A quantum–quantum Metropolis algorithm

Man-Hong Yung (翁文康) and Alán Aspuru-Guzik[1]

Department of Chemistry and Chemical Biology, Harvard University, Cambridge MA, 02138

Edited by Frank A. Wilczek, Center for Theoretical Physics, Cambridge, MA, and approved November 16, 2011 (received for review July 20, 2011)

The classical Metropolis sampling method is a cornerstone of many statistical modeling applications that range from physics, chemistry, and biology to economics. This method is particularly suitable for sampling the thermal distributions of classical systems. The challenge of extending this method to the simulation of arbitrary quantum systems is that, in general, eigenstates of quantum Hamiltonians cannot be obtained efficiently with a classical computer. However, this challenge can be overcome by quantum computers. Here, we present a quantum algorithm which fully generalizes the classical Metropolis algorithm to the quantum domain. The meaning of quantum generalization is twofold: The proposed algorithm is not only applicable to both classical and quantum systems, but also offers a quantum speedup relative to the classical counterpart. Furthermore, unlike the classical method of quantum Monte Carlo, this quantum algorithm does not suffer from the negative-sign problem associated with fermionic systems. Applications of this algorithm include the study of low-temperature properties of quantum systems, such as the Hubbard model, and preparing the thermal states of sizable molecules to simulate, for example, chemical reactions at an arbitrary temperature.

Metropolis method | statistical physics | quantum computing | quantum simulation | simulated annealing

Interacting many-body problems cover a wide range of applications in many areas in science and technology. These are best exemplified by the Ising spin model, which was originally developed for explaining ferromagnetism and its associated phase transitions (1). The Ising model was later found to be related to many practical applications, for example, error-correcting codes, image restoration, associative memory, and optimization problems (see, e.g., ref. 2). However, there is no efficient method for finding solutions to many-body problems in general. For example, the problem of finding the ground state of the Ising model with arbitrary couplings is known to be a nondeterministic polynomial (NP)-complete problem, meaning that, if an efficient algorithm for solving the Ising model exists, then all of the problems in the class of NP, for example the graph isomorphism problem and some variants of the traveling salesman problem, can be solved efficiently as well (see, e.g., ref. 3).

The most fundamental challenge for solving many-body problems is that they generally require an exponentially large amount of spatial and temporal computing resources to find the solutions as the system size increases (4). Although no universal method for solving general many-body problems has been found, powerful classical methods such as Markov-chain Monte Carlo (MCMC) or quantum Monte Carlo (QMC) have been invented and proven to be successful in many applications (see, e.g., ref. 5). These methods, however, have certain limitations. For example, the running time of MCMC scales as $O(1/\delta)$ (6), where δ is the gap of the transition matrix. For problems such as spin glasses (7) where δ is vanishingly small as a function of system size, MCMC becomes computationally inefficient. QMC methods, on the other hand, suffer from the negative-sign problem (8), making them inefficient for problems involving fermions, e.g., electronic problems. Despite the many efforts toward improvement that have been made (9), this limitation is still one of the biggest challenges in QMC (10).

Background of Quantum Simulation

One of the most important goals in the field of quantum computation, as proposed by Feynman (11), is to look for methods or algorithms that can solve these many-body problems more efficiently. A promising solution is quantum simulation, which aims at employing a controllable quantum system to simulate the behaviors of the target quantum system. Quantum simulation can solve the problem of the spatial requirement for solving many-body problems. Quantum simulation can be implemented either by dedicated (or analog) quantum simulators (12) or with universal (or digital) quantum computers (13). For analog quantum simulation, high-precision experimental techniques are required for faithful simulation, and therefore involves many engineering challenges. An example is the use of neutral atoms trapped in optical lattices (14) to simulate the low-temperature properties of the Hubbard or Bose–Hubbard model.

