Lecture 12: Quantum imaginary time evolution (QITE), etc
Reading: Motta et al, 2019.

1 Introduction

The past few algorithms we've seen have been variational algorithms in that they required specification of an ansatz with some parameters that were varied to yield the desired result, e.g. the minimum energy as an estimate of the ground state energy. Wouldn't it be nice if there were a way to access the ground state energy without requiring specification of an ansatz?

In classical simulations, there are methods to estimate extremal eigenvalues using the exact Hamiltonian. One of these methods is imaginary time evolution, as we saw before. To recap, the Schrodinger equation in imaginary time is:

$$-\frac{\partial |\Phi(\beta)\rangle}{\partial \beta} = H |\Phi(\beta)\rangle \tag{1.1}$$

with a solution of $\Phi(\beta) = \exp(-\beta H) |\Phi(\beta)\rangle$. As $\beta \to \infty$, high energy states are projected out at a rate that depends on the gap between the ground states and excited states. Therefore, ITE eventually leads to the ground state without requiring any ansatz.

Various classical algorithms exist to perform time evolution, particularly stochastic methods like diffusion Monte Carlo. The Lanczos method that we will discuss in detail is effectively another type of imaginary time evolution that works by constructing H in a so-called Krylov subspace formed from powers of H applied to a starting state.

The trouble with these classical algorithms is that, as expected, they suffer from exponential complexity in space and time. The exponential space originates from storing the state vector, and the exponential time stems from the cost of multiplying the Hamiltonian. We would like a quantum version of these algorithms that could potentially solve some of these problems. The problem is, how to implement the non-unitary operation on a quantum device that requires unitary operators? We can of course add ancilla qubits to make the operation unitary, but we generally don't have many extra qubits available in near-term devices. Plus, non-unitary operations mean that the operation has some probability of failure, necessitating increased numbers of experiments.

Quantum imaginary time evolution (QITE) and QLanczos are quantum algorithms that address these problems. Per iteration, they provide exponential reductions in space and time requirements compared to the classical algorithms. Further, they do not require deep circuits or ancilla qubits. Finally, they are guaranteed to converge to the ground state by the properties of ITE, unlike variational algorithms based on classical nonlinear optimization. Despite these favorable properties, note that the algorithm cannot solve every ground state problem in polynomial time as this problem is QMA-complete. However, the algorithms can address many physical problems for which correlations between sites are bounded in spatial extent, as we will see.

2 Quantum imaginary time evolution (QITE)

Let's start by assuming we have a k-local $H = \sum_m h_m$. If we Trotterize the imaginary time evolution, we obtain:

$$e^{-\beta H} = (e^{-\Delta\tau h_1} ... e^{-\Delta\tau h_l})^n + O(\Delta\tau) \tag{2.1}$$

where $n = \beta/\Delta\tau$. After the application of one term in the Trotter step, we have $|\Psi'\rangle = e^{-\Delta\tau h_m}|\Psi\rangle$. As mentioned above, this operation is not unitary and thus does not appear to be implementable on a QC without ancilla qubits.

It turns out that we can actually find a unitary operator between $|\Psi\rangle$ and a normalized final state, $|\bar{\Psi}'\rangle = |\Psi'\rangle / \|\Psi'\|$. Uhlmann's theorem guarantees that we can find such a unitary so that (proof given in the SI of the Motta paper):

$$|\bar{\Psi}'\rangle = e^{-i\Delta\tau A[m]}|\Psi\rangle \tag{2.2}$$

where $A[m]$ is a Hermitian operator that acts on a neighborhood of the qubits of h_m. The operator can be determined by tomography on $|\Psi\rangle$.

How to do that? Let's start by taking $|\Psi\rangle$ to be a product state. One quantity we need is the normalization constant. We can get that by:

$$\begin{aligned} c &= \|\Psi'\|^2 = \langle\Psi'|\Psi'\rangle = \langle\Psi|e^{-\Delta\tau h_m}e^{-\Delta\tau h_m}|\Psi\rangle = \langle\Psi|e^{-2\Delta\tau h_m}|\Psi\rangle & (2.3)\\ &= 1 - 2\Delta\tau\langle\Psi|h_m|\Psi\rangle + O(\Delta\tau^2) & (2.4)\end{aligned}$$

Given $|\Psi\rangle$, we can measure the operator h_m to figure out c.

