Optimum and Efficient sampling for Variational Quantum Monte Carlo

J.R. Trail and R. Maezono

School of Information Science, JAIST,

Nomi, Ishikawa 923-1292, Japan

July 2009

- Only Variational Monte Carlo considered (for now...)
- Monte Carlo can be implemented for any choice of sample distribution - $P = \psi^2$ is just convenient
- When is the CLT valid?
- What is the *optimum* choice of sample distribution?
- What is an *efficient* choice of sample distribution?
- Results for isolated atom/diatomic molecules - comparison of ‘optimum’, ‘efficient’, and ‘standard’ sampling

VMC and 'Standard' Sampling

- For $P = \psi^2$

$$\text{Est}_r\left[E_{tot}\right] = \frac{1}{N}\sum_{i=1}^{N} E_L(\mathsf{R}_i)$$

- CLT $\Rightarrow$ distributed Normally with [a]

$$\mu = \frac{\int \psi^2 E_L d\mathbf{R}}{\int \psi^2 d\mathbf{R}} \quad , \quad \sigma^2 = \frac{1}{N}\frac{\int \psi^2 \left(E_L - \mu\right)^2 d\mathbf{R}}{\int \psi^2 d\mathbf{R}}$$

- Estimates are available:

$$\overline{\mu} = \frac{1}{N}\sum_{i=1}^{N} E_L(\mathbf{R}_i) \quad , \quad \overline{\sigma}^2 = \frac{1}{N.(N-1)}\sum_{i=1}^{N} \left(E_L(\mathbf{R}_i) - \overline{\mu}\right)^2$$

- Total energy is a sample drawn from a Normal distribution whose shape we can estimate,

$\longrightarrow$ The error is **controlled** if the CLT is **valid**

[a] We also require that the variance is finite, and N is large enough

VMC and General Sampling

- For $P = \psi^2/w$, what is the distribution of

$$\mathsf{Est}_r\left[E_{tot}\right] = \frac{\sum w(\mathsf{R}_i) E_L(\mathsf{R}_i)}{\sum w(\mathsf{R}_i)}$$

- We **cannot** normalise wrt the sum of weights and use the CLT, ie

$$\overline{\mu} \neq \frac{1}{N} \sum w_i' E_{L,i} \quad , \quad \overline{\sigma}^2 \neq \frac{1}{N.(N-1)} \sum \left(w_i' E_L(\mathbf{R}_i) - \overline{\mu}\right)^2$$

Because:

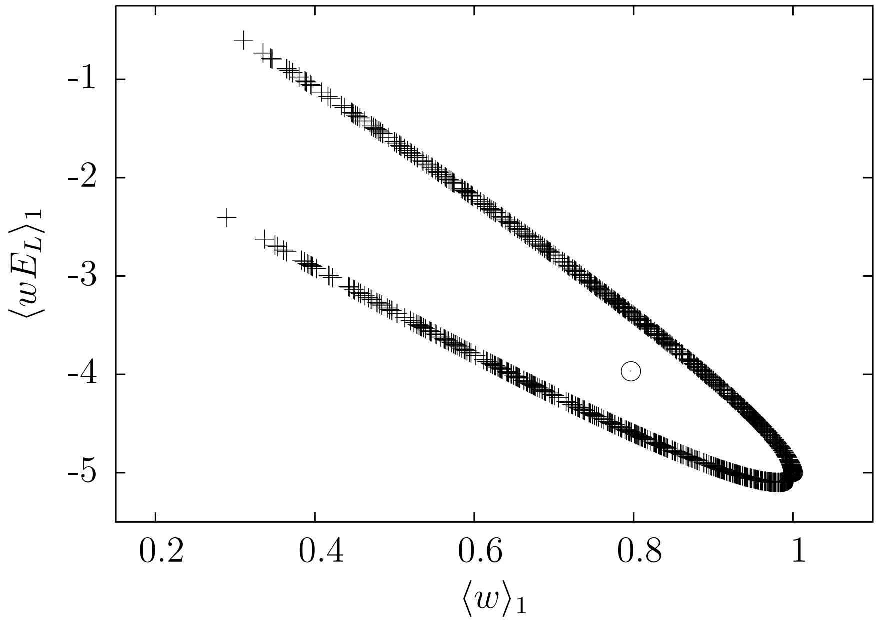
- CLT is true for sums of *independent*, *identically distributed* random variables
  - $w_1/(w_1+w_2)$ is correlated with $w_2/(w_1+w_2)$ $\Rightarrow$ not *independent*
  - $w_1/(w_1+w_2)$ has a different distribution to $w_1/(w_1+w_2+w_3)$ $\Rightarrow$ not *identically distributed*
- There is no reason for this to provide a good approximation

