Finite-temperature properties of strongly correlated systems via variational Monte Carlo

Jahan Claes and Bryan K. Clark

Finite temperature properties of strongly correlated systems via variational Monte Carlo

Jahan Claes[1] and Bryan K. Clark[1]

[1]*Department of Physics, University of Illinois at Urbana-Champaign*

Variational methods are a common approach for computing properties of ground states but have not yet found analogous success in finite temperature calculations. In this work we develop a new variational finite temperature algorithm (VAFT) which combines ideas from minimally entangled typical thermal states (METTS), variational Monte Carlo (VMC) optimization and path integral Monte Carlo (PIMC). This allows us to define an implicit variational density matrix to estimate finite temperature properties in two and three dimensions. We benchmark the algorithm on the bipartite Heisenberg model and compare to exact results.

Strongly correlated fermionic and frustrated spin systems span many interesting physical systems. Computing properties of these systems is difficult, particularly in dimensions greater then one where density matrix renormalization group (DMRG) [1] methods are not particularly effective, and the fermion sign problem [2] renders exact quantum Monte Carlo methods such as Path Integral Monte Carlo (PIMC) [3, 4] exponential. However, for ground state properties, there has been significant progress using variational Monte Carlo (VMC) methods which, while approximate, often give qualitatively and quantitatively detailed information [5, 6]. In VMC, one optimizes for the lowest energy state over a set of variational wave-functions $|\Psi[\vec{\alpha}]\rangle$ parameterized by $\vec{\alpha}$. The approximation improves as the set of variational wave-functions expands to include wavefunctions that have greater overlap with the true ground state $|\Psi_0\rangle$, becoming exact in the limit where $|\Psi_0\rangle$ is included in the space. There exist various forms of variational wave-functions including the Slater-Jastrow [7], backflow [8, 9], Huse-Elser [10] (equivalently correlator product states [11], entanglement plaquette states [12], or graph tensor network states [13]), BDG states, projected entangled pair states [14, 15], etc. At finite temperature, variational techniques have been less useful, except again in one-dimension where minimally entangled typical thermal states (METTS) [16, 17] and finite-temperature DMRG [18] have proved valuable.

Of course, it is also possible to use the variational technique at finite temperature. Most naturally one might parameterize the finite temperature many body density matrix (FTDM) $\hat{\rho} \equiv \exp[-\beta\hat{H}]/\mathrm{Tr}(\exp[-\beta\hat{H}])$ as $\hat{\rho}[\vec{\alpha}]$, optimizing again over the parameters $\vec{\alpha}$. However, this requires specifying a set of FTDM over which to optimize, a task that is naturally more difficult than specifying a set of wavefunctions. Additionally, this optimization needs to minimize the free energy of the system, a quantity which is harder to evaluate, requiring techniques such as coupling constant integration. For one example of this approach, see ref. [19].

Producing a good variational ansatz for very high or low temperature density matrices is straightforward.

At high temperature (small β), one can approximate $\exp[-\beta\hat{H}]$ by either Taylor expanding as $(1 - \beta\hat{H})$ or using the Trotter break-up. At low temperature, one can instead expand in terms of the low-energy excitations, writing $\hat{\rho} = \sum_{i=1}^{k} \exp[-\beta E_i]|\Psi_i\rangle\langle\Psi_i|$ where $|\Psi_i\rangle$ and E_i are variational estimates for the ith eigenvector and eigenvalue respectively and k is some cutoff. Unfortunately, this can be difficult, requiring orthogonal variational estimates of a number of excited states, and as one goes to higher temperature the number of eigenstates increases exponentially. Nonetheless, this framework is often applied in the context of density functional theory (DFT) to produce finite temperature functionals [20].

At any temperature, one can approximate the density matrix as a purification $\hat{\rho} \approx \sum_i \lambda_i|\Psi_i\rangle\langle\Psi_i|$ over a set $|\Psi_i\rangle$ of variational wavefunctions. However, one cannot efficiently enumerate more than a polynomial number of $|\Psi_i\rangle$, nor is optimizing these wavefunctions straightforward. See ref [21] for an example of working in a truncated Hilbert space of ten such wave-functions.

In spite of these difficulties, it would be useful if we could use known classes of variational wave-functions to compute finite temperature properties. In this paper, we describe a new variational finite temperature approach (VAFT), in which we implicitly define a variational finite temperature density matrix. Our variational FTDM is a purification of an *exponentially* large number of variational wavefunctions. We compute properties of this FTDM via an algorithm which stochastically samples from its diagonal without ever needing to explicitly represent it. The set of variational wavefunctions need not accurately represent highly excited states, instead needing only to accurately represent imaginary time propagation of many-body basis elements.

VAFT combines ideas from METTS, VMC, and PIMC to produce this stochastic sample. Our algorithm can be viewed either as an approximate generalization of METTS to wave-functions beyond matrix product states (MPS) or a modification of two-bead PIMC which replaces the stochastic evolution over imaginary time with an approximate deterministic one. In fact, we will see that, in this limit, METTS and PIMC are very similar.

The overall outline of this paper is as follows. We first describe an idealized purification of the FTDM. We then introduce a variational approximation to the FTDM which is a sum over an exponential number of variational wavefunctions, and explain how to efficiently sample from the diagonal of our variational FTDM and calculate finite temperature observables. Finally, we test the VAFT algorithm using Huse-Elser variational states on the Heisenberg model, and compare to exact results. For completeness, we describe all the necessary steps for our algorithm even when they have partial overlap with methods that already exist in the literature.

Algorithm: In what follows, $\{|c\rangle\}$ is some many-body basis for our Hilbert space, and $\hat{U}$ is a unitary operator that satisfies $[\hat{U}, \hat{H}] = 0$. Most commonly, $\{|c\rangle\}$ is a basis of product states. $\hat{U}$ may be thought of as a change of basis that respects the symmetries of the Hamiltonian. For simplicity, the reader may take $\hat{U} = \mathbf{1}$, although we will see that other choices of $\hat{U}$ can be used to mitigate various technical issues. - We want to represent $\hat{\rho}$ as a purification over a set of wavefunctions. One way to do this is to write

$$
\hat{\rho} = \frac{e^{-\beta \hat{H}}}{\text{Tr}(e^{-\beta \hat{H}})}
$$

$$
= \sum_c e^{-\beta \hat{H}/2} \hat{U} |c\rangle \langle c| \hat{U}^\dagger e^{-\beta \hat{H}/2} \frac{1}{\text{Tr}(e^{-\beta \hat{H}})}
$$

$$
= \sum_c \frac{e^{-\beta \hat{H}/2} \hat{U} |c\rangle \langle c| \hat{U}^\dagger e^{-\beta \hat{H}/2}}{\langle c| e^{-\beta \hat{H}} |c\rangle} \frac{\langle c| e^{-\beta \hat{H}} |c\rangle}{\text{Tr}(e^{-\beta \hat{H}})}
$$

$$
= \sum_c \frac{|\tilde{\Psi}[\beta/2; c]\rangle \langle \tilde{\Psi}[\beta/2; c]|}{\langle \tilde{\Psi}[\beta/2; c]|\tilde{\Psi}[\beta/2; c]\rangle} \tilde{p}(c)
$$

where in the last line we let $\tilde{p}(c) = \langle c|\hat{\rho}|c\rangle$ be the diagonal of the FTDM, and $|\tilde{\Psi}[\beta/2; c]\rangle$ be the (unnormalized) wavefunction $\exp[-\beta \hat{H}/2]\hat{U}|c\rangle$.