Digital Quantum Simulation. To look for algorithms based on the special properties of the laws of quantum mechanics, top-down approaches aim to improve the existing classical algorithms by combining them with elementary quantum algorithms. For the bottom-up approach, in the context of quantum simulation, the only known example that quantum algorithms can achieve exponential gain over classical methods is the simulation of time dynamics (15, 16). It was also discovered (17) that such an approach is particularly suitable for simulating single or many-particle systems where the Hamiltonians consist of two terms, namely kinetic energy and potential energy, for example, atomic or molecular systems (18). On the other hand, however, attempt to create quantum algorithms to solve the ground-state (19, 20) or thermal-state (21–23) problems of unstructured Hamiltonians failed to show an exponential gain over classical approaches. However, this lack of an exponential advantage does not mean that quantum computers fail to show advantages over classical computers in quantum simulation.

Variational Methods. In fact, many problems in physics and chemistry do exhibit certain symmetries or structures. These features allow us to take the top-down approach to generalize methods that are proven to be successful in classic computing to design quantum algorithms. For example, the variational wavefunctions, such as the Bardeen–Cooper–Schrieffer wavefunction in superconductivity and the Hartree–Fock solutions in the electronic problems of atomic and molecular systems, are good approximations to the exact ground states of the corresponding Hamiltonians to some extent. With the quantum phase estimation algorithm (24, 25), it is possible to project the exact ground states from the approximate solutions with high probabilities. Further theoretical investigations on the efficiency of this method on molecular systems have been made (26–28), and a related experimental implementation has been realized (29).

Author contributions: M.-H.Y. and A.A.-G. designed research, performed research, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]To whom correspondence should be addressed. E-mail: aspuru@chemistry.harvard.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1111758109/-/DCSupplemental.

Markov-Chain-Based Methods. For problems where clever approximating solutions are unavailable, one may still be able to solve them efficiently by using Markov-chain-based Monte Carlo methods. In the context of quantum computation, Szegedy described a method to quantize classical Markov chains (30). The key result is that a quadratic speedup $O(1/\sqrt{\delta})$ in the gap δ of the transition matrix is possible. This idea was later adapted to other algorithms (31, 32) for preparing thermal states of classical Hamiltonians, based on the idea of quantum simulated annealing (QSA) (31).

In fact, Markov chains that correspond to classical Hamiltonians are relatively easy to construct, as all the eigenstates of classical Hamiltonians are in the computational basis $|x\rangle$. However, preparation of the thermal states of quantum Hamiltonians (20, 21, 33, 34), where the eigenvectors are in nontrivial superposition of the computational basis vectors, is a much more challenging problem for classical computation. Classically, one would need to solve for the full eigenvalue problem (i.e., look for all eigenvalues and eigenvectors) for the quantum Hamiltonian first, which often requires more computational resources than directly computing the corresponding partition functions. Quantum mechanically, any attempt for a straightforward generalization of the classical method by Szegedy would encounter the problem due to the inability to clone quantum states (35).

Quantum Metropolis Methods. The key strategy that has been employed so far to overcome the difficulty of obtaining information about eigenvectors of quantum Hamiltonians with a quantum computer is the use of the quantum phase estimation algorithm (24); the eigenvalues of a quantum Hamiltonian can be recorded without explicitly knowing the detailed structure of the eigenvectors. Based on this idea, Terhal and DiVincenzo (33) made an attempt to extend the Metropolis algorithm (5) to the quantum domain in order to avoid the negative-sign problem in QMC. However, one limitation of their result is that the key step involves too many energy nonlocal transitions, making it inefficient for solving practical problems. Recently, Temme et al. (34) addressed this problem by introducing a measurement-based rejection method, which makes the algorithm more realistic for simulating many-body problems. This result shows that quantum computers can solve many-body problems beyond the limits of classical computers. However, the underlying Markov chains of this algorithm are still classical in nature, which means that the scaling of the running time is still $O(1/\delta)$.

Motivation, Results, and Outline of the Quantum–Quantum Metropolis Algorithm (Q^2MA)

To summarize, before this work, the existing quantum algorithms were capable of either (*i*) preparing thermal states of classical systems with quantum speedup (31, 32), or (*ii*) preparing thermal states of quantum systems without quantum speedup (33, 34); this suggests that the potential of quantum computers may not have been fully exploited. This problem motivates us to further explore the power of quantum computation by providing a quantum algorithm that allows us to prepare thermal states of quantum systems with quantum speedup.

