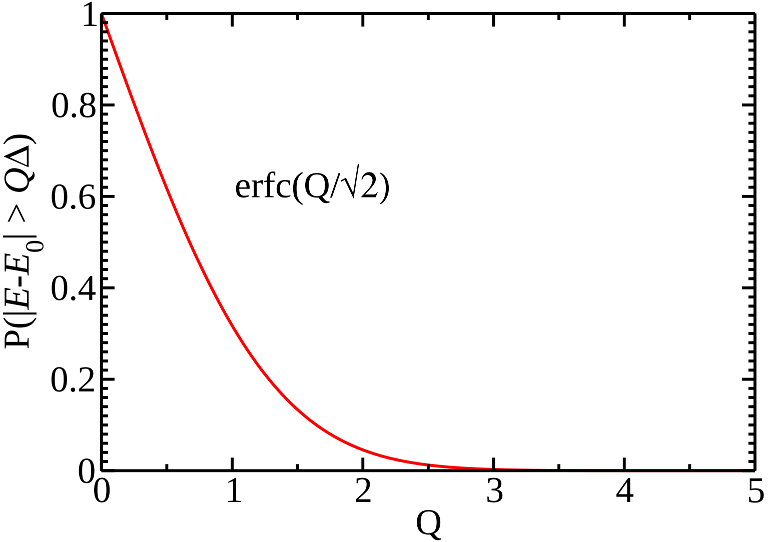
U. Wolff, Comput. Phys. Commun. **156**, 143 (2004).

Gathering some statistics

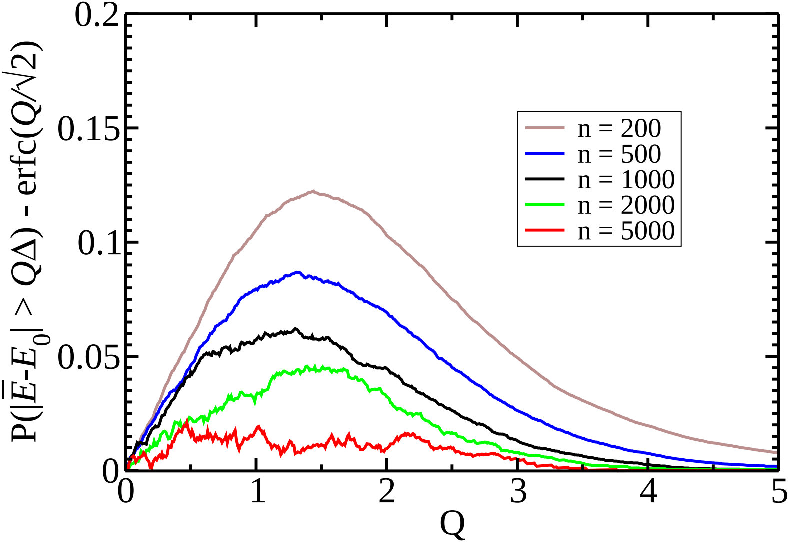
Example: The C atom

- Perform a VMC run consisting of 10^7 steps
- Split the data up into $N_{\rm runs}$ 'runs' of length $10^7/N_{\rm runs}$
- Estimate the mean, correlation length $n_{\rm corr}$ and corrected error for each run separately
- Observe how the uncertainty in $n_{\rm corr}$ affects the probability of observing an energy more than Q error bars from the underlying mean

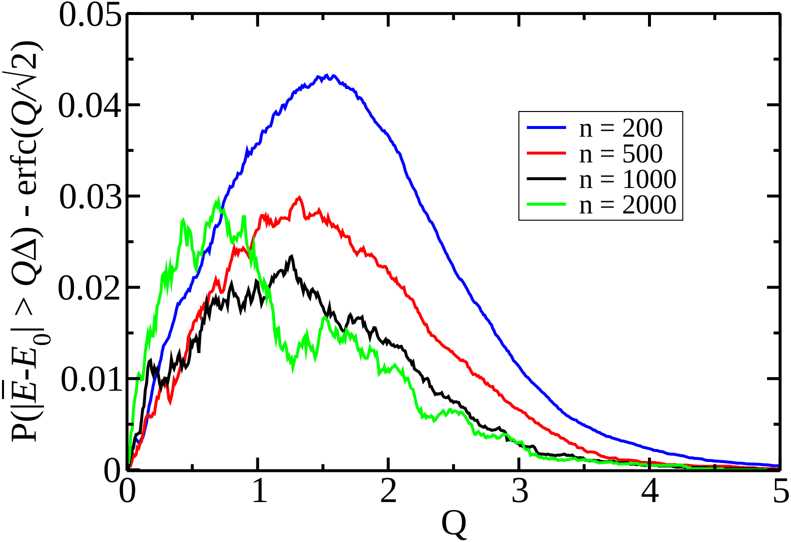
If we knew the error bar exactly...