Next we need to get the unitary that implements the ITE. Generally, $A[m]$ as a Hermitian operator can be expanded as a sum of Pauli strings:

$$A[m] = \sum_{i_1,i_2,...i_k} a[m]_{i_1...i_k}\sigma_{i_1}...\sigma_{i_k} \tag{2.5}$$

In this example, since $|\Psi\rangle$ is a product state the unitary only needs to act over the k qubits of h_m. In general, the unitary would need to act over all qubits that are entangled in the neighborhood of h_m qubits, which is in general a bigger region.

To find the unitary, we want to find the coefficients such that the unitary carries out the ITE while including the normalization. Let's define

$$|\Delta_0\rangle = \frac{|\bar{\Psi}'\rangle - |\Psi\rangle}{\Delta\tau} \tag{2.6}$$

representing the difference between the desired state and the initial state. We can also define the difference between the state after unitary evolution and the initial state as $|\Delta\rangle = -iA[m]|\Psi\rangle$. Therefore, we want to minimize $\|\Delta_0 - \Delta\|$. Define $f = \|\Delta_0 - \Delta\|^2$. Then,

$$\begin{aligned} f &= (\langle\Delta_0| - \langle\Delta|)(|\Delta_0\rangle - |\Delta\rangle) = \langle\Delta_0|\Delta_0\rangle - \langle\Delta|\Delta_0\rangle - \langle\Delta_0|\Delta\rangle + \langle\Delta|\Delta\rangle & (2.7)\\ &= f_0 + ic^{-1/2}\langle\Delta_0|\sum_I \sigma_I|\Psi\rangle - ic^{-1/2}\langle\Psi|\sum_{a_I}\sigma_I^\dagger|\Delta_0\rangle + \langle\Psi|\sum_I a_I\sigma_I^\dagger\sum_J a_J\sigma_J|\Psi\rangle & (2.8)\end{aligned}$$

Since we want to minimize this functional with respect to variations in the trial wavefunction, we take a derivative in the coefficients a_I to yield a linear system of equations that must be satisfied. The result is:

$$\begin{aligned} (S + S^T)x &= b & (2.9)\\ S_{IJ} &= \langle\Psi|\sigma_I\sigma_J|\Psi\rangle & (2.10)\\ b_I &= 2c^{-1/2}Im\langle\Psi|\sigma_I h_m|\Psi\rangle & (2.11)\end{aligned}$$

where x is a vector of desired coefficients a_I. Solving this system thus requires obtaining the entries of the matrix, which in turn requires measuring expectation values of Pauli strings over the present state Ψ. Although it appears we need to fill the entire matrix S_{IJ} thus requiring many measurements, remember that the Pauli operators form an algebra such that the product of any two operators equals another operator. Therefore, measuring the Pauli strings in the correlation length of the given site (e.g. performing tomography on the state) is sufficient to fill both S and b.

Since $S+S^T$ in general has a non-trivial null space, we use least-squares to obtain the coefficients $\vec{a}$. These coefficients define the expansion of the unitary that we need to evolve the state forward in imaginary time. Once we obtain the coefficients, we apply the unitary, evolve to a new state, and perform the same set of steps to find the next unitary, and so on. After some number of steps, we will arrive at a step with good overlap with the ground state, without requiring any ansatz or assumption about the state for the general case.

However, starting from a product state over k qubits, the support of the unitaries increases in imaginary time as adjacent sites become entangled. Examination of the scheme shows that if the correlation length in imaginary time is unbounded, the resources required to fill the matrices grow exponentially. In general, this outcome must occur because finding the ground state of a Hamiltonian is generically a hard (QMA) problem. Therefore, the method is efficient only if the Hamiltonian exhibits a correlation length that saturates in imaginary time. That assumption is equivalent to assuming that the Hamiltonian is local and gapped, as for a gapped system correlations between observables separated by distance l are bounded by $e^{-l/C}$ where C is the correlation length.