Instead of following the path developed in the work of Temme et al. (34), we took the following strategy: First of all, a natural question to ask is whether Szegedy's quantum speedup is restricted to employing classical Hamiltonians, where the eigenstates simply coincide with the computational basis, making cloning of the classical information possible. In this manuscript, we extend the Szegedy approach to be able to deal with quantum Hamiltonians; we introduce a quantum version of the method of Markov-chain quantization which is combined with the QSA procedure (31). Formally, we propose a quantum Metropolis algorithm that exhibits a quadratic quantum speedup as a function of the eigenvalue gap δ of the corresponding Metropolis Markov matrix M for any quantum Hamiltonian H. Note that, because classical Hamiltonians belong to a subclass of quantum Hamiltonians, our algorithm is also applicable to classical systems.

Lastly, we note that this algorithm is still limited by the Markov-chain gap δ, which implies that there are problems, especially ground-state problems of frustrated systems, where this algorithm cannot solve efficiently. This result is expected, as otherwise, hard problems (20) in the quantum Merlin–Arthur complexity class (and even NP) would be solved by this algorithm. Whether such a quantum algorithm could exist is still one of the major open questions in quantum information science.

Summary of the Main Results. The quantum Metropolis algorithm constructs the thermal state, $\rho_{\text{th}} = e^{-\beta H}/\text{Tr}(e^{-\beta H})$, for any given quantum (or classical) Hamiltonian H, where the system to be simulated is prepared at a given temperature T ($\beta \equiv 1/k_{\text{B}}T$), with a quadratic quantum speedup $O(1/\sqrt{\delta})$ in the gap δ of the corresponding Markov matrix M. We call this Q^2MA because it shows a quantum speedup for the Markov chain of thermal-state preparation for a quantum system being simulated. For comparison, the construction method by Temme et al. (34) is restricted by the no-cloning theorem insofar as the information of a eigenstate cannot be retrieved after accepting the proposed move in the Metropolis step. We relax this restriction by adopting a dual representation where the basis states consist of pairs of eigenstates related by time-reversal operation (see Eq. **8**). For the cases where the physical system being simulated is time-reversal invariant, the pair of the eigenstates forming the basis vector become identical. This construction effectively circumvents the restriction imposed by the no-cloning theorem.

Outline of the Key Ideas of Q^2MA. The goal of the proposed Q^2MA is to prepare the coherent encoding of the thermal state (CETS),

$$|\alpha_0\rangle = \sum_{i=0}^{N-1} \sqrt{e^{-\beta E_i}/\mathcal{Z}}\,|i\rangle, \qquad [1]$$

where $|i\rangle \equiv |\varphi_i\rangle|\tilde{\varphi}_i\rangle|0\rangle$, the state $|\varphi_i\rangle$ is the eigenstate of a quantum Hamiltonian H, associated with the eigenvalue E_i, and $\mathcal{Z} = \text{Tr}(e^{-\beta H})$ is the partition function. The state $|\tilde{\varphi}_i\rangle$ is the time-reversal counterpart (complex conjugate) of $|\varphi_i\rangle$, which is the eigenstate of the corresponding time-reversal Hamiltonian $\tilde{H} \equiv H^*$ with the same eigenenergy E_i. After tracing out all of the ancilla quantum bits (qubits), the CETS is equivalent to $\rho_{\text{th}} = (1/\mathcal{Z})\sum_i e^{-\beta E_i}|\varphi_i\rangle\langle\varphi_i|$. An outline of this algorithm is shown in Fig. 1. The steps go as follows: A maximally entangled state $N^{-1/2}\sum_{x=0}^{N-1}|x\rangle|x\rangle$ is the input state of the algorithm. Alternative input states are possible as long as they have a finite overlap with the CETS, but for simplicity we will focus on only one particular type of input state here. This maximally entangled state can be mathematically transformed into a state (see Eq. **8**) that spans the dual basis related to the Hamiltonian H to be simulated. Note that this state represents the infinite temperature $\beta = 0$ state. Next, a generalization of Szegedy's operator W (see Eq. **12**) will be constructed and applied to the input state. The main purpose of W is to project out the CETS at a lower temperature through the quantum phase estimation algorithm. A similar procedure is then repeated for projecting out a CETS at an even lower temperature, and so on. The whole procedure for "adiabatically" preparing from a high-temperature CETS to a lower-temperature one is known as QSA (31). When the ancilla qubits are traced out, we obtain the thermal density matrix $\rho_{\text{th}} = e^{-\beta H}/\text{Tr}(e^{-\beta H})$, which accomplishes the goal of the Q^2MA.