Trail JR, Phys. Rev. E. **77**, 016703,016704 (2008)

VMC and General Sampling

- What is distribution of $(\overline{wE_l}, \overline{w})$?
- *Bivariate Central Limit Theorem*

An example: N=1

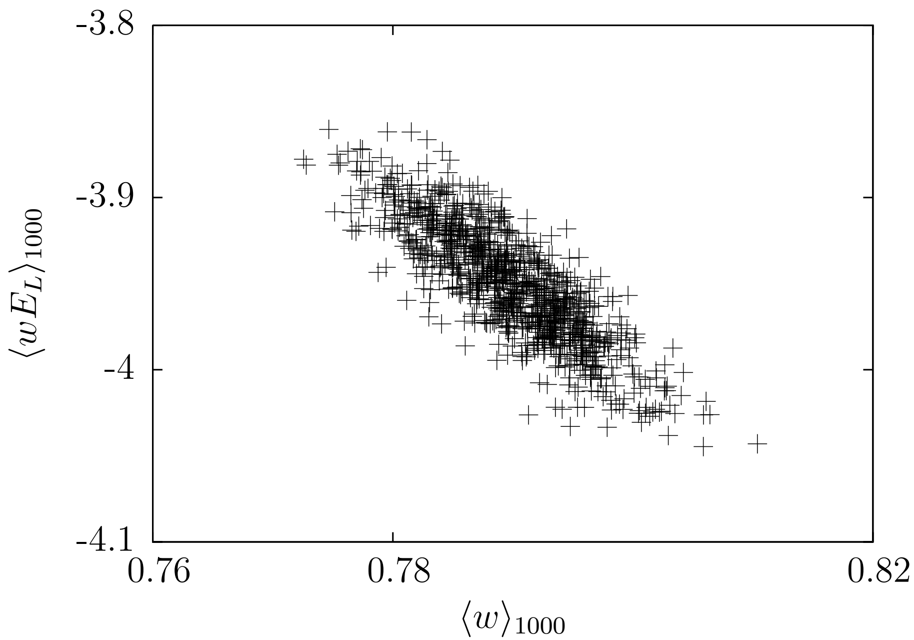


- 1000 estimates of $(\overline{wE_L}, \overline{w})$ each constructed from 1 sample of R

VMC and General Sampling

- What is distribution of $(\overline{wE_l}, \overline{w})$?
- *Bivariate Central Limit Theorem*

An example: N=1000



- 1000 estimates of $(\overline{wE_L}, \overline{w})$ each constructed from 1000 samples of R

$\longrightarrow$ Bivariate Normal distribution: mean is a **vector** + Covariance is a **matrix**

VMC and General Sampling

- Convert the distribution of (y, x) to a distribution of y/x using Fiellers theorem [a]
- $\text{Est}_r\left[E_{tot}\right]$ is distrbuted Normally with

$$\mu = \frac{\int \psi^2 E_L d\mathbf{R}}{\int \psi^2 d\mathbf{R}} \quad , \quad \sigma^2 = \frac{1}{N} \frac{\int \psi^2 / w d\mathbf{R} \int w \psi^2 (E_L - \mu)^2 d\mathbf{R}}{\left[\int \psi^2 d\mathbf{R}\right]^2}$$