Unfortunately, there is no obvious polynomial algorithm to compute either $|\tilde{\Psi}[\beta/2; c]\rangle$ or $\tilde{p}(c)$. Instead, in this paper we work with approximate versions of these quantities $|\Psi[\beta/2; c]\rangle$ and probability distribution $p(c)$.

We define our approximation $|\Psi[\beta/2; c]\rangle$ as

$$
|\Psi[\beta/2; c]\rangle \equiv [\hat{P}(1 - \tau \hat{H})]^{\beta/(2\tau)} \hat{U}|c\rangle
$$

where the operator $\hat{P}$ projects the wavefuction into a variational space defined by the set of wave-functions $|\Psi[\vec{\alpha}]\rangle$. Notice that $|\Psi[\beta/2; c]\rangle \approx |\tilde{\Psi}[\beta/2; c]\rangle$ and approaches it exactly in the limit of a large enough variational space and small enough τ.

We can compute $|\Psi[\beta/2; c]\rangle$ via (projected) imaginary time propagation by applying the operator $\hat{P}(1 - \tau \hat{H})$ repeatedly, $\beta/(2\tau)$ times. At each application, we select a new wave-function from the variational space by selecting parameters $\vec{\alpha}'$ such that the overlap

$$
\frac{\langle \Psi[\vec{\alpha}]|(1 - \tau H)|\Psi[\vec{\alpha}']\rangle}{\sqrt{\langle \Psi[\vec{\alpha}']|\Psi[\vec{\alpha}']\rangle}}
\tag{1}
$$

is maximized. Eqn 1 is maximized by choosing $\vec{\alpha}'$ such that

$$
\vec{\alpha}' = \vec{\alpha} - \tau \mathbf{S}^{-1} \vec{h}
\tag{2}
$$

where, defining $|\Psi_i\rangle \equiv \frac{\partial}{\partial \alpha_i} |\Psi\rangle$, we have

$$
h_i = \frac{\langle \Psi|\hat{H}|\Psi_i\rangle}{\langle \Psi|\Psi\rangle} - \frac{\langle \Psi|\hat{H}|\Psi\rangle}{\langle \Psi|\Psi\rangle} \frac{\langle \Psi|\Psi_i\rangle}{\langle \Psi|\Psi\rangle}
\tag{3}
$$

$$
(\mathbf{S})_{ij} = \frac{\langle \Psi_j|\Psi_i\rangle}{\langle \Psi|\Psi\rangle} - \frac{\langle \Psi_j|\Psi\rangle}{\langle \Psi|\Psi\rangle} \frac{\langle \Psi|\Psi_i\rangle}{\langle \Psi|\Psi\rangle}
\tag{4}
$$

Each of these quantities can be evaluated via Monte Carlo (see Appendix). This process is identical to the one used in the stochastic reconfiguration optimization method [22].

We implicitly define our approximate probability $p(c)$ as the stationary distribution of a Markov chain whose state space is the many-body basis elements $|c\rangle$ and whose transition rule is defined by

$$
\text{Pr}(c \to c') \propto |\langle \Psi[\beta/2; c]|c'\rangle|^2
$$

We can now write a concrete algorithm to do a walk over this Markov chain. To take a Markov step from $|c\rangle \to |c'\rangle$, we

- Write $\hat{U}|c\rangle$ as $|\Psi[\vec{\alpha}_0]\rangle$.

- Use projected imaginary time propagation to compute $|\Psi[\beta/2; c]\rangle$

- Use variational Monte Carlo to select a configuration $|c'\rangle$ from the distribution $|\langle c'|\Psi[\beta/2; c]\rangle|^2$.

Note that this Markov chain may alternately be viewed as a walk over the space of wavefunctions $|\Psi[\beta/2; c]\rangle$ where $\text{Pr}(|\Psi[\beta/2; c]\rangle \to |\Psi[\beta/2; c']\rangle) \propto |\langle \Psi[\beta/2; c]|c'\rangle|^2$. The Markov chain used in our approach, restricted to the space of MPS states and without the projection or operator $\hat{U}$, is the one used in METTS; there, though, steps on the Markov chain are implemented differently. Notice this is a legitimate Markov chain for any set of wave-functions as it is memoryless and therefore is guaranteed to reach its stationary distribution $p(c)$ in the long term limit. In the case where $|\Psi[\beta/2; c]\rangle$ is exactly equal to $|\tilde{\Psi}[\beta/2; c]\rangle$, this Markov chain has the stationary distribution $\tilde{p}(c)$ (see Appendix for proof). Thus, when $|\Psi[\beta/2; c]\rangle \approx |\tilde{\Psi}[\beta/2; c]\rangle$, we expect $p(c) \approx \tilde{p}(c)$.

It is worth stepping back a moment and recognizing what we've achieved. Starting with some variational space defined by $|\Psi[\vec{\alpha}]\rangle$, we've implicitly defined an approximate FTDM that is a purification over an exponentially large number of variational wavefunctions.

$$
\hat{\rho} = \sum_c \frac{|\Psi[\beta/2; c]\rangle \langle \Psi[\beta/2; c]|}{\langle \Psi[\beta/2; c]|\Psi[\beta/2; c]\rangle} p(c)
$$

Note that this is still a valid density matrix, in that $\hat{\rho}$ is Hermitian and satisfies $\text{Tr}(\hat{\rho}) = 1$. In the limit where our space includes all wave-functions of the form $\exp[-\tau\hat{H}/2]\hat{U}|c\rangle$ for all $|c\rangle$ in the basis and all $0 \leq \tau \leq \beta$, this reduces to the exact FTDM, in the same spirit that variational Monte Carlo becomes exact when the space includes the true ground state.

This purification is too large to explicitly enumerate. However, we can nonetheless compute the wavefunctions $|\Psi[\beta/2; c]\rangle$ as well as sample from $p(c)$ by doing a random walk over the Markov chain described above; as we'll see below, this is all that is necessary for computing finite temperature properties of this density matrix.