Organization of the Paper. We first summarize the features of Markov chains, Metropolis sampling, and Szegedy's method that are related to this work, in a way that become applicable to both

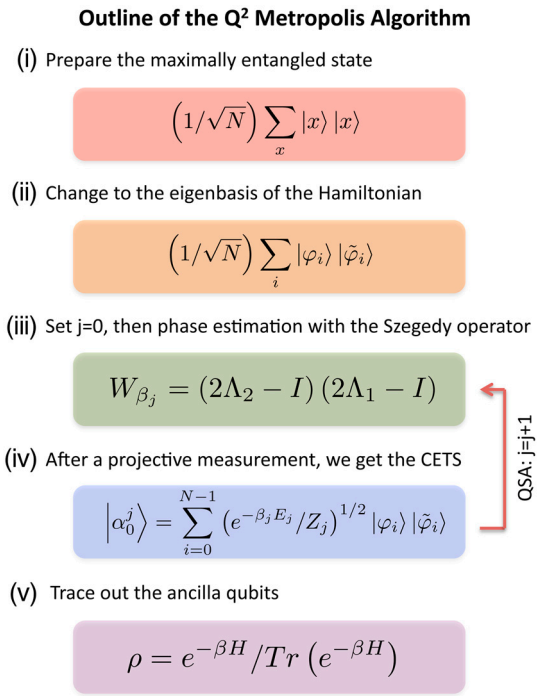

Fig. 1. Outline of the Q²MA for preparing thermal states of quantum Hamiltonians H. (*i*) A maximally entangled state, $(1/\sqrt{N})\sum_x|x\rangle|x\rangle$, of n qubits is first prepared. (*ii*) Then, it is mathematically transformed into a maximally entangled state, $(1/\sqrt{N})\sum_i|\varphi_i\rangle|\tilde{\varphi}_i\rangle$ (see Eq. **8**), associated with the Hamiltonian H. (*iii*) A schedule of QSA (31) is then applied to this state, lowering the simulated inverse temperature from zero to β. The QSA algorithm is based on the phase estimation algorithm using the Szegedy operator W defined in Eq. **12**. (*iv*) The resulting state is a CETS described in Eq. **1**. (*v*) The final state becomes the thermal state $\rho = e^{-\beta H}/Tr(e^{-\beta H})$, after tracing out all of the ancilla qubits.

classical and quantum systems. Next, we outline the generalization of the method of Markov-chain quantization by Szegedy (30) to make it applicable to quantum Hamiltonians H. Then, we analyze the structure of the Szegedy operator W in detail. Finally, we combine the results above with the method of quantum simulated annealing (31) to generate thermal states of quantum systems at any finite temperature T. We describe nonessential technical details in the *SI Text*.