- Estimates are available:

$$\overline{\mu} = \frac{\sum w_i E_L(\mathbf{R}_i)}{\sum w_i} \quad , \quad \overline{\sigma}^2 = \frac{N}{N-1} \frac{\sum w_i^2 \left(E_L(\mathbf{R}_i) - \overline{\mu}\right)^2}{\left(\sum w_i\right)^2}$$

- $\overline{\sigma}^2 \neq (\text{sample variance})/N$
- These equations do not follow from the usual (univariate) Central Limit Theorem
- *Zero Variance Principle* is still valid - for exact $\psi \Rightarrow \overline{\sigma} = 0$

$\longrightarrow$ The error is **controlled** if the bivariate CLT is **valid** and $\langle w \rangle \neq 0$

[a] We also require that the covariance is finite, $\langle w \rangle \neq 0$, and N is large enough

VMC and General Sampling

We already do generalised sampling:

- Correlated sampling in VMC optimisation
- Population control in weighted DMC

BUT we can choose w (equivalently P) specifically to improve performace and statistics:

- It changes the size of the error
- It can reinstate the CLT where it is invalid for standard sampling

VMC Total Energy estimate: standard sampling

- $P(E_L) \propto 1/x^4 \Rightarrow$ CLT is valid for local energy
- For correlated sampling CLT is not valid
- For most estimates, CLT is not valid
- Standard error, σ^2, is fixed for each system
- Can we improve on this?

Consider two possibilities:

1) Optimum sampling
2) Efficient Sampling

Optimum sampling

- What is the lowest statistical estimate possible for N samples?
- Minimise σ^2 wrt function w (or P)
- Solve $\frac{\delta\sigma^2}{\delta w} = 0$, where

$$\sigma^2 = \frac{1}{N} \frac{\int \psi^2 / w d\mathbf{R} \int w \psi^2 (E_L - \mu)^2 d\mathbf{R}}{\left[\int \psi^2 d\mathbf{R}\right]^2}$$

Optimum Sampling

- For given $(\psi, \hat{H}, N)$ lowest statistical error provided by

$$w = \frac{1}{|E_L - \mu|} \quad \text{or} \quad P_{opt} = \psi^2 |E_L - \mu|$$

- This gives the *optimum* error

$$\begin{aligned} \sigma_{opt} &= \frac{1}{N^{\frac{1}{2}}} \int \psi^2 |E_L - E_{tot}| d\mathbf{R} \\ &= \text{MAD}/N^{\frac{1}{2}} \end{aligned}$$

- Compare with *standard sampling* error

$$\begin{aligned} \sigma_{std} &= \frac{1}{\sqrt{N}} \left[\int \psi^2 (E_L - E_{tot})^2 d\mathbf{R} \right]^{1/2} \\ &= \text{S.D.}/N^{\frac{1}{2}} \end{aligned}$$

- For any calculation we can estimate a lower limit for the error
- Non-statistical estimates can have higher accuracy (eg one sample at $E_l = E_{tot}$)
- Cannot use $\mu \approx \overline{\mu}$ (CLT becomes invalid)

Optimum Sampling

- Use a random estimate of μ
- Normally distributed with mean,variance E_0, ϵ^2
- Minimise the mean value of σ^2

$$w = \frac{1}{\left[(E_L - E_0)^2 + \epsilon^2\right]^{\frac{1}{2}}} \quad \text{or } P_{opt} = \psi^2 \left[(E_L - E_0)^2 + \epsilon^2\right]^{\frac{1}{2}}$$

- (E_0, ϵ) does not bias estimates
- (E_0, ϵ) does not have to be accurate
- $\epsilon \longrightarrow \infty$ gives standard sampling
- Good starting values are $(E_{HF}, E_{HF}/10)$