2.1 Running time

We need to perform full tomography on R_v to obtain the matrix elements needed to solve the least-squares problem and figure out the unitary. Therefore, the cost grows exponentially with the correlation length. As a result, the method can only be efficient if the correlation length is bounded by the physical interactions of the problem.

A bound for the running time is $mn\exp(O(k(2C)^d\ln^d(2\sqrt{2}nm\epsilon^{-1}))$. So the cost is exponential in C and quasipolynomial in n and m. We can generally beat this bound if we assume that the error from applying local Hamiltonian terms does not add to the error of the entire state. Specifically, the quasipolynomial dependence on the number of sites originates from the error term containing a factor of m (see statement of the theorem in the SI). If we assume that this error does not add to the error for a local state probed only by local observables, then this m dependence is removed from the exponential term and the error scales linearly with $m \sim O(N)$, where N is the total number of sites.

3 Q-Lanczos

Now that we can do ITE efficiently for many physical problems, what else can we do? Another algorithm we can make based on QITE is the quantum version of the Lanczos algorithm. The Lanczos method is a classical algorithm for finding extremal eigenvalues exactly that is typically applied for large matrices. A fast way to do that is given by the Lanczos method which provides extremal eigenstates and eigenvalues of a matrix H.

3.1 Aside - review of classical Lanczos

It is due to Lanczos 1950 with improvements reported by Ojalvo 1970, Paige 1972. The only input is an algorithm for computing $H|\psi\rangle$ for any $|\psi\rangle$.

By expanding an arbitrary state $|\psi\rangle$ in the basis of (unknown) eigenstates of a matrix H, one can show that the problem of finding extremal eigenvalues exactly corresponds to the problem of finding extremal values of a functional $E[\psi]$:

$$E[\psi] = \frac{\langle\psi|H|\psi\rangle}{\langle\psi|\psi\rangle} \tag{3.1}$$

Let us call the extremal value $E_g = min\ E[\psi]$, and the corresponding state $E[\psi_g] = E_g$. We can take a functional gradient of Eq. 3.1 to obtain

$$\frac{\delta E[\psi]}{\delta \langle\psi|} = \frac{H|\psi\rangle - E|\psi\rangle}{\langle\psi|\psi\rangle} \equiv |\psi_a\rangle \tag{3.2}$$

By the properties of a gradient, $|\psi_a\rangle$ gives the direction of the maximum rate of change of the functional. Therefore, moving in the opposite direction will decrease the functional $E[\psi]$. The next natural question is how far we should move in that direction? As an illustration of a way to answer that question, let us set up another variational problem in which we seek to find an $\alpha > 0$ such that $E[\psi - \alpha\psi_a]$ is lowered by a maximal amount. We therefore seek to minimize $E[\psi - \alpha\psi_a]$ by varying α in the space $K_2 = \{|\psi\rangle, |\psi_a\rangle\} = \{|\psi\rangle, H|\psi\rangle\}$ (recalling that $|\psi_a\rangle$ was obtained by applying H).

To start the process, we define our first normalized basis vector as $|v_0\rangle = |\psi\rangle / \langle\psi|\psi\rangle$ where $|\psi\rangle$ is a random initial state. The second basis vector is chosen to be in the space $H|v_0\rangle$ but orthogonal to the first one:

$$|\tilde{v}_1\rangle = H|v_0\rangle - |v_0\rangle\langle v_0|H|v_0\rangle \tag{3.3}$$

