Sampling from the Thermal Quantum Gibbs State and Evaluating Partition Functions with a Quantum Computer

David Poulin\textsuperscript{1} and Pawel Wocjan\textsuperscript{2}

\textsuperscript{1}Département de Physique, Université de Sherbrooke, Québec, Canada, J1K 2R1

\textsuperscript{2}School of Electrical Engineering and Computer Science, University of Central Florida, Florida 32816-2362, USA

(Received 15 June 2009; published 24 November 2009)

We present a quantum algorithm to prepare the thermal Gibbs state of interacting quantum systems. This algorithm sets a universal upper bound $D^\alpha$ on the thermalization time of a quantum system, where $D$ is the system’s Hilbert space dimension and $\alpha \leq \frac{1}{2}$ is proportional to the Helmholtz free energy density. We also derive an algorithm to evaluate the partition function of a quantum system in a time proportional to the system’s thermalization time and inversely proportional to the targeted accuracy squared.

We call a algorithm

\[ \rho(\beta) = e^{-\beta H}Z(\beta) \]

of any locally interacting quantum system, where $Z(\beta) = \text{Tr}(e^{-\beta H})$ is the partition function. The problem of thermalizing a quantum system with a quantum computer was first raised in [4] where heuristic methods were proposed, but no rigorous upper bounds have been derived. Our results set a universal upper bound $D^\alpha$ for the thermalization time of a quantum system, where $D$ is the system’s Hilbert space dimension and $\alpha \leq \frac{1}{2}$ is proportional to the Helmholtz free energy density.

Second, we present a quantum algorithm to evaluate the partition function of quantum systems within relative accuracy $\epsilon$. The complexity of the algorithm scales like the inverse of $\epsilon^2$ (when the system is classical) and proportionally to the thermalization time of the system. This algorithm is not affected by the sign problem and works for any locally interacting quantum system. It is a full quantum generalization of the algorithm presented in [1,2] for the evaluation of partitions functions of classical systems. The combination of our two algorithms yields a universal upper bound for the time needed to evaluate a partition function with a quantum computer.

We consider a system composed of $n$ interacting $d$-level particles with a $k$-local Hamiltonian $H = \sum_i h_i$, where each term $h_i$ has a bounded norm and acts on at most $k$ particles ($k$ is a constant). It is one of quantum information science’s most celebrated result that $k$-local Hamiltonians can be simulated efficiently on a quantum computer. More precisely, for those Hamiltonians, the time evolution operator $U(t) = e^{-itH}$ can be approximately expressed as the product of a sequence of discrete one- and two-qubit gates, and the number of gates scales essentially linearly with the duration $t$ of the simulated process and polynomially with the system size $n$ [5–8].

Given the system’s Hamiltonian and a temperature $1/\beta$, we call a thermalization process any (possibly time-dependent) local Hamiltonian on the system and a bath (of size scaling polynomially with the number of particles in the system) that generates, after time $T$, a unitary transformation $V$ such that

\[ \text{Tr}_{\text{bath}} \left\{ V \left( \frac{I}{D} \otimes \lvert 0 \rangle \langle 0 \rvert_{\text{bath}} \right) V^\dagger \right\} = \rho(\beta), \tag{1} \]

or more generally that this equation holds within some accuracy $\delta$. The system’s thermalization time $T_{\text{th}}(\beta)$ is...
the duration of the shortest thermalization process. Note
that in our definition, the temperature dependence is en-
coded in the details of \( V \) instead of in the initial state of the
bath as one would expect physically. This is to rule out
trivial thermalization processes where the bath is prepared
in the desired state \( \rho(\beta) \) and \( V \) simply swaps the system
and the bath. Nevertheless, the temperature dependence of
the thermalization process we present could equally well be
encoded in the initial state of the bath with a simple modifica-
tion to our algorithm.

Thermalization processes that occur in nature constitute
of local interactions (all forces of nature are two-body), so
 can be efficiently simulated by a quantum computer, i.e., in
a time proportional to the duration of the thermalization
process \([5–8]\). In what follows, we present a universal
thermalization process and characterize its complexity.
Then, we show how to use this—or any other—thermal-
ization process to evaluate partition functions with a quan-
tum computer.