Review of the Classical Metropolis Method

In the standard Metropolis scheme as employed in classical computation (5), the thermal Gibbs distribution $e^{-\beta E_i}/\mathcal{Z}$ of certain a eigenstate $|\varphi_i\rangle$ of a certain Hamiltonian H, associated with the eigenenergy E_i, at an inverse temperature $\beta = 1/k_B T$, is generated through a Markov-chain procedure, in which the matrix element m_{ij} of the corresponding Markov transition matrix M refers to the transition probability from the eigenstate $|\varphi_i\rangle$ to $|\varphi_j\rangle$. The equilibrium (stationary) distribution of the Markov-chain procedure, $\pi_i \equiv e^{-\beta E_i}/\mathcal{Z}$, satisfies the following detailed balance condition:

$$\pi_i m_{ij} = \pi_j m_{ji}. \qquad [2]$$

For classical Hamiltonians, the eigenstate vector $|\varphi_i\rangle$ is simply one of the vectors in the computational basis $\{|x\rangle\}$, but for quantum Hamiltonians, $|\varphi_i\rangle$ is, in general, a nontrivial superposition of the computational basis vectors.

The Metropolis Filter. The Metropolis method solves problems by using an efficient way to construct a Markov chain so that the final distribution of certain random variables is identical to the distribution $\{\pi_i\}$. A well-known solution that can satisfy the detailed balance condition [**2**] is given by the following choice: $m_{ji} = s_{ij}z_{ji}$, where $s_{ij} = s_{ji}$ is any symmetrical transition probability, and

$$z_{ij} = \min\{1, e^{-\beta(E_j - E_i)}\} \qquad [3]$$

is sometimes called the Metropolis filter, which determines the acceptance or rejection of a proposed change of the eigenstates by comparing the energy difference between them.

In a practical implementation of the Metropolis method (e.g., when applied to the Ising model), one starts with an arbitrary initial configuration of spins and then applies a random transition (e.g., single spin-flip) from one configuration (eigenstate) to another, and compares the energy between the new eigenstate and the old one. According to the rule in Eq. **3**, if the new eigenstate has a lower energy, then the move is accepted. Otherwise, the move is accepted only with a probability distribution given by the ratio of the corresponding Boltzmann factors—i.e., $e^{-\beta(E_j - E_i)}$. It is important to note that, in the whole operation, the explicit knowledge of the partition function $\mathcal{Z}$, which usually cannot be computed efficiently, is not required.

Performance of the Metropolis Method. The performance of the Metropolis method depends on many factors. For many cases, it depends on the properties of the underlying transition matrix M of the Markov chain, especially the eigenvalues λ_k, which are all positive and bounded by one, and the largest eigenvalue is always equal to one. For convenience, we order them as $\lambda_0 = 1 > \lambda_1 \geq \cdots \geq \lambda_{N-1} > 0$. The problem is solved when the Markov-chain procedure produces the stationary distribution $\{\pi_i\}$. The convergence time of a Markov chain is limited by the eigenvalue gap $\delta \equiv 1 - \lambda_1$ of the transition matrix M as $O(1/\delta)$ (6). For problems such as spin glasses (7), this gap can be very small, and hence it will necessarily take a long time to solve. The purpose of this work is to show that the running time of quantum algorithms can be improved to $O(1/\sqrt{\delta})$, and that we can fully extend the Metropolis sampling algorithm into the quantum domain (i.e., for quantum Hamiltonians and with quantum speedup).

Summary of Szegedy's Method

The idea of Szegedy's method (30) of Markov-chain quantization is to construct a unitary operator W that contains the information of a Markov matrix, which turns out to have two important properties: (*i*) the CETS of Eq. **1** is one of the eigenstates of W, and (*ii*) the eigenvalue gap scales $O(\sqrt{\delta})$. These two properties are important for generating classical thermal states through annealing (31, 32).

The first step of Szegedy's method is to encode the information about the transition matrix elements m_{ij} into a pair of unitary operators, which acts on two sets of qubits:

$$U_X|i\rangle|0\rangle = \sum_j \sqrt{m_{ij}}\,|i\rangle|j\rangle, \qquad [4]$$

$$U_Y|j\rangle|0\rangle = \sum_i \sqrt{m_{ji}}\,|i\rangle|j\rangle, \qquad [5]$$

where $|i\rangle$ and $|0\rangle \equiv |000\ldots\rangle$ are state vectors in the computational basis. Let us define a more compact notation $|i,0\rangle \equiv |i\rangle|0\rangle$. Using the detailed balance condition in Eq. **2**, we have $\langle i,0|U_X^\dagger U_Y|j,0\rangle = \sqrt{m_{ij}m_{ji}} = \sqrt{\pi_i}m_{ij}/\sqrt{\pi_j}$, which implies that, within the subspace $\{|i,0\rangle\}$, the matrix $U_X^\dagger U_Y$ is therefore a similar matrix to the Markov matrix M (i.e., same eigenvalue spectrum $\{\lambda_k\}$). Based on this property, the Szegedy operator is defined as