Note that, unlike in ground state VMC, there is no optimization step; the approximate FTDM is naturally close to the actual FTDM, and the only adjustment that can be made to our approximation is changing the underlying variational space. Where ground state VMC takes a set of variational wavefunctions as input and optimizes to select the best one, the VAFT algorithm takes a set of variational wavefunctions as input and produces a density matrix that is a weighted linear combination of all of them. (Nonetheless, we will see later that we can still choose the best set of wave-functions among many such sets).

Up to reasonable assumptions about mixing times, this sampling is polynomial and therefore scales well with system size. The VMC steps involved in the projected imaginary propagation, the VMC in sampling the new $|c'\rangle$, and the overall Markov chain can be readily parallelized. The VMC steps can be parallelized by distributing the VMC walkers across multiple processors, while the overall Markov chain can be run independently on multiple processors in order to obtain independent samples of $p(c)$. Since the main aspects of the algorithm are applying $\hat{P}(1 - \tau\hat{H})$ to a variational wavefunction and sampling from a variational wavefunction, working at finite temperature requires minimal modifications to an existing VMC code that does optimization using the stochastic reconfiguration method.

In fig 1, we plot the energy of the state as our algorithm runs. There are two things going on here. The overall Markov chain takes place over the basis elements $\hat{U}|c\rangle$ denoted by the green circles, or equivalently over the wavefunctions $|\Psi[\beta/2; c]\rangle$ denoted by the red squares. To get from the green circles to the red squares, we use projected imaginary time propagation (analogous to stochastic reconfiguration optimization) to apply $\hat{P}(1 - \tau\hat{H})$ repeatedly. As we apply $\hat{P}(1 - \tau\hat{H})$, the energy steadily decreases, until we reach the red squares, where the state has been evolved for an imaginary time of $\beta/2$. Note that while we use Monte Carlo to apply the time evolution, the time evolution is, up to statistical noise, deterministic. We calculate observables at the red squares (see below). For example, to calculate the energy, we average the energies of the wavefunctions at the red squares; this
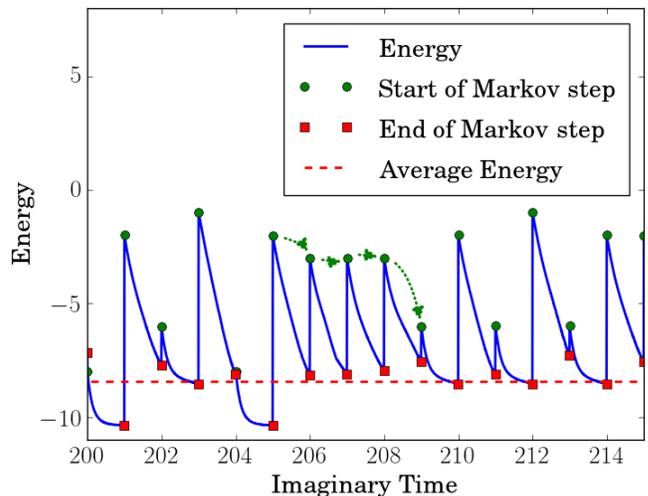

Figure 1: Energy as a function of imaginary time for a 4×4 square Heisenberg model, $\beta = 1$, using nearest neighbor Huse-Elser states (see below). The green circles indicate the beginning of each Markov step, while the red squares indicate the end of each Markov step. The arrows show the progression of the Markov chain from one step to the next. Each Markov step lasts $\beta/2 = .5$. Observables are calculated at the red circles. The dashed red line indicates the average energy of the system.

average energy is shown as the dashed line. After each imaginary time propagation, we select a new $|c\rangle$ from our wavefunction, which causes the energy to jump discontinuously to the next green circle.

There are a number of approximations in our simulation. The fundamental approximation comes from assuming some variational space of wavefunctions. There are two additional systematic approximations. First, we have introduced a time step error τ coming from approximating $\exp[-\tau\hat{H}] \approx (1 - \tau\hat{H})$. Second, we have estimated the values of $(\mathbf{S})_{ij}$ and h_j using VMC, so these quantities have statistical error associated with them which go away as the Monte Carlo run becomes arbitrarily long.

Observables: To compute the finite temperature expectation value of an observable $\hat{\mathcal{O}}$ in our algorithm, we compute $\text{Tr}(\hat{\rho}\hat{\mathcal{O}})$ using our approximation to the FTDM:

$$\langle\hat{\mathcal{O}}\rangle = \text{Tr}(\hat{\rho}\hat{\mathcal{O}}) = \sum_c \langle\hat{\mathcal{O}}\rangle_c p(c)$$

where

$$\langle\hat{\mathcal{O}}\rangle_c \equiv \frac{\langle\Psi[\beta/2; c]|\hat{\mathcal{O}}|\Psi[\beta/2; c]\rangle}{\langle\Psi[\beta/2; c]|\Psi[\beta/2; c]\rangle}$$

is simply the expectation value of $\hat{\mathcal{O}}$ in the state $|\Psi[\beta/2; c]\rangle$. Thus, to compute $\langle\hat{\mathcal{O}}\rangle$, we simply average the expectation values $\langle\hat{\mathcal{O}}\rangle_c$ across the samples

$|\Psi[\beta/2; c]\rangle$ generated by our Markov chain. To compute $\langle\hat{\mathcal{O}}\rangle_c$ we use the Metropolis algorithm [23] to sample basis elements $|c'\rangle$ from $|\langle c'|\Psi[\beta/2; c]\rangle|^2$, then average $\langle c'|\hat{\mathcal{O}}|\Psi[\beta/2; c]\rangle/\langle c'|\Psi[\beta/2; c]\rangle$ over these samples. We can compute both diagonal and off-diagonal observables with this method; for example, in fig 1, we've computed the expectation value of the off-diagonal observable $\hat{H}$.

In the special case where $\hat{\mathcal{O}}$ is diagonal in the basis $\{|c\rangle\}$, we can also write

$$\langle\hat{\mathcal{O}}\rangle = \sum_c \langle c|\hat{\mathcal{O}}|c\rangle p(c)$$

Thus, rather than averaging the $\langle\hat{\mathcal{O}}\rangle_c$ over the samples, we may also estimate $\langle\hat{\mathcal{O}}\rangle$ by averaging $\langle c|\hat{\mathcal{O}}|c\rangle$ over the samples. However, this latter estimate for $\hat{\mathcal{O}}$ will generally have greater variance.