With this definition, $\langle v_0|\tilde{v}_1\rangle = 0$. We then normalize to get $|v_1\rangle = |\tilde{v}_1\rangle / \langle\tilde{v}_1|\tilde{v}_1\rangle$. If we define $a_0 = \langle v_0|H|v_0\rangle$ and $b_1^2 = \langle\tilde{v}_1|\tilde{v}_1\rangle$, we can rewrite the equation above as:

$$H|v_0\rangle = b_1|v_1\rangle + a_0|v_0\rangle \tag{3.4}$$

Remember that H can be interpreted as consisting of projector elements like $|v_1\rangle\langle v_0|$. Therefore, b_1 can also be interpreted as the matrix element for the given projector, which is $b_1 = \langle v_1|H|v_0\rangle$. In the space K_2, H has the representation

$$H_{K_2} = \begin{pmatrix} a_0 & b_1 \\ b_1 & a_1 \end{pmatrix} \tag{3.5}$$

where $a_1 = \langle v_1|H|v_1\rangle$. If we find the ground state of H_{K_2}, yielding a state $|g\rangle_{K_2}$, we find the minimum energy within the space $\{|\psi\rangle, H|\psi\rangle\}$ and therefore the optimal α. We can now iterate the process: use $|g\rangle_{K_2}$ as the initial state for another iteration and perform the same procedure. We find that convergence is rapid. To assess convergence, we monitor the residual $r[\psi] = ||(H - E[\psi])|\psi\rangle||^2 = \langle\psi|H^2|\psi\rangle - E^2\langle\psi|\psi\rangle$.

If we perform the procedure above L times, starting from $|v_0\rangle$, the resulting vector will lie in the Krylov space $K_L = \{|v_0\rangle, H|v_0\rangle, H^2|v_0\rangle, ..., H^L|v_0\rangle\}$. This vector space is referred to as the Krylov space of H over $|v_0\rangle$ and has dimension $L+1$. We can use this space to find extremal eigenvalues by constructing K_L and then minimizing H within this space to look for the ground state. To do so, we construct the Krylov basis iteratively. For this discussion, let $a_n = \langle v_n|H|v_n\rangle$. $|v_1\rangle$ is defined exactly as above. For $|v_2\rangle$, we want to lie in the space $H|v_1\rangle = H^2|v_0\rangle$ but be orthogonal to the other vectors in the Krylov space. Defining in general $b_n = \sqrt{\langle\tilde{v}_n|\tilde{v}_n\rangle}$, we get

$$b_2|v_2\rangle = |\tilde{v}_2\rangle = H|v_1\rangle - \sum_{i=0}^{1}|v_i\rangle\langle v_i|H|v_i\rangle = H|v_1\rangle - a_1|v_1\rangle - b_1|v_0\rangle \tag{3.6}$$

Note that $\langle v_2|H|v_0\rangle = 0$ because $H|v_0\rangle \in K_1$ and $|v_2\rangle$ is designed to be orthogonal to vectors in K_1. Proceeding further, we get

$$b_3 |v_3\rangle = |\tilde{v}_3\rangle = H |v_2\rangle - \sum_{i=0}^{2} |v_i\rangle \langle v_i|H|v_i\rangle = H |v_2\rangle - a_2 |v_2\rangle - b_2 |v_1\rangle \tag{3.7}$$

where we used that $\langle v_2|H|v_0\rangle = 0$. We have therefore established a two-term recursion relation that in general looks like:

$$b_{n+1} |v_{n+1}\rangle = |\tilde{v}_{n+1}\rangle = H |v_n\rangle - \sum_{i=0}^{n} |v_i\rangle \langle v_i|Hv_n\rangle = H |v_n\rangle - a_n |v_n\rangle - b_n |v_{n-1}\rangle \tag{3.8}$$

where we always have $\langle v_{n+1}|H|v_i\rangle = 0$ if $i = 0, ..., n-1$ by the same logic as above. If we rearrange this equation, we find that $H |v_n\rangle = b_{n+1} |v_{n+1}\rangle + a_n |v_n\rangle + b_n |v_{n-1}\rangle$. Therefore in the Krylov basis H_{K_L} is tridiagonal. This matrix is on the order of tens $\times$ tens in dimension and thus can be easily diagonalized to obtain the ground state ψ_g^L which obeys the equation:

$$(H_{K_L})_{ij}(\psi_g^L)_j = E_g^L(\psi_g^L)_i \tag{3.9}$$

where E_g^L is our best estimate for E_g in the Krylov space K_L. The best estimate of the ground state of the initial matrix is:

$$|\psi_g^L\rangle = \sum_j (\psi_g^L)_j |v_j\rangle \tag{3.10}$$

For computational efficiency (particularly regarding memory), it can be constructed on the fly as the Krylov space is gradually increased. Remarkably, this scheme converges exponentially with a rate constant proportional to the square root of the gap to the first excited state!