For sake of analysis, we shift the system’s Hamiltonian
by a constant \( E^\ast \) to ensure that \( H \) is positive. This changes
the partition function by a factor \( e^{-\beta E^\ast} \) that we keep track
of implicitly: we henceforth assume \( 0 < H < E_{\text{max}} \) where
\( E_{\text{max}} \) is a known (polynomial) upper bound of the
Hamiltonian. We denote the eigenvalues and eigenvectors of
the Hamiltonian \( H(a) = E_a a \).

We will make extensive use of the system’s time evolu-
tion operator with \( t = \pi/(4E_{\text{max}}) \), that we simply denote \( U \).
As mentioned above, \( U \) cannot be simulated exactly
with a quantum computer. Instead, we can produce \( \tilde{U} \),
which is a good approximation to it. This will unavoidably
limit the accuracy of our thermalization process and of the
estimated partition function. Indeed, defining the effective
limit the accuracy of our thermalization process and of the
Hamiltonian \( \tilde{H} = - \frac{1}{t} \log \tilde{U} \) (by definition of \( t \), there is no
multiple value problem), our algorithms will prepare the Gibbs
state \( \tilde{\rho}(\beta) \) and evaluate the partition function \( \tilde{Z}(\beta) \) asso-
ciated with \( \tilde{H} \), not \( H \). However, we show in Appendix A of
the supplementary material \([9]\) that an accurate simulation
\( ||U - \tilde{U}|| \leq \epsilon \) leads to an accurate effective Hamiltonian
\( ||H - \tilde{H}|| \leq K E_{\text{max}} \epsilon \) (where \( K \) is a small constant). This,
in turn, implies an accurate estimate of the partition func-
tion \( |\tilde{Z}(\beta) - \tilde{Z}(\beta)| \leq 2K \beta E_{\text{max}} \epsilon \tilde{Z}(\beta) \) (see Appendix B
in the supplementary material \([9]\)) and a high fidelity Gibbs
state \( F(\tilde{\rho}(\beta), \tilde{\rho}(\beta)) \geq 1 - K' E_{\text{max}} \epsilon \) (see Appendix C in
the supplementary material \([9]\)).

Because the exact time required to simulate \( U \) within
accuracy \( \epsilon \) depends on the details of the system, we mea-
sure the complexity of our algorithms in terms of the
number of times they need to implement \( U \). Hence, the
true running time of our algorithms have an additional
arbitrarily weak dependence \([6]\) on the targeted accuracy
\( \epsilon \) of the partition function, and Gibbs state.

Our algorithm requires a cooling schedule \( 0 = \beta_0 < \beta_1 < \ldots < \beta_\ell = \beta \) such that \( F_k := \frac{Z(\beta_{k+1})}{Z(\beta_k)} \geq 1/2 \)
for all \( k \). Writing \( F_k = \sum_a e^{-\beta_k E_a} e^{-\Delta \beta_k E_a} = (e^{-\Delta \beta_k H})_{\beta_k} \)
where \( \Delta \beta_k = \beta_{k+1} - \beta_k \), we see that a polynomial-length
cooling schedule can always be constructed by choosing
\( \Delta \beta_k = \log(2)/E_{\text{max}} \), which implies a cooling schedule of
length \( \ell = E_{\text{max}} \beta / \log 2 \), but shorter schedules can be used
in most cases. Notice how the choice of a lower bound on the
Hamiltonian \( E^\ast \) affects these ratios. For this reason, it is
desirable that \( -E^\ast \) be as close as possible to the true
ground state energy of the system to decrease the length of
the cooling schedule.

The partition function is expressed as a product
\[
Z(\beta_k) = Z(\beta_0) F_0 F_1 \ldots F_{k-1}
\]
and will be computed by evaluating each fraction and using
the fact that \( Z(\beta_0) = D \), the dimension of the system’s
Hilbert space. If each ratio \( F_k \) is evaluated within accuracy
\( \epsilon \), the resulting relative error on \( Z(\beta) \) will be \( O(\epsilon) \). Our
universal thermalization process requires evaluating the \( F_k \)
sequentially: the fraction \( F_1 \) is evaluated from the Gibbs
state \( \rho(\beta_k) \), and the value of \( Z(\beta_k) \) is needed to prepare
\( \rho(\beta_{k+1}) \).