Low Temperature Limit: While our algorithm is designed to compute finite-temperature properties, we can consider whether it smoothly interpolates to the ground state density matrix $|\Psi[\vec{\alpha}_g]\rangle\langle\Psi[\vec{\alpha}_g]|$ where $|\Psi[\vec{\alpha}_g]\rangle$ is the wave-function in our variational space with lowest energy. It is straightforward to see that if $|\Psi[\vec{\alpha}_g]\rangle$ can be reached from any configuration $\hat{U}|c\rangle$ via projected imaginary time evolution, then this variational ground state limit emerges naturally. At low temperature $(\beta \to \infty)$ any initial basis configuration $\hat{U}|c\rangle$ will likely have overlap with the ground state. Thus, $\hat{U}|c\rangle$ will propagate down to the variational ground state in one application of $\exp[-\beta\hat{H}/2]$. However, depending on the variational space, it may be the case that the ground state can't be reached from all basis elements $\hat{U}|c\rangle$; instead the imaginary time evolution $\exp[-\beta\hat{H}/2]\hat{U}|c\rangle$ may get stuck in local energy minima. In such a case, our algorithm does not cleanly extrapolate to the variational limit and must have worse average energy (although it is less obvious whether other properties will be better or worse). As the variational space gets sufficiently large, however, this difference disappears.

Ansatz: The VAFT algorithm is general and can work with any variational space $\Psi[\vec{\alpha}]$, provided it includes the product states $\{\hat{U}|c\rangle\}$. The technically challenging part of the algorithm involves computing the imaginary time propagation step. There are two technical pitfalls in this implementation that must be considered when implementing VAFT in a new variational space: linear dependence of the derivatives, and undersampling in the VMC.

A single imaginary-time step is propagated in the subspace of derivatives $\{|\Psi_i\rangle\}$. These derivatives may form a linearly dependent set causing $\mathbf{S}^{-1}$ to be singular. Even when the $\{|\Psi_i\rangle\}$ are only approximately linearly dependent, the time evolution may fail; in this case, $\mathbf{S}^{-1}$ will have very large entries, which causes the parameters $\vec{\alpha}$ to change by a large amount (see eqn 2). Since the propagation is based on first-order approximations to $\delta\vec{\alpha}$, we

may have to choose a very small τ for the approximation to still be valid. This can be formally resolved in various ways such as dropping parameters or using a pseudo-inverse. It should be pointed out that in the case where we have significant linear dependence the approximate projected imaginary time evolution may be less accurate because of the lower-dimensional space.

For most ansatz, we find this linear dependence to be most severe near a product state, $|c\rangle$, where many of the derivatives $|\Psi_i\rangle$ become identically zero. To solve this linear dependence, one can choose a $\hat{U}$ such that $\hat{U}|c\rangle$ has all nonzero derivatives.

Undersampling during the VMC generally causes the estimates of $(\mathbf{S})_{ij}$ to not converge properly. To estimate $\mathbf{S}_{ij}$, we use VMC to estimate the overlap $\langle\Psi_j|\Psi_i\rangle/\langle\Psi|\Psi\rangle$ according to (see Appendix)

$$\frac{\langle\Psi_j|\Psi_i\rangle}{\langle\Psi|\Psi\rangle} = \frac{\sum_c \frac{\Psi_j(c)}{\Psi(c)}\frac{\Psi_i(c)}{\Psi(c)}|\Psi(c)|^2}{\sum_c |\Psi(c)|^2}$$

We compute this sum by sampling basis elements $|c\rangle$ from $|\Psi\rangle$. However, if there exist $|c\rangle$ such that $\Psi(c) \approx 0$, but $\Psi_j(c)\Psi_i(c) \neq 0$, our sampling will essentially ignore this $|c\rangle$, even though it is important for estimating the overlap. Thus, we undersample relevant basis elements.

Undersampling, like linear dependence of derivatives, is most severe near a product state $|c\rangle$. In this case, every configuration $|c'\rangle \neq |c\rangle$ satisfies $\Psi(c') \approx 0$, though many of these $|c'\rangle$ are relevant for computing the overlap. There are two possible remedies to the undersampling problem. The classical approach is to sample using another probability distribution which doesn't undersample these configuration while modifying the measured properties to keep the integral the same. For example, one might sample $|\Psi\rangle + \sum_i a_i|\Psi_i\rangle$ rather than $|\Psi\rangle$. Alternatively, one can choose a $\hat{U}$ such that $\hat{U}|c\rangle$ has weight on all relevant basis elements. In both cases, we end up sampling from something far from a product state, as desired.

We will see an example below of choosing a $\hat{U}$ which solves both of these problems simultaneously for the Huse-Elser ansatz.

Example: As a test, we apply our algorithm to a Heisenberg antiferromagnet on a two-dimensional square lattice, with Hamiltonian

$$\hat{H} = \sum_{<i,j>} \vec{\hat{S}}_i \cdot \vec{\hat{S}}_j$$

where the sum runs over all nearest neighbor sites of the lattice. We sample from the grand canonical ensemble, in which the total S_z is allowed to fluctuate. The finite temperature properties of this model can be efficiently simulated by the stochastic series expansion (SSE) method[24]; thus, it provides an ideal test for our approximate algorithm. Note that while the SSE algorithm does not have a sign problem for this particular
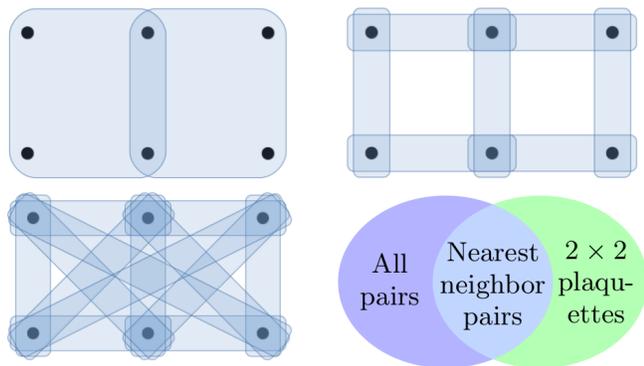
Figure 2: Three possible arrangements of correlators on a lattice. Each shaded bubble represents the support of one correlator. Counterclockwise from the upper right: Nearest neighbor pairs, 2×2 plaquettes, all pairs. Bottom right: The variational space defined by "nearest neighbor pairs" is a subset of both "all pairs" and "2×2 plaquettes", while "all pairs" and "2×2 plaquettes" are, naively, incomparable.

model, our algorithm never has a sign problem. We calculate the squared staggered magnetization, defined by

$$\hat{\mathcal{O}} = \left(\sum_i \sigma_i \vec{\hat{S}}_i \right)^2$$

with $\sigma_i = \pm 1$ alternating between adjacent sites. We also calculate the energy, and an estimate of the free energy.

For this example, we use Huse-Elser variational states [10]. If we let $c_i = \pm 1$ label the spin of the ith particle in our lattice, the most general state $|\Psi\rangle$ can be written as $\sum_{c_1,..,c_n} \Psi^{c_1,...,c_n} |c_1,...,c_n\rangle$, where $\Psi^{c_1,...,c_n}$ is an arbitrary set of complex coefficients. In a Huse-Elser variational space, $\Psi^{c_1,...,c_n}$ is restricted to be of the form

$$\Psi^{c_1,...,c_n} = \prod_j C_j^{c_{j_1},...,c_{j_k}}$$

where each $C_j^{c_{j_1},...,c_{j_k}}$ is an arbitrary complex number that depends only on some subset of the lattice $\{c_{j_1}, ..., c_{j_k}\}$. Each C_j is called a correlator, and the subset $\{c_{j_1}, ..., c_{j_k}\}$ is called the support of the correlator. We can specify a Huse-Elser variational space by specifying the supports of all the correlators. Three examples of possible correlator arrangements are shown in fig. 2. As we increase the number of wavefunctions in our variational space, by either increasing the number of correlators or increasing the size of the correlators, we will more accurately reproduce the imaginary time evolution.