$$W \equiv (2\Lambda_2 - \mathcal{I})(2\Lambda_1 - \mathcal{I}), \quad [6]$$

where $\mathcal{I}$ is the identity operator, $\Lambda_1 \equiv \sum_i |i,0\rangle\langle i,0|$, and $\Lambda_2 \equiv U_X^\dagger U_Y \Lambda_1 U_Y^\dagger U_X$. Now define a set of eigenvectors $|\phi_k\rangle$ such that $\Lambda_1 U_X^\dagger U_Y \Lambda_1 |\phi_k\rangle = \lambda_k |\phi_k\rangle$. By direct substitution, the Szegedy operator is block diagonal and can be decomposed into 2×2 matrices in the subspace $\{|\phi_k\rangle, U_X^\dagger U_Y |\phi_k\rangle\}$:

$$w_k = \begin{bmatrix} \cos(2\theta_k) & -\sin(2\theta_k) \\ \sin(2\theta_k) & \cos(2\theta_k) \end{bmatrix}, \quad [7]$$

where $\cos\theta_k \equiv \lambda_k$. For $k = 0$ (i.e., $\lambda_0 = 1$), we have $w_0|\phi_0\rangle = |\phi_0\rangle$, where $|\phi_0\rangle = \sum_i \sqrt{\pi_i}\,|i,0\rangle$ contains the stationary distribution of the Markov chain (see *SI Text*), which is point (*i*) above. On the other hand, the eigenvalues of w_k are $\exp(\pm 2i\theta_k)$, and the eigenvalue gap $\Delta \equiv |1 - e^{2i\theta_1}|$ satisfies the following inequality: $\Delta = 2\sqrt{1-\lambda_1^2} \geq 2\sqrt{\delta}$, which corresponds to point (*ii*) and leads to the quadratic speedup $O(1/\sqrt{\delta})$ (31, 32).

Generalization of the Markov-Chain Quantization Method to Quantum Hamiltonians

The Markov-chain quantization method by Szegedy (30) discussed above is applicable to classical Hamiltonians only, because it was formulated in the computational basis. To extend Szegedy's method to make it applicable to quantum Hamiltonians, we start with the following procedure: first we prepare a set of n qubits initialized in the state $(1/\sqrt{2})^n(|0\rangle + |1\rangle)^{\otimes n}$, or equivalently $N^{-1/2}\sum_{x=0}^{N-1}|x\rangle$, where $N = 2^n$. By performing a bit-by-bit controlled-NOT gate on a set of n ancilla qubits initialized in the state $|000...0\rangle$, we obtain the maximally entangled state $N^{-1/2}\sum_{x=0}^{N-1}|x\rangle|x\rangle$, which can be formally expressed as an entangled state of the form (*SI Text*)

$$|\phi_0\rangle \equiv \frac{1}{\sqrt{N}}\sum_{i=0}^{N-1}|\varphi_i\rangle|\tilde{\varphi}_i\rangle, \quad [8]$$