Efficient Sampling

- Often the wavefunction is complex and involves many flops to evaluate
- Markovian chain using Metropolis algorithm has long correlation times
- Expensive for complex wavefunctions/long correlation times (eg atoms)
- Less expensive for simple wavefunctions/short correlation times (eg HEG)

$\longrightarrow$ Reduce computational cost of random walk *between* samples of E_L

Efficient Sampling

- Choose a simplified distribution, P_{sim} by excluding Jastrow, Backflow, Multideterminents . . .
- Make sure the CLT remains valid for the accompanying total energy estimate

Example: Use a HF determinant, with an arbitrary power:

$$P_{sim} = |D_0(\mathbf{R})|^p$$

Analysis of the distribution at the nodal surface:

$P_{sim}(E_L) \sim 1/x^{2+1/p} \Rightarrow$

- CLT invalid for $p \geq 1$
- error increased by an order of magnitude for $p < 1$

. . . not good enough

Efficient Sampling

Desirable features of P_{sim}:

- $\mathsf{Est}_r\,[E_{tot}]$ is Normal
- P_{sim} is computationally cheap
- P_{sim} that is not too far from optimum
- Reproduces exponental tails of ψ^2
- Has no nodal surface

Efficient Sampling

Final choice:

$$P_{sim} = |D_0|^2 + |D_1|^2$$

- No singularites introduced in averaged quantity $\longrightarrow$ CLT is valid
- Cheap to calculate (few determinants, no Jastrow, no Backflow)
- Accurate tails
- $P_{sim} \neq 0$ on nodal surface
- $P_{sim} = 0$ on coalescence planes only

$\mathrm{E_{VMC}}$ for an isolated O atom

24h on processors desktop:

Sampling	$E/(a.u.)$	N
std	-75.0610(3)	2,246,400
opt	-75.0610(7)	230,400
sopt	-75.0607(1)	19,968,000
sim	-75.06058(5)	78,720,000

- Efficient sampling reduces error by $\times\frac{1}{7}$
- Reduces cpu-hours by $\times\frac{1}{49}$
- Equivalent to a Moore's-law-timespan of 8 years

Efficient optimisation

- $E_{VMC} = E_0 + \epsilon_{VMC} + \epsilon_{opt}$
- Details of particular methods unimportant - we use matrix energy minimisation
- Draw a random curve from a 'random curve generator' with initial wavefunction parameters $\{\alpha_{init}\}$
- Find an improved set of parameters $\{\alpha_{min}\}$ - a sample value of a random variable
- Iterate...
- What is the random error due to the random position of the minimum?

Efficient optimisation

Exact curve, with minimum at α_0

$$f = f^0 + \frac{1}{2} f^2 (\alpha - \alpha_0)^2 + \ldots$$

Available is a random curve (ie $\langle \mathsf{f} \rangle = f$)

$$\mathsf{f} = \mathsf{f}^{(0)} + \mathsf{f}^{(1)}.(\alpha - \alpha_0) + \frac{1}{2} \mathsf{f}^{(2)} (\alpha - \alpha_0)^2 + \ldots$$

- Random minimum at $\mathsf{a}_0 = \alpha_0 - \frac{\mathsf{f}^{(1)}}{\mathsf{f}^{(2)}}$
- a_0 is Normal if $\left(\mathsf{f}^{(1)}, \mathsf{f}^{(2)}\right)$ are bivariate Normal

$$\epsilon_{opt} = \frac{1}{2} \left[\frac{\mathsf{f}^{(1)}}{\mathsf{f}^{(2)}} \right]^2 \langle \mathsf{f}^{(2)} \rangle$$