The most natural basis $\{|c\rangle\}$ for the Huse-Elser states is the basis of $\hat{S}_z$ product states, since these basis elements can be sampled in the VMC. However, in this basis, we suffer from both undersampling and linearly dependent derivatives. In fact, for Huse-Elser states, when

$|\Psi\rangle = |c\rangle$, many of the $|\Psi_i\rangle$ are identically zero. We solve both of these problems by choosing $\hat{U}$ to be the operator that rotates $\hat{S}_z$ product states to $\hat{S}_x$ product states. This choice of $\hat{U}$ commutes with $\hat{H}$ since $\hat{H}$ is invariant under rotations, and it is clearly unitary. To be able to cleanly represent the states $\hat{U}|c\rangle$ as Huse-Elser states, we add a set of single site correlators to our correlator arrangements.

After this rotation no derivatives are zero, and $\hat{U}|\Psi\rangle$ has weight on all basis elements $|c\rangle$, so both problems are solved simultaneously. An additional effect of this rotation is to mix S_z spin sectors, so we are indeed sampling from the grand canonical ensemble.

We plot our results in figs 3 and 4. In fig 3, we plot the results for three different classes of Huse-Elser states on a 4×4 Heisenberg lattice, as well as the (essentially) exact result from SSE. In fig 4 we plot the results for one class of Huse-Elser states on an 8×8 lattice, and the corresponding SSE results. In both cases, it is necessary to use a smaller timestep τ at lower β. It is surprising to note that despite the restriction to a variational space, for $\beta \leq 1.4$, we accurately predict the staggered magnetization, while as β grows, our algorithm still produces a qualitatively similar curve as the true result. The divergence from the true result occurs because we can no longer represent $e^{-\tau \hat{H}}|c\rangle$ in our variational space for all $0 < \tau < \beta/2$. We also see in fig 3 that as we add wavefunctions to our variational space, the variational estimate approaches the true answer.

To find the best variational space, we can compare the free energies. Each run of our algorithm gives us an approximate density matrix $\hat{\rho}(\beta) = \sum_c |\Psi[\beta/2; c]\rangle\langle\Psi[\beta/2; c]|$, where the sum runs over the samples $|c\rangle$ from the Markov chain. The "best" density matrix should be the one that minimizes the free energy, $F(\beta) = E(\beta) - S(\beta)/\beta$. We don't compute the entropy $S(\beta)$ directly; instead, we estimate it from the $E(\beta)$ curve via the equation

$$S(\beta) = S(0) - \left[\int_0^\beta E(\beta')d\beta' - \beta E(\beta) \right]$$

This will not produce the exact entropy $S(\beta) = \text{Tr}(\hat{\rho}(\beta)\ln(\hat{\rho}(\beta)))$ of our approximate density matrices, but it will produce an approximation to $S(\beta)$ based on the entire E vs β curve. We will then get an approximation to $F(\beta)$ which also depends on the whole curve.

To estimate $F(\beta)$, we fit a curve of the form

$$E(\beta) = \frac{\sum_{i=1}^N D_i E_i e^{-E_i \beta}}{\sum_{i=1}^N D_i e^{-E_i \beta}}$$

to our energy data, where the E_i and D_i are parameters representing energies and densities of states, respectively. The results of this fit for the 4×4 lattice are shown in fig 5. The free energy systematically decreases as we increase the number of parameters in our variational space,
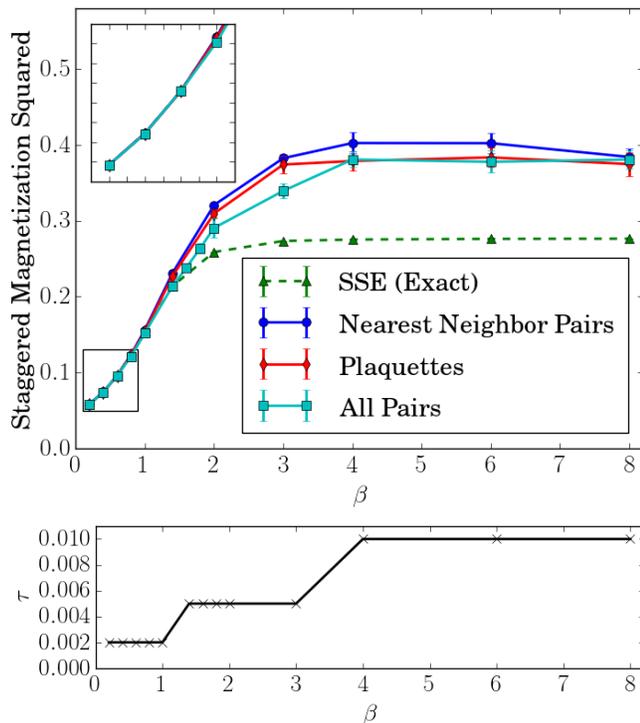
Figure 3: Above: Squared staggered magnetization as a function of β for a 4×4 square Heisenberg model. Inset: Enlarged version of the boxed region. Below: The timestep τ used for approximating $\exp[-\beta \hat{H}/2]$

Figure 4: Above: Squared staggered magnetization as a function of β for a 8×8 square Heisenberg model. Inset: Enlarged version of the boxed region. Below: The timestep τ used for approximating $\exp[-\beta \hat{H}/2]$

indicating that "all pairs" gives a better density matrix than "plaquettes", which gives a better density matrix than "nearest neighbor pairs." Note that this ordering is consistent with the fact that "plaquettes" and "all-pairs" are supersets of nearest-neighbor pairs. This approach allows us to compare our estimates, even in the absence of the exact SSE result.

Figure 5 also shows that for large β, our algorithm gives us an estimate of the ground state energy. However, this estimate is strictly worse than ground state VMC. This is because not every configuration $|c\rangle$ in our Markov chain is able to propagate down to the lowest energy state in the variational space. Instead, many get stuck in local minima. Thus, the estimate for the energy at large β is an average over local energy minima, and not the minimum energy wavefunction in our variational space.

PIMC and METTS: VAFT (when $\hat{U} = \mathbf{1}$) can either be viewed as a generalization of METTS to variational spaces besides matrix product states, or a modification of two-bead PIMC where one replaces the stochastic imaginary time evolution with a variational time evolution.