Results - C atom, 10^7 steps

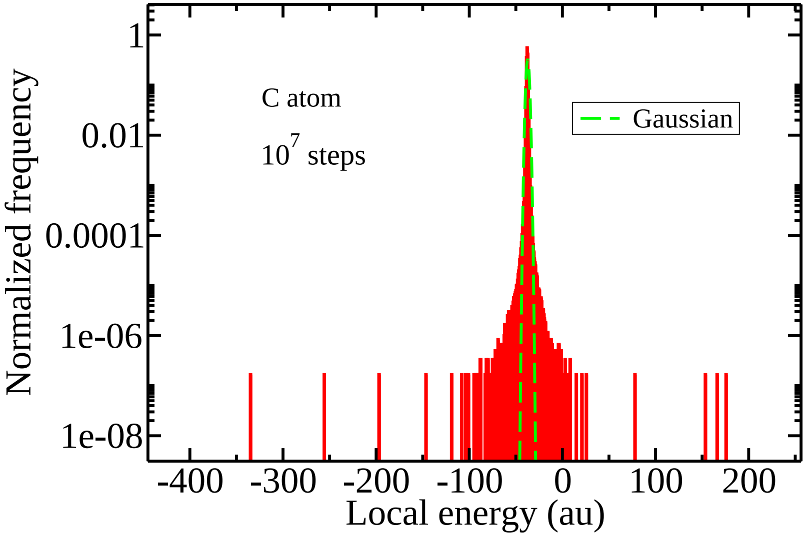


Results - bulk Si, 1.5×10^7 steps



Non-Gaussian distributions

J. R. Trail, Phys. Rev. E **77**, 016703 (2008).



Outliers when we have a Gaussian $\mathrm{P}_{\mathrm{loc}}(E_L)$

We can say something more about the result when the distribution of local energies is Gaussian.
The distribution of mean energies is

$$\mathrm{p}_{\mathrm{ave}}(\bar{E}) = \sqrt{\frac{\nu_0}{2\pi\sigma_0^2}} \exp\left[\frac{-(\bar{E}-E_0)^2}{2\sigma_0^2/\nu_0}\right]$$

The distribution of errors is

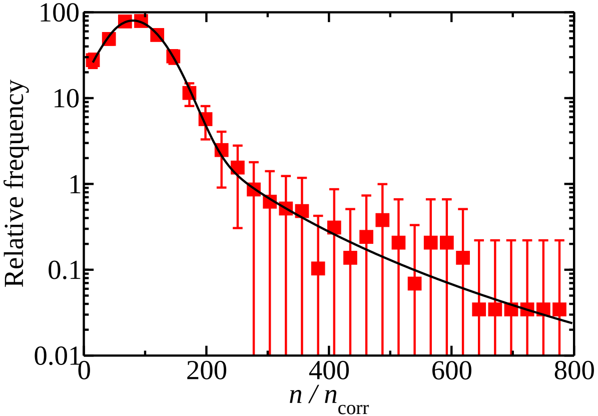
$$\mathrm{p}_{\mathrm{err}}(\Delta,\nu) = \frac{\Delta^{\nu-2} \exp\left[-\frac{\nu(\nu-1)\Delta^2}{2\sigma_0^2}\right] \mathrm{p}_{\mathrm{ind}}(\nu)}{\left(\frac{\nu(\nu-1)}{\sigma_0^2}\right)^{\frac{1-\nu}{2}} 2^{\frac{\nu-3}{2}}\ \Gamma\left(\frac{\nu-1}{2}\right)},$$

where Δ is the estimated error bar, $\nu = n/n_{\mathrm{corr}}$, and ν_0 is the 'true' value of n/n_{corr}.