Summary of classical algorithm

1. Start with arbitrary $|v_0\rangle$.
2. Perform an abbreviated initial step:
   (a) $|\tilde{v}_1\rangle = H |v_0\rangle$
   (b) $a_0 = \langle \tilde{v}_1|v_0\rangle$
   (c) $|\tilde{v}_2\rangle = |\tilde{v}_1\rangle - a_0 |v_1\rangle$ (to make new vector orthogonal to previous one)
3. Ground state iteration ($n \geq 1$):
   (a) $b_n = \sqrt{\langle \tilde{v}_n|\tilde{v}_n\rangle}$
   (b) if $b_n \neq 0$, then $|v_n\rangle = |v_n\rangle / b_n$, else pick $|v_n\rangle$ as an arbitrary vector that is orthogonal to $|v_1\rangle ... |v_{n-1}\rangle$.
   (c) $|\tilde{\tilde{v}}_{n+1}\rangle = H |v_n\rangle$
   (d) $a_n = \langle \tilde{\tilde{v}}_{n+1}|v_n\rangle$
   (e) $|\tilde{v}_{n+1}\rangle = |\tilde{\tilde{v}}_{n+1}\rangle - a_n |v_n\rangle - b_n |v_{n-1}\rangle$
   (f) Return to (a).

There are other ways of organizing this iteration loop, but the one shown here is numerically the most stable (see Paige 1972 thesis).

3.2 Back to Q-Lanczos

How can we get a quantum version of this algorithm? Well, we just figured out a way to generate $|\Phi_l\rangle = e^{-l\Delta\tau H}|\Phi\rangle$. For sufficiently small $\Delta\tau$, applying that operator is equivalent to applying H to a state. So in the big picture, if we keep applying $e^{-l\Delta\tau H}$ for various l, we should be able to generate the Lanczos matrix. The smallest eigenvalue of this matrix yields an estimate of the ground state energy.

Let's go over the algorithm in more detail. To perform Lanczos, we need the expectation value of the energy, $\langle\Phi_l|H|\Phi_{l'}\rangle$, and the overlap matrix $\langle\Phi_l|\Phi_{l'}\rangle \equiv c_l^2$. (A difference with classical Lanczos is that we cannot impose orthogonality of the vectors so we need to keep track of the overlaps). To facilitate taking expectation values by measurement, we will build these matrices in a basis formed by applying ITE an even number of time steps.

Consider the overlap first. We have:

$$\langle\Phi_l|\Phi_{l'}\rangle = \langle\Phi|e^{-l\Delta\tau H}e^{-l'\Delta\tau H}|\Phi\rangle = \langle\Phi|e^{-(l+l')\Delta\tau H/2}e^{-(l+l')\Delta\tau H}|\Phi/2\rangle \tag{3.11}$$

This matrix can be computed by evolving in imaginary time for $(l+l')/2$ steps and measuring to get $c_{(l+l')/2}$. Therefore, $S_{ll'} = c^2_{(l+l')/2}$, and we only need to compute the diagonal of the matrix to get all the entries.

Now let's try the Hamiltonian.

$$H_{ll'} = \langle\Phi_l|H|\Phi_{l'}\rangle = \langle\Phi|e^{-l\Delta\tau H}He^{-l'\Delta\tau H}|\Phi\rangle \tag{3.12}$$

Since all the operators involved commute we can play a similar trick as for the overlap matrix so that $H_{ll'} = H_{(l+l')/2,(l+l')/2}$. Once we compute the diagonal of this matrix, we have all the elements. We then solve a generalized eigenvalue problem ($Hx = ESx$, x = eigenvector and E = eigenvalue = energy) to get an estimate of the ground and excited states.