Our method makes use of quantum phase estimation
(QPE) \([10,11]\). This is a transformation that operates on
the “system” register and an \( m \)-qubit “energy” register initialized
to the state \( |0_m \rangle \) as follows:
\[
|\alpha\rangle \overset{U^*}{\longrightarrow} |a\rangle \overset{W}{\longrightarrow} \overset{\text{FT}}{\longrightarrow} \sum_E f(E, E_\alpha)|E\rangle
\]
Here, \( W \) denotes the Hadamard transform \( W|0_m \rangle = \frac{\sqrt{m}}{2} \sum_{r=0}^{2^m-1} |r\rangle \) and FT denotes the Fourier transform.
The central gate applies \( r \) repetitions of the time evolution
operator \( U \) to the system register, where \( r \) is dictated by the
state of the lower register. Since \( r \) can be as large as \( 2^m \),
the running time of this procedure is \( 2^m \). The label \( E \) runs
over the discretization of the interval \( [0, E_{\text{max}}) \), i.e., \( E \in \frac{E_{\text{max}}}{2^m} \times \{0, 1, \ldots, 2^m - 1\} \). The function \( f(E, E_\alpha) \) is the
discrete Fourier transform of \( g(r) = e^{-\mathrm{i} E r} \), so it is highly
peaked around the value \( E = E_\alpha \), i.e.,
\[
|f(E, E_\alpha)|^2 = \frac{\sin(E - E_\alpha)}{E_{\text{max}}} \frac{\pi 2^m}{\sin(\frac{E_{\text{max}}}{4})}.
\]

Thus, QPE essentially measures the energy of the system,
and writes it down on the energy register, up to some
fluctuations associated to the width \( f(E, E_\alpha) \).

Let us first describe our algorithm to prepare Gibbs
states by ignoring these fluctuations and return to them
later. In fact, our algorithm will prepare purified marked
Gibbs states
\[
|\beta_k\rangle = \sum_a \sqrt{Z(\beta_k)} |\rho_a\rangle \otimes |E_a\rangle
\]
that are defined on one system register, one scratchpad
register, and one energy register. The term “purified”
refers to the fact that the state $|\beta_i\rangle$ is pure and has a reduced system density matrix equal to the Gibbs state at inverse temperature $\beta_i$. The term “marked” means that the energy register encodes the energy of each system eigenvalue.

To obtain the state $|\beta_0\rangle$, we must first prepare the system and scratchpad in any maximally entangled state

$$D^{-1/2} \sum_a |a\rangle \otimes |\phi_a\rangle$$

and then apply QPE to the system register (ignoring fluctuations of $f$). The finite temperature quantum Gibbs state $|\beta_j\rangle$ is obtained from $|\beta_0\rangle$ and an ancillary qubit initially in the state $|0\rangle$ that we rotated by an angle $\theta(E, \beta_j) = \arcsin(e^{-(\beta_j E/2)})$ conditioned on the energy register, resulting in

$$|\Psi_j\rangle = D^{-1/2} \sum_a |a\rangle \otimes |\phi_a\rangle \otimes |E_a\rangle \otimes |\theta(E_a, \beta_j)\rangle$$

(4)

$$= \sqrt{Z(\beta_j) / D} |\beta_j\rangle |0\rangle + \ldots$$

(5)

where $|\theta\rangle = \sin\theta|0\rangle + \cos\theta|1\rangle$ and the ellipsis represents terms in which the ancilla qubit is in the state $|1\rangle$.

We can use Grover’s algorithm [12] to amplify the overlap of $|\Psi_j\rangle$ with the projector $\Pi_0 = I \otimes |0\rangle\langle 0|$ associated to the subspace where the ancillary qubit is in the state $|0\rangle$. Indeed, applying $q = \lceil D/Z(\beta_j) \rceil$ times the sequence of two reflections $(I - 2 |\Psi_j\rangle\langle \Psi_j|)(I - 2 \Pi_0)$ to the state $|\Psi_j\rangle$ will boost the amplitude of the term $|\beta_j\rangle \otimes |0\rangle$ from its initial value $\sqrt{Z(\beta_j) / D}$ to nearly 1. However, $q$ cannot be computed exactly because at this stage of the algorithm, $Z(\beta_j)$ is known only within relative accuracy $\frac{1}{2\epsilon}$. This is nonetheless sufficient to amplify the amplitude of the term $|\beta_j\rangle \otimes |0\rangle$ to $O(1)$. An amplitude $1 - O(\frac{1}{\epsilon})$ can be achieved using fixed-point search [13,14] at an additional multiplicative cost of $\log \epsilon$.