As mentioned previously, METTS uses the same Markov chain to sample from $p(c)$ as in our algorithm, with the restriction that the wavefunctions $|\Psi[\beta/2; c]\rangle$ are written as matrix product states. With this restriction,
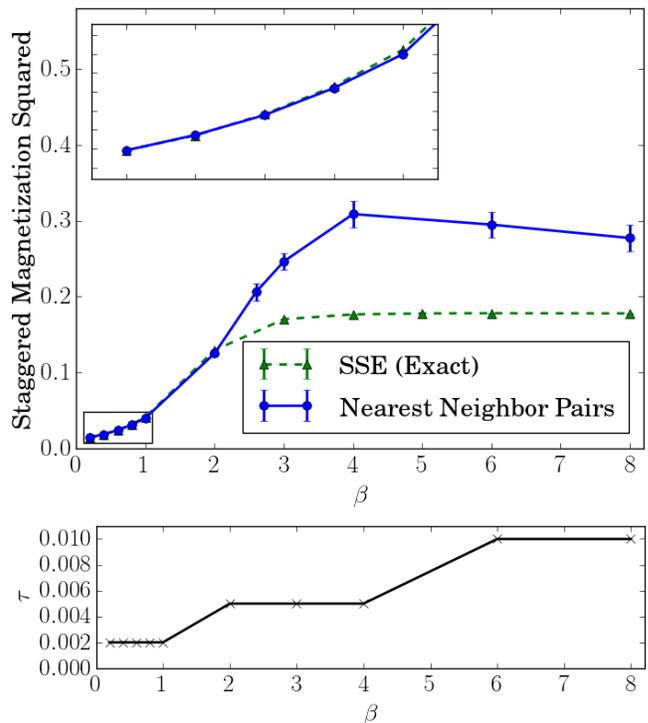
$(1 - \tau \hat{H})$ can be efficiently applied using DMRG techniques rather than Monte Carlo. Thus, our algorithm may be viewed as an extension of METTS to general variational states via VMC methods.

PIMC samples from $\tilde{p}(c)$ by rewriting $\text{Tr}(\exp[-\beta \hat{H}])$ as

$$\text{Tr}(\exp[-\beta \hat{H}]) = \sum_{\{c_i\}} \prod_i \langle c_i | \exp[-\tau \hat{H}] | c_{i+1} \rangle$$

where each i represents a bead on the path. PIMC works by sampling paths $\{c_i\}$ with appropriate weights via the Metropolis algorithm. In the case where we have only two beads, we write

$$\text{Tr}(\exp[-\beta \hat{H}]) = \sum_{c_1 c_2} |\langle c_1 | \exp[-\beta \hat{H}/2] | c_2 \rangle|^2$$

A local update that samples this distribution is

- Chose a new $|c_1\rangle$ with probability proportional to $|\langle c_1 | \exp[-\beta \hat{H}/2] | c_2 \rangle|^2$

- Chose a new $|c_2\rangle$ with probability proportional to $|\langle c_1 | \exp[-\beta \hat{H}/2] | c_2 \rangle|^2$

This is identical to our Markov chain, provided we calculate $|\langle c_1 | \exp[-\beta \hat{H}/2] | c_2 \rangle|^2$ using our variational approximation to $\exp[-\beta \hat{H}/2]$; thus, VAFT may be viewed as
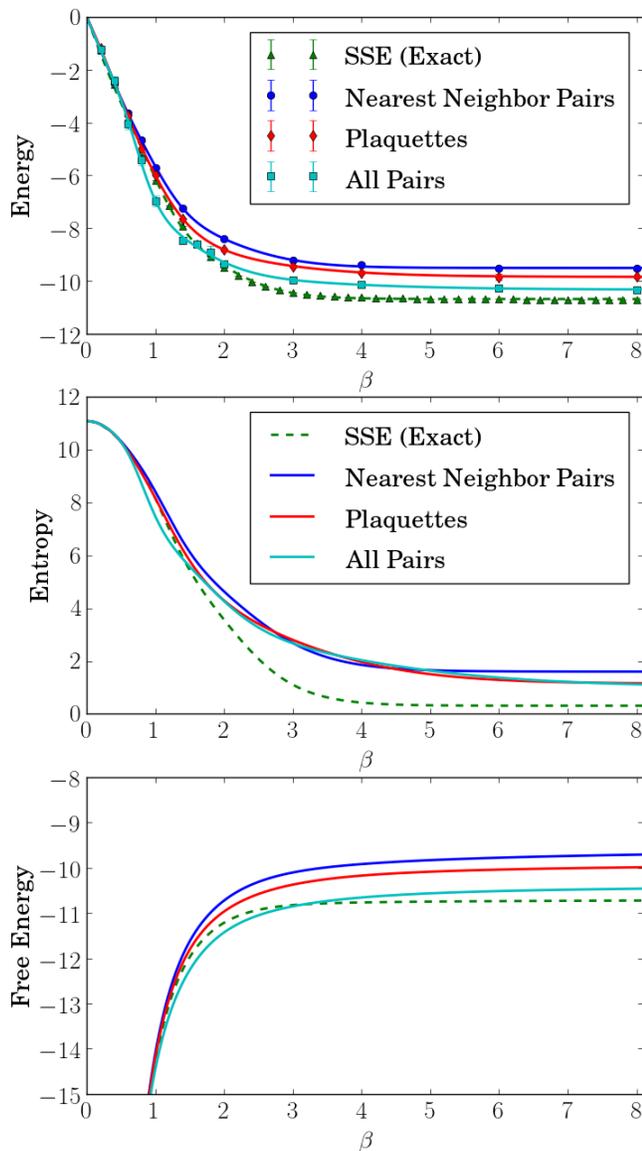
Figure 5: Energy, entropy, and free energy as a function of β for the 4×4 Heisenberg lattice. The free energy is lowest for the "all pairs" correlator arrangement, indicating this correlator arrangement gives the best estimate for the density matrix.

a two-bead PIMC in which $\exp[-\beta\hat{H}/2]$ is applied variationally.

We also see that in this point of view, METTS can be viewed as a two-bead PIMC in which $\exp[-\beta\hat{H}/2]$ is applied using matrix product states.

Discussion: While we benchmarked VAFT on the bipartite Heisenberg lattice, VAFT is general and can be applied in any setting, including continuum models, fermions, and frustrated magnets. In particular, since VAFT has no sign problem, it can be applied in cases where many other Monte Carlo methods fail. Variational

Monte Carlo has been the de-facto standard for understanding strongly-correlated ground states and VAFT allows this success to translate to finite temperature. Notice, that VAFT also can work with any ansatz.

While this work has been focused on a variational approach, it is interesting to note that the ideas presented here can be used to generate an alternative finite temperature projector QMC approach complementing approaches such as DMQMC [25–27] and finite temperature AFQMC [28, 29] as well as improve calculations in path-integral Monte Carlo. While the details of these methods will be outlined in upcoming works [30], here we mention some general aspects of the basic approach.