- ϵ_{opt} distributed as square of Normal random variable

Efficient optimisation

For N_{param} parameters, diagonalise $\mathsf{f}^{(2)}$

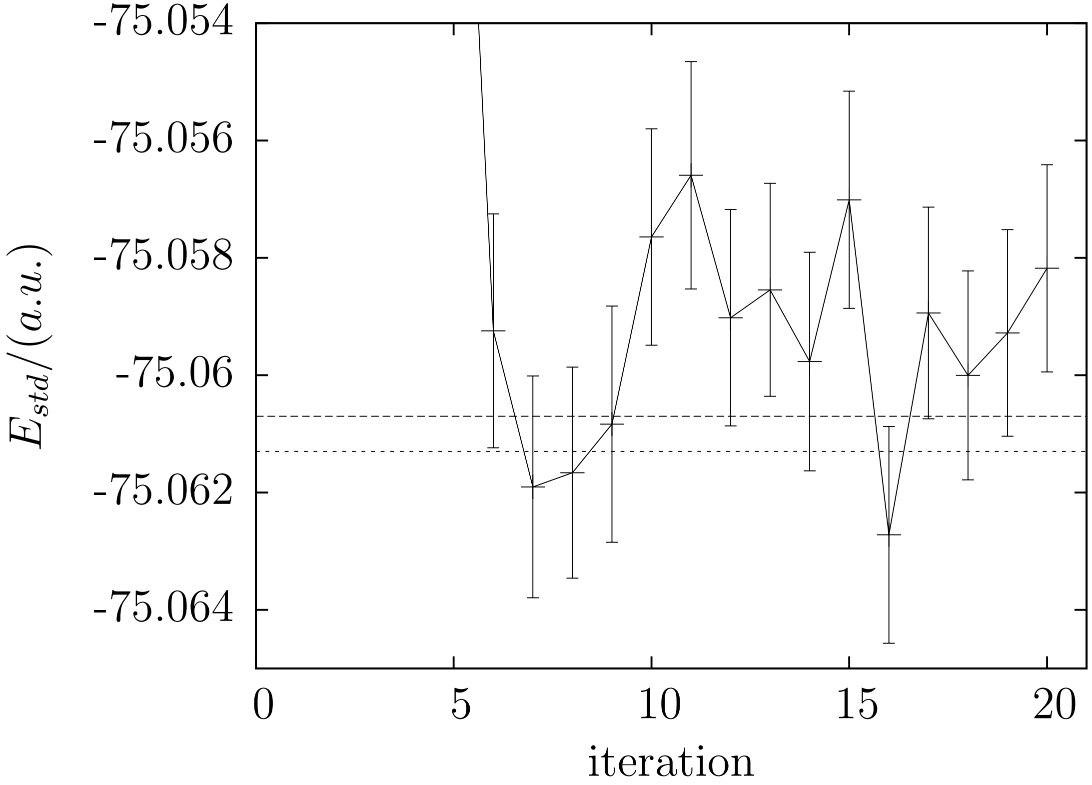
- $\chi^2_{N_{param}}$ distribution of order N_{param}
- $\longrightarrow$ normal for large N_{param}
- Mean $\propto N_{param}/N$
- Variance $\propto (N_{param}/N)^2$

$\longrightarrow$ Offset error proportional to number of parameters in the trial wavefunction