The fluctuations of $f(E_a, E_a)$ will in general invalidate the procedure we have described. If we perform QPE with a slightly larger number of qubits—i.e., if we use an $(m + 4)$-qubit energy register—the probability that the estimated energy $E$ deviates from its true value $E_a$ by more that $2^{-m}/t$ will be less than $1/16$ [11]. To further suppress these fluctuations, we can perform $\eta$ independent QPE procedures and compute the median $M$ of the $\eta$ results. The probability that this median value deviates from the true energy by more than $2^{-m}/t$ is less than $2^{-\eta}$ [15].

The evaluation of the median can be done by a coherent quantum process. There is an efficient quantum circuit on $\eta$ energy registers and one $m$-qubit “median” register that maps $|E(1)\rangle \otimes |E(2)\rangle \otimes \ldots \otimes |E(\eta)\rangle \otimes |0_m\rangle$ to $|E(1)\rangle \otimes |E(2)\rangle \otimes \ldots \otimes |E(\eta)\rangle \otimes |M(E(1), E(2), \ldots)\rangle$. From the result stated above, we know that on input eigenstate $|a\rangle$, only two median register states $|E_a\rangle$ and $|E_{a'}\rangle$ encoding the closest $m$-bit estimates of $E_a$ can have an amplitude of magnitude larger than $\sqrt{2^{-\eta}}$. This can be used to prepare a “good enough” version of the infinite temperature purified marked Gibbs state

$$|\tilde{\beta}_0\rangle = D^{-1/2} \sum_a |a\rangle \otimes (\alpha_a |\phi_a\rangle |E_a\rangle + \alpha_{a'} |\phi_{a'}\rangle |E_{a'}\rangle) + |\text{bad}\rangle$$

(6)

where $|\alpha_a|^2 + |\alpha_{a'}|^2 \geq 1 - 2^{-\eta}$, $|E_a - E_{a'}| \leq 2^{-m}/t$, and $|\text{bad}\rangle$ represent the components where the median deviates from the true energy by more than $2^{-m}/t$, so $|\langle\text{bad}|\text{bad}\rangle| \leq 2^{-\eta}$. The scratchpad states $|\phi_{a'}\rangle$ contain the original states $|\phi_a\rangle$ used to purify the infinite temperature Gibbs state and the $\eta$ energy registers used for the computation of the median, whose value is encoded in the third register.

With $m = \log_2(\beta t/\epsilon)$ the error caused on the Gibbs state by the roundoff of the energy estimate will be $O(\frac{1}{\epsilon})$ (see Appendix C in the supplementary material [9]). Similarly, this roundoff error will lead to a $O(\frac{1}{\epsilon})$ relative error on our estimate of $F_k$ (see Appendix B in the supplementary material [9]). The error due to the $|\text{bad}\rangle$ component of Eq. (6) requires a bit more attention. The reason is that the relative weight of the good and the bad component [i.e., first and second term of Eq. (6)] is in general not preserved by Grover’s amplification. Although the norm of $|\text{bad}\rangle$ starts out small, it can increase by a factor $e^{\beta |E_{\text{max}}|/2}$ during the amplification process. Setting $\eta = \ln(1 + |\beta_{E_{\text{max}}}|)/(\ln 2)$ insures that the norm of $|\text{bad}\rangle$ will remain bounded by $\frac{1}{\epsilon}$ after the amplification process, so can safely be ignored. This completes our presentation of the universal thermalization process, and demonstrates that the Gibbs state can be prepared with accuracy $\delta = \frac{1}{\epsilon}$ by a local process in time $T_{\text{th}}(\beta) \equiv O(1/|\ln(\frac{1}{\delta} + 2\beta |E_{\text{max}}|)|)$. We now describe how the state $|\beta_k\rangle$ can be used to estimate the ratio $F_k$. Similarly to the procedure that led to Eq. (5), we append to the quantum Gibbs state $|\beta_k\rangle$ yet another ancillary qubit in the state $|0\rangle$ and rotate it by an angle $\theta(E, \Delta \beta_{k+1})$ conditioned on the energy register, resulting in

$$|\Phi_k\rangle = \sum_a \sqrt{e^{-\beta_k E_a} Z(\beta_k)} |a\rangle \otimes |\phi_a\rangle \otimes |E_a\rangle \otimes |\theta(E_a, \Delta \beta_{k+1})\rangle$$