Estimating n/n_{corr}

Fit the distribution of $\nu = n/n_{\mathrm{corr}}$ to the VMC data with the form

$$\begin{aligned} \mathrm{p}_{\mathrm{ind}}(\nu) &= A\exp\left(\frac{-(\nu-\mu_\nu)^2}{2\sigma_\nu^2}\right)\left[1+\mathrm{erf}\left(\frac{\alpha(\nu-\mu_\nu)}{\sqrt{2\sigma_\nu^2}}\right)\right] \\ &+ B\exp\left(\frac{-C}{\nu}\right)|\nu|^{-D} \end{aligned}$$



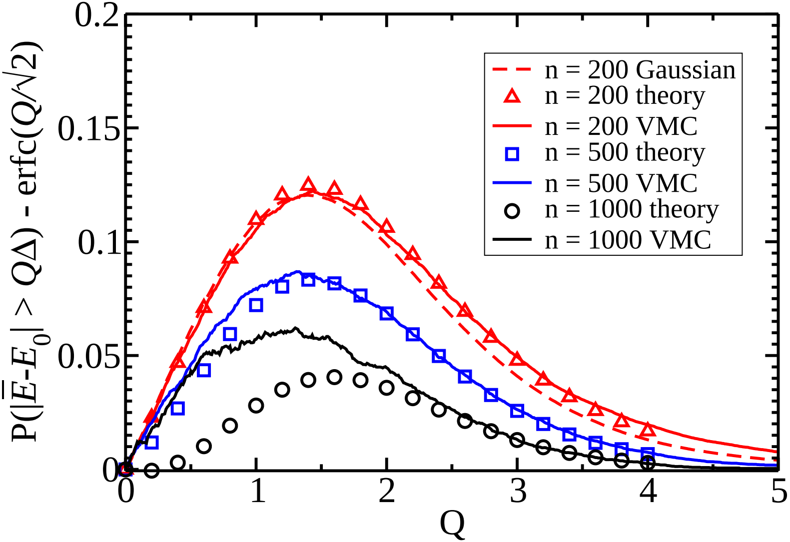
Frequency of outliers

The probability of observing a mean energy more than Q error bars from the underlying mean may then be found,

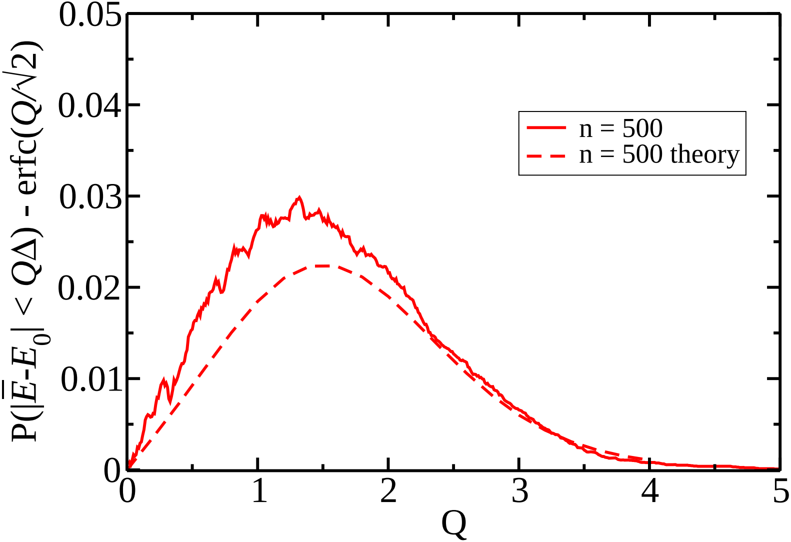
$$\begin{aligned}\mathrm{P}\left(\delta\bar{E} > Q\Delta\right) &= \int_2^\infty \mathrm{d}\nu \int_0^\infty \mathrm{d}\Delta \left[\int_{E_0+Q\Delta}^\infty \mathrm{d}\bar{E}\ \mathrm{p}_{\mathrm{ave}}(\bar{E})\right. \\ &+ \left.\int_{-\infty}^{E_0-Q\Delta} \mathrm{d}\bar{E}\ \mathrm{p}_{\mathrm{ave}}(\bar{E})\right]\mathrm{p}_{\mathrm{err}}(\Delta,\nu)\ ,\end{aligned}$$

where $\delta\bar{E} = |\bar{E} - E_0|$

Results (C atom)



Results (bulk Si)



Conclusions

- One must exercise caution when there is little data because the conventional interpretation of error bars may not hold.
- My data indicate that most of this effect is due to uncertainty in the correlation length.
- Obtaining an accurate estimate of the correlation length from elsewhere (a larger data set on a similar system) could help.

Acknowledgements

Richard Needs
Neil Drummond
EPSRC UK

TCM