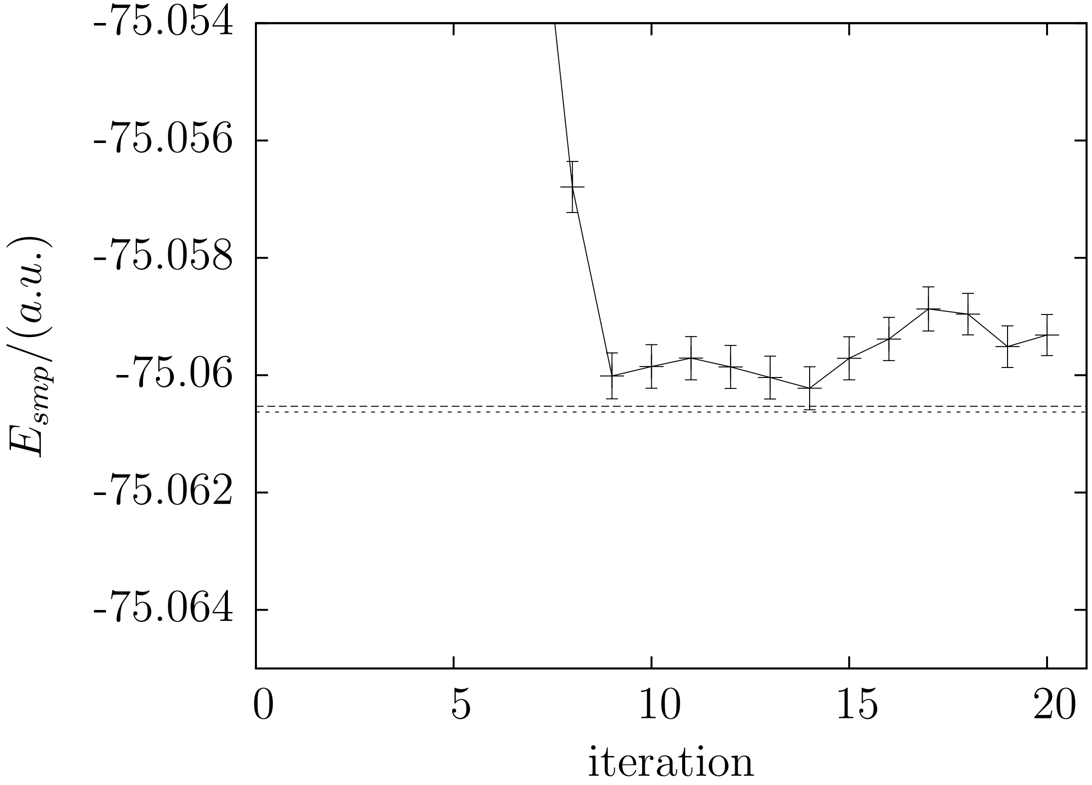
$\longrightarrow$ Requires $\left(\mathsf{f}^{(1)}, \mathsf{f}^{(2)}\right)$ to be normal - not true for standard sampling

So . . . non-standard sampling + average parameters

Standard optimisation - O atom, 24h



Efficient optimisation - O atom, 24h



1st row atoms

- ~ 500 dets. Jastrow+Backflow
- 24h runtime

	E_{VMC}	E_{VMC}(prev)	Exact
Li	-7.478052(2)	-7.47799(1)	-7.47806032
Be	-14.667243(3)	-14.66716(2)	-14.66736
B	-24.65329(1)	-24.65254(4)	-24.65391
C	-37.84361(2)	-37.84199(7)	-37.8450
N	-54.58641(4)	-54.5840(1)	-54.5892
O	-75.06058(5)	-75.0566(2)	-75.0673
F	-99.72623(8)	-99.7220(2)	-99.7339
Ne	-128.9299(1)	-128.9246(4)	-128.9376

(prev) Brown MD *et al.* J. Chem. Phys. **126**, 224110 (2007)

1st row diatomic molecules

- ~ 100 dets., numerical orbitals, Jastrow+Backflow
- 0.5h runtime

	E_{VMC}	E_{VMC}(prev)	Exact
Li2	-14.9839(2)	-14.99229(5)	-14.9951
C2	-75.881(1)	-75.8862(2)	-75.9265
N2	-109.494(2)	-109.4851(3)	-109.5421
Ne2	-257.854(3)	-257.80956(2)	-257.8753

(prev) Toulouse F and Umrigar CJ, J. Chem. Phys. **1**28, 174101 (2008)

Conclusions

- $P = \psi^2$ is an ad-hoc choice
- This choice introduces singularities and non-Normal distributions that don't have to be there
- Other P is possible
- *Optimum* and *efficient* choice can be made that improve on the standard method
- A simpler P can provide a Normal error for all estimates
- A simpler P can allow considerably larger sample sizes