Here is a little more detail about how everything is actually computed. Let

$$|\Phi_l\rangle = n_l e^{-l\Delta\tau H}|\Psi_T\rangle \tag{3.13}$$

where n_l is the normalization factor (norm of the state on RHS) and $|\Psi_T\rangle$ is a trial state. Let l range from 0 to L_{max} and set $2r = l + l'$. Then

$$S_{ll'} = n_l n_l' \langle\Psi_T|e^{-l\Delta\tau H}e^{-l'\Delta\tau H}|\Psi_T\rangle = \frac{n_l n_l'}{n_r^2} \tag{3.14}$$

$$H_{ll'} = n_l n_l' \langle\Psi_T|e^{-l\Delta\tau H}He^{-l'\Delta\tau H}|\Psi_T\rangle = \frac{n_l n_l'}{n_r^2}\langle\Phi_r|H|\Phi_r\rangle = S_{ll'}\langle\Phi_r|H|\Phi_r\rangle \tag{3.15}$$

We obtain the normalization factor n_r recursively as:

$$\frac{1}{n_{r+1}^2} = \langle\Psi_T|e^{-(r+1)\Delta\tau H}e^{-(r+1)\Delta\tau H}|\Psi_T\rangle = \frac{\langle\Phi_r|e^{-2\Delta\tau H}|\Phi_r\rangle}{n_r^2} \tag{3.16}$$

So by performing sequential imaginary time steps and storing the relevant measurements at each time step, we generate the necessary matrix elements needed to solve the linear system on a classical computer.

A relevant note: when this algorithm is actually implemented on a QC, the measurements you get back are not perfect measurements but rather contain noise that can complicate solving the linear system. A way around that is to add a diagonal matrix to the Lanczos matrix before the linear solve to make the matrix diagonally dominant. That has the effect of stabilizing the linear system.

4 Q-METTS for thermal averages

4.1 Classical algorithm

We will be able to use QITE to obtain thermal averages of observables also. But most of the section will be spent explaining the classical version of Minimally Entangled Typical Thermal States (METTS) originally described in Refs [1, 2]. Once that is done, the quantum version is easy.

In quantum statistical mechanics, we often want to compute averages of the form:

$$\langle A \rangle = Tr[\rho A]/Tr[\rho] = \frac{1}{Z} Tr[e^{-\beta H} A] \tag{4.1}$$

Trying to compute this formula by brute force is usually impossible because we can't compute $e^{\beta H}$ (if we could, we could solve any other problem involving H also, which we know we can't). Prior approaches to computing thermal averages were based on quantum Monte Carlo to integrate over the very high dimensional space. Although quantum Metropolis algorithms have been proposed, [3, 4] sampling and rejecting are not natural operations on a QC and require large circuit depth.

METTS was proposed as an alternate approach to compute these thermal averages [1]. The goal of the original work was to find states $|\phi\rangle$ that are "typical" of a quantum system at finite temperature. In more detail, let's rewrite our thermal average in the following way:

$$\begin{aligned} \langle A \rangle &= Tr[\rho A]/Tr[\rho] = Z^{-1} Tr[e^{-\beta H} A] = Z^{-1} Tr[e^{-\beta H/2} A e^{-\beta H/2}] && (4.2) \\ &= Z^{-1} \sum_i \langle i | e^{-\beta H/2} A e^{-\beta H/2} | i \rangle && (4.3) \\ &= Z^{-1} \sum_i p(i) \langle \phi(i) | A | \phi(i) \rangle && (4.4) \end{aligned}$$

where $p(i) = \langle i | \exp -\beta H | i \rangle$ and we defined new states as:

$$|\phi(i)\rangle \equiv p(i)^{-1/2} e^{-\beta H/2} |i\rangle \tag{4.5}$$

With this definition, the partition function $Z = Tr\rho = \sum_i p(i)$. We see that the new states $|\phi(i)\rangle$ we have defined have the following property:

$$\sum_i p(i) |\phi(i)\rangle \langle \phi(i)| = e^{-\beta H} \tag{4.6}$$

So if we had such states that would be great, but there are two problems we have to solve. First, we would like an orthonormal basis $|i\rangle$ that is distributed according to $p(i)$ so that we obtain the desired weighted average (we have not yet specified what the basis is; that is the next problem). We can arrange for this distribution to occur by a suitable Markov chain. Recall that a Markov chain is a stochastic process in which the probability to transition to any state depends only on the

present state. Therefore, we need to define a rule to transition from state $|i\rangle$ to state $|j\rangle$. The most important requirement of this transition rule is that the limiting distribution of this transition rule is the desired one for computing thermal averages, $p(i)/Z \equiv p'(i)$.