With respect to projector quantum Monte Carlo, one can replace the variational imaginary time evolution in VAFT with a stochastic time evolution. This replacement re-introduces a sign problem which then must be separately attenuated by, for example, annihilation in FCIQMC [31] or fixed-node/phase approaches in other flavors of QMC [32–34]. Such attenuation techniques have been very successful in the ground state. Generalizing VAFT to projector QMC requires additional effort in the sampling of $|\langle c|\Psi\rangle|^2$.

This approach may also be used to improve the restricted path integral Monte Carlo algorithm (RPIMC) [35, 36]. In RPIMC, one avoids the sign problem by assuming a sign structure for the density matrix $\hat{\rho}$. However, the results of RPIMC depend on accurately determining the sign structure. We can compute $\exp[-\beta\hat{H}]|R\rangle$ within our variational space for any configuration $|R\rangle$, therefore letting us evaluate the sign of the many body density matrix $\hat{\rho}(R, R'; \beta)$. This technique also has positive implications for the low-temperature ergodicity problems seen in PIMC.

Appendix: Here we prove several results stated without proof in the text. Namely, we show that the stationary distribution of the ideal Markov chain is $\tilde{p}(c)$, that eqn 1 is maximized by choosing $\vec{\alpha}'$ according to eqn 2, and that both h_i and $(\mathbf{S})_{ij}$ from eqns 3 and 4 can be

calculated via Monte Carlo.

First, we show that the ideal Markov chain has the stationary distribution $\tilde{p}(c)$. To accomplish this, we assume that we begin in the distribution

$$\tilde{p}(c) = \frac{\langle c|\exp[-\beta\hat{H}]|c\rangle}{\sum_{c''}\langle c''|\exp[-\beta\hat{H}]|c''\rangle}$$

and show that a single step of the Markov chain leaves the system in the same probability distribution. In other words, we show

$$\sum_c \tilde{p}(c)\Pr(c \to c') = \tilde{p}(c')$$

where $\Pr(c \to c')$ is the transition probability of our Markov chain. We first note that

$$\Pr(c \to c')$$

$$= \frac{|\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c'\rangle|^2}{\sum_{c''}|\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c''\rangle|^2}$$

$$= \frac{|\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c'\rangle|^2}{\sum_{c''}\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c''\rangle\langle c''|\exp[-\beta\hat{H}/2]\hat{U}|c\rangle}$$

$$= \frac{|\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c'\rangle|^2}{\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}]\hat{U}|c\rangle}$$

$$= \frac{|\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c'\rangle|^2}{\langle c|\exp[-\beta\hat{H}]|c\rangle}$$

Thus, we may write

$$\sum_c \tilde{p}(c)\Pr(c \to c')$$

$$= \sum_c \frac{\langle c|\exp[-\beta\hat{H}]|c\rangle}{\sum_{c''}\langle c''|\exp[-\beta\hat{H}]|c''\rangle} \frac{|\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c'\rangle|^2}{\langle c|\exp[-\beta\hat{H}]|c\rangle}$$

$$= \sum_c \frac{|\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c'\rangle|^2}{\sum_{c''}\langle c''|\exp[-\beta\hat{H}]|c''\rangle}$$

$$= \frac{\sum_c \langle c'|\exp[-\beta\hat{H}/2]\hat{U}|c\rangle\langle c|\hat{U}^\dagger \exp[-\beta\hat{H}/2]|c'\rangle}{\sum_{c''}\langle c''|\exp[-\beta\hat{H}]|c''\rangle}$$

$$= \frac{\langle c'|\exp[-\beta\hat{H}]|c'\rangle}{\sum_{c''}\langle c''|\exp[-\beta\hat{H}]|c''\rangle}$$

$$= \tilde{p}(c')$$

Therefore, the stationary distribution is indeed $\tilde{p}(c)$.

Next, we prove eqn 1 is maximized by eqn 2. Maximizing eqn 1 is equivalent to setting the derivatives with respect to $\vec{\alpha}'$ to zero. Writing $\vec{\alpha}' = \vec{\alpha} + \delta\vec{\alpha}$, we work to first order in $\delta\vec{\alpha}$ and τ. For notational convenience, we'll denote $|\Psi[\vec{\alpha}]\rangle$ by $|\Psi\rangle$, and $|\Psi[\vec{\alpha}']\rangle$ by $|\Psi'\rangle$,

$$0 = \frac{\partial}{\partial\alpha_i'} \frac{\langle\Psi|(1-\tau\hat{H})|\Psi'\rangle}{\sqrt{\langle\Psi'|\Psi'\rangle}}$$

$$= \frac{\langle\Psi|(1-\tau\hat{H})|\Psi_i'\rangle}{\sqrt{\langle\Psi'|\Psi'\rangle}} - \frac{\langle\Psi|(1-\tau\hat{H})|\Psi'\rangle\langle\Psi'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle^{3/2}}$$

Dividing both sides by $\sqrt{\langle\Psi'|\Psi'\rangle}$ gives

$$0 = \frac{\langle\Psi|(1-\tau\hat{H})|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle} - \frac{\langle\Psi|(1-\tau\hat{H})|\Psi'\rangle}{\langle\Psi'|\Psi'\rangle}\frac{\langle\Psi'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle}$$

Expanding the first term to first order gives

$$\frac{\langle\Psi'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle} - \frac{\langle\Psi_j'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle}\delta\alpha_j - \tau\frac{\langle\Psi|\hat{H}|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle}$$

Expanding the second term to first order gives

$$-\frac{\langle\Psi'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle} + \frac{\langle\Psi_j'|\Psi'\rangle}{\langle\Psi'|\Psi'\rangle}\frac{\langle\Psi'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle}\delta\alpha_j + \tau\frac{\langle\Psi|\hat{H}|\Psi'\rangle}{\langle\Psi'|\Psi'\rangle}\frac{\langle\Psi'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle}$$

Thus, in total we have

$$0 = -\frac{\langle\Psi_j'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle}\delta\alpha_j - \tau\frac{\langle\Psi|\hat{H}|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle}$$
$$+ \frac{\langle\Psi_j'|\Psi'\rangle}{\langle\Psi'|\Psi'\rangle}\frac{\langle\Psi'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle}\delta\alpha_j + \tau\frac{\langle\Psi|\hat{H}|\Psi'\rangle}{\langle\Psi'|\Psi'\rangle}\frac{\langle\Psi'|\Psi_i'\rangle}{\langle\Psi'|\Psi'\rangle} \tag{5}$$

Since everything in eqn 5 is first order, we can replace $|\Psi'\rangle$ with $|\Psi\rangle$ wherever it occurs. Rearranging the equation, we get

$$\left(\frac{\langle\Psi_j|\Psi_i\rangle}{\langle\Psi|\Psi\rangle} + \frac{\langle\Psi_j|\Psi\rangle}{\langle\Psi|\Psi\rangle}\frac{\langle\Psi|\Psi_i\rangle}{\langle\Psi|\Psi\rangle}\right)\delta\alpha_j$$
$$= -\tau\left(\frac{\langle\Psi|\hat{H}|\Psi_i\rangle}{\langle\Psi|\Psi\rangle} - \frac{\langle\Psi|\hat{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle}\frac{\langle\Psi|\Psi_i\rangle}{\langle\Psi|\Psi\rangle}\right) \tag{6}$$

Comparing this to eqn 3 and eqn 4, we see this is equivalent to

$$(\mathbf{S})_{ij}\delta\alpha_j = -\tau h_i \tag{7}$$

from which eqn 2 immediately follows. We'll note that for this derivation to imply 2, $\mathbf{S}$ need not be two-sided invertible; it suffices for $\mathbf{S}^{-1}$ to be a right inverse to $\mathbf{S}$.