(7)

$$= \sqrt{F_k} |\beta_{k+1}\rangle |0\rangle + \ldots$$

where the ellipsis represent terms in which the ancilla qubit is in the state $|1\rangle$. Clearly, the squared norm of $\Pi_0|\Phi_k\rangle$ is equal to the quantity $F_k$ we are trying to estimate. Thus, $F_k$ can be estimated by quantum counting [16], i.e., using phase estimation to estimate the eigenvalue of the unitary matrix $(I - 2 |\Phi_k\rangle\langle \Phi_k|)(1 - 2 \Pi_0)$ associated to the Jordan block that supports $|\Phi_k\rangle$. An accuracy $\frac{1}{\epsilon}$ requires $\frac{1}{\epsilon}$ uses of the circuit that prepares $|\Phi_k\rangle$, so the total time required to evaluate the partition function with accuracy $\epsilon$ with constant success probability is $O(\sqrt{D \beta |E_{\text{max}}| / \epsilon})$.

More generally, any thermalization process can be used to evaluate $F_k$. Indeed, a purified Gibbs state can be prepared by simulating the thermalization process with a quantum computer, substituting the system’s maximally
mixed state $\frac{1}{D} \sum_i |a_i\rangle \otimes |\phi_{ac}\rangle$. Hence, this requires a time proportional to the thermalization time of the system. This state can be marked using multiple QPEs and evaluating their median to the thermalization time of the system. This state can be naturally coincides with the running time found in [17] to compute the ground state energy of the system. At higher temperatures however, the algorithm can be significantly faster. Surprisingly, the running time is shortest at the critical field $g/J = 1$, where the correlations in the system are the strongest.

For a classical Hamiltonian, the energy of a given configuration can be computed exactly. In that case, our thermalization process produces a Gibbs state with accuracy $\delta$ in time $\sqrt{\frac{D}{Z(\beta_0)}} \log \frac{1}{\delta}$. The only approximation to the partition function involved in that case comes from quantum counting, so $Z(\beta)$ can be evaluated with accuracy $\epsilon$ in time $O(\sum_i \frac{\epsilon}{\delta} \sqrt{\frac{D}{Z(\beta_i)}})$. This should be compared to the time $O(\sum_i \frac{1}{\delta_i} \sqrt{\frac{D}{Z(\beta_i)}})$ required by the quantum simulated annealing algorithm [1,2], where $\delta_i$ is the spectral gap of a Markov chain whose fixed point is the Gibbs state of the system at inverse temperature $\beta_i$. Thus, our method offers a speed-up for the computation of a classical partition function whenever no such Markov chain can be found with $\delta \leq Z(\beta)/D$.

Conclusion.—We have presented a quantum algorithm to prepare the Gibbs state of a quantum system in a time that grows like the $\alpha$th power of the system’s Hilbert space dimension, where $\alpha \leq \frac{1}{2}$ is proportional to the Helmholtz free energy density $h(\beta)$ at inverse temperature $\beta$. This sets a universal upper bound on the thermalization time of quantum systems. However, our universal thermalization process fails to recognize special properties of the system—e.g., large energy gap, absence of long range correlations—that could potentially speed-up the computation. We are actively investigating this problem.

We have demonstrated how the ability to thermalize a quantum system on a quantum computer leads to an algorithm to estimate its partition function with accuracy $\epsilon$ in a time proportional to the system’s thermalization time and $\epsilon^{-2}$. Our method provides a complete quantum generalization of the work reported in [1,2] for partitions functions of classical systems and may provide an additional speed-up when there exist no rapidly mixing Markov chain to prepare the system’s Gibbs state.