To figure out the rule, consider the transition rate of states from an arbitrary state i. We can write it as:

$$q(i) = -p'(i)\sum_j g(i \to j) + \sum_j p'(j)g(j \to i) \tag{4.7}$$

where $g(...)$ is the transition rule we need. If $p'(i)$ is a stationary distribution, by definition this flux should be zero, indicating that $p'(i)$ is the limiting distribution to which the Markov chain tends. For this outcome to always occur, we need to enforce a detailed balance:

$$p'(i)g(i \to j) = p'(j)g(j \to i) \tag{4.8}$$

which sets one of the conditions of the transition rule. The other requirement is that $g(...)$ should be a probability between 0 and 1. With only these requirements, there is a natural choice for the transition rate:

$$g(i \to j) = |\langle j|\phi(i)\rangle|^2 = \frac{|\langle j|e^{-\beta H/2}|i\rangle|^2}{p(i)} \tag{4.9}$$

You can check that this distribution indeed does satisfy detailed balance and is a probability.

The next question is what are the basis states $|i\rangle$? There are a few choices. We could first take the energy eigenstates of H as $|i\rangle$ in which case $|\phi(i)\rangle = |i\rangle$. There are obvious problems with this choice. First, getting the states requires full diagonalization. Second, it is a poor choice on physical grounds if the state we're interested in has a broken symmetry that is not present in the original Hamiltonian. Another choice is random vectors. However, this choice can lead to highly entangled states of two subsystems that may not be entangled.

A better choice is to take $|i\rangle$ as a CPS! These states start with entanglement entropy of 0. After ITE to obtain $|\phi\rangle$, we expect them to have minimum entanglement as they started with zero entanglement. This choice has a number of nice properties, such as (1) METTS for a system with non-interacting systems is the tensor product of the METTS; (2) since the CPS is chosen randomly, METTS can break symmetries.

With all these considerations, here is the classical algorithm:

1. Choose a CPS $|i\rangle$
2. Compute METTS $|\phi(i)\rangle$ and calculate observables.
3. Collapse a new CPS $|i'\rangle$ from $|\phi(i)\rangle$ with the probability $p(i \to i') = |\langle i'|\phi(i)\rangle|^2$.

In the original papers, this algorithm was implemented using the matrix product state ansatz.

4.1.1 Quantum algorithm

With the classical algorithm figured out, the adaptation is easy. Here is the algorithm:

1. On a QC, prepare a product state $|i\rangle$.
2. Use QITE to compute $|\phi_i\rangle = p_i^{-1/2} \exp(-\beta H/2) |i\rangle$ and measure the desired observable
3. Perform another iteration of QITE from the same product state, measure a product operator like $Z_1 \otimes Z_2 \otimes ...$ to collapse to another CPS, repeat.

Measurement naturally chooses a product state with exactly the probability we want, e.g. $|\langle i'|\phi(i)\rangle|^2$. To avoid statistical correlations between samples, we can alternate the basis we measure the state in by performing basis rotations prior to measurement (e.g. to measure in X apply H before measurement).

References

[1]S. R. White, "Minimally entangled typical quantum states at finite temperature.", Physical Review Letters **102**, 190601 (2009).

[2]E. M. Stoudenmire and S. R. White, "Minimally entangled typical thermal state algorithms", New journal of physics **12**, 055026 (2010).

[3]K. Temme, T. J. Osborne, K. G. Vollbrecht, D. Poulin, and F. Verstraete, "Quantum metropolis sampling.", Nature **471**, 87–90 (2011).

[4]M.-H. Yung and A. Aspuru-Guzik, "A quantum-quantum metropolis algorithm.", Proceedings of the National Academy of Sciences of the United States of America **109**, 754–759 (2012).