Finally, we show that both h_i and $(\mathbf{S})_{ij}$ can be evaluated using Monte Carlo. We rewrite each term in eqns 3

and 4 in a form amenable to Monte Carlo calculations.

$$\frac{\langle\Psi|\Psi_i\rangle}{\langle\Psi|\Psi\rangle} = \frac{\sum_c \frac{\Psi_i(c)}{\Psi(c)}|\Psi(c)|^2}{\sum_c|\Psi(c)|^2}$$

$$\frac{\langle\Psi_j|\Psi_i\rangle}{\langle\Psi|\Psi\rangle} = \frac{\sum_c \frac{\Psi_j(c)}{\Psi(c)}\frac{\Psi_i(c)}{\Psi(c)}|\Psi(c)|^2}{\sum_c|\Psi(c)|^2}$$

$$\frac{\langle\Psi_i|\hat{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle} = \frac{\sum_c \frac{\hat{H}\Psi(c)}{\Psi_i(c)}|\Psi(c)|^2}{\sum_c|\Psi(c)|^2}$$

$$\frac{\langle\Psi|\hat{H}|\Psi\rangle}{\langle\Psi|\Psi\rangle} = \frac{\sum_c \frac{\hat{H}\Psi(c)}{\Psi(c)}|\Psi(c)|^2}{\sum_c|\Psi(c)|^2}$$

From these equations, it is clear that each term can be calculated by sampling the distribution $|\Psi(c)|^2$. To produce these samples, we again use the Metropolis algorithm.

[1] S. R. White, Phys. Rev. Lett. **69**, 2863 (1992).
[2] E. Loh Jr, J. Gubernatis, R. Scalettar, S. White, D. Scalapino, and R. Sugar, Phys. Rev. B **41**, 9301 (1990).
[3] M. Suzuki, S. Miyashita, and A. Kuroda, Progr. Theor. Phys. **58**, 1377 (1977).
[4] D. M. Ceperley, Rev. Mod. Phys. **67**, 279 (1995).
[5] W. L. McMillan, Phys. Rev. **138**, A442 (1965).
[6] D. Ceperley, G. Chester, and M. Kalos, Phys. Rev. B **16**, 3081 (1977).
[7] R. Jastrow, Phys. Rev. **98**, 1479 (1955).
[8] R. Feynman and M. Cohen, Phys. Rev. **102**, 1189 (1956).
[9] M. A. Lee, K. Schmidt, M. Kalos, and G. Chester, Phys. Rev. Lett. **46**, 728 (1981).
[10] D. A. Huse and V. Elser, Phys. Rev. Lett. **60**, 2531 (1988).
[11] H. J. Changlani, J. M. Kinder, C. J. Umrigar, and G. K.-L. Chan, Phys. Rev. B **80**, 245116 (2009).
[12] F. Mezzacapo, N. Schuch, M. Boninsegni, and J. I. Cirac, New J. Phys. **11**, 083026 (2009).
[13] K. H. Marti, B. Bauer, M. Reiher, M. Troyer, and F. Verstraete, New J. Phys. **12**, 103008 (2010).
[14] F. Verstraete and J. I. Cirac, arXiv:cond-mat/0407066 (2004).
[15] V. Murg, F. Verstraete, and J. I. Cirac, Phys. Rev. A **75**, 033605 (2007).
[16] S. R. White, Phys. Rev. Lett. **102**, 190601 (2009).
[17] E. Stoudenmire and S. R. White, New J. Phys. **12**, 055026 (2010).
[18] A. E. Feiguin and S. R. White, Phys. Rev. B **72**, 220401 (2005).
[19] B. Militzer and E. Pollock, Phys. Rev. E **61**, 3470 (2000).
[20] W. Kohn and L. J. Sham, Phys. Rev. **140**, A1133 (1965).
[21] K. Takai, K. Ido, T. Misawa, Y. Yamaji, and M. Imada, Journal of the Physical Society of Japan **85**, 034601 (2016).
[22] S. Sorella, Phys. Rev. B **64**, 024512 (2001).
[23] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, J. Chem. Phys. **21**, 1087 (1953).
[24] A. W. Sandvik, Physical Review B **59**, R14157 (1999).
[25] N. Blunt, T. Rogers, J. Spencer, and W. Foulkes, Physical Review B **89**, 245124 (2014).
[26] F. D. Malone, N. Blunt, J. J. Shepherd, D. Lee, J. Spencer, and W. Foulkes, The Journal of chemical physics **143**, 044116 (2015).
[27] F. D. Malone, N. Blunt, E. W. Brown, D. Lee, J. Spencer, W. Foulkes, and J. J. Shepherd, arXiv preprint arXiv:1602.05104 (2016).
[28] S. Zhang, Physical review letters **83**, 2777 (1999).
[29] B. M. Rubenstein, S. Zhang, and D. R. Reichman, Physical Review A **86**, 053606 (2012).
[30] B. Clark, "Algorithms for finite temperature QMC," (2013), presented at Advances in quantum Monte Carlo techniques for non-relativistic many-body systems.
[31] G. H. Booth, A. J. Thom, and A. Alavi, J. Chem. Phys. **131**, 054106 (2009).
[32] R. Blankenbecler, D. J. Scalapino, and R. L. Sugar, Phys. Rev. D **24**, 2278 (1981).
[33] G. Sugiyama and S. Koonin, Ann. Phys. **168**, 1 (1986).
[34] D. Ten Haaf, H. Van Bemmel, J. Van Leeuwen, W. Van Saarloos, and D. Ceperley, Physical Review B **51**, 13039 (1995).
[35] D. Ceperley, J. Stat. Phys. **63**, 1237 (1991).
[36] D. Ceperley, Phys. Rev. Lett. **69**, 331 (1992).
[37] A. W. Sandvik, http://physics.bu.edu/ sandvik/vietri/sse/results.html Computer code SSE.