where $|\varphi_i\rangle = \sum_{x=0}^{N-1}\langle x|\varphi_i\rangle|x\rangle$ is an energy eigenstate of H (i.e., $H|\varphi_i\rangle = E_i|\varphi_i\rangle$), and $|\tilde{\varphi}_i\rangle \equiv \sum_{x=0}^{N-1}\langle\varphi_i|x\rangle|x\rangle$ is the time-reversal counterpart (complex conjugate) of $|\varphi_i\rangle$, which is the eigenstate of the corresponding time-reversal Hamiltonian $\tilde{H} \equiv H^*$ with the same eigenenergy E_i (i.e., $\tilde{H}|\tilde{\varphi}_i\rangle = E_i|\tilde{\varphi}_i\rangle$). As we shall see, the pair structure $|\varphi_i\rangle|\tilde{\varphi}_i\rangle$ of the eigenvectors provides a unique feature for our construction of the quantum algorithm. For example, through the phase estimation algorithm (PEA), the value of eigenvalue E_i can be obtained either from $|\varphi_i\rangle$ or $|\tilde{\varphi}_i\rangle$; this is the key property introduced in this paper, which relaxes the constraints of the previous quantum Metropolis algorithm (34) in applying it to Szegedy's method.

In the following, for the purpose of clarity, we shall assume that the PEA can be applied perfectly, in the sense that each eigenstate can be uniquely identified by a unique eigenvalue. It turns out that the problem of degeneracy of the eigenenergies of a quantum Hamiltonian alone is not necessarily a hurdle for the construction of our algorithm. We leave our discussion on the effects of degeneracy on this algorithm to the *SI Text*.

Encoding the Information of the Markov Chain. We have prepared the initial input state **[8]** of the Q^{2}MA. Now, we have to define the required operations: Let us include in our Hilbert space an extra qubit initialized in the state $|0\rangle$ and define a more compact notation, $|i\rangle \equiv |\varphi_i\rangle|\tilde{\varphi}_i\rangle|0\rangle$. In the Q^{2}MA, the information of a Markov chain (i.e., z_{ij}) is encoded in a pair of unitary operators U_X and U_Y in the following way (see Eqs. **4** and **5**, and see *SI Text* for their detailed construction):

$$U_X|i\rangle = \sum_k(\sigma_{ik}|\varphi_k\rangle|\varphi_i\rangle|0\rangle + \gamma_{ik}|\varphi_i\rangle|\varphi_k\rangle|1\rangle), \quad [9]$$

$$U_Y|j\rangle = \sum_m(\sigma_{jm}|\varphi_m\rangle|\varphi_j\rangle|0\rangle + \gamma_{jm}|\varphi_m\rangle|\varphi_j\rangle|1\rangle), \quad [10]$$

where $\sigma_{ik} \equiv \alpha_{k\tilde{i}}\sqrt{z_{ik}}$, $\gamma_{ik} \equiv \alpha_{k\tilde{i}}\sqrt{1-z_{ik}}$, and $\alpha_{k\tilde{i}} \equiv \langle\varphi_k|K|\tilde{\varphi}_i\rangle$. Here, K is an unitary operator, and like the spin-flip operator σ_x in the classical Metropolis method, K causes transitions from one eigenstate to another. Here, z_{ik} is the Metropolis filter defined in Eq. **3**. Note that U_Y is related to U_X by a controlled-SWAP operation.

The physical meaning of U_X is that, when the system is in the eigenstate $|\varphi_i\rangle$, the transition probability for it to make a transition to another eigenstate $|\varphi_k\rangle$ is $|\sigma_{ik}|^2 \propto z_{ik}$. This operation is consistent with the rules in the classical Metropolis method. The meaning of U_Y is not as clear as U_X; the way U_Y is constructed is mainly to make the combination $U_X^\dagger U_Y$ contain the same eigenvalue spectrum as the Markov matrix M (see Eq. **11**).

Connection with the Markov Matrix. To see an explicit connection between the operators U_X and U_Y with the Metropolis Markov matrix M, we will apply the detailed balance condition **[2]** to the product of $U_X^\dagger$ and U_Y, which are the key elements in constructing Szegedy's operator W (Eq. **12**). From Eqs. **9** and **10**, for $j \neq i$, we have $\langle j|U_X^\dagger U_Y|i\rangle = |\alpha_{\tilde{j}\tilde{i}}|^2(z_{ij}z_{ji})^{1/2}$, where we used $\langle\varphi_i|\tilde{\varphi}_j\rangle = \langle\varphi_j|\tilde{\varphi}_i\rangle$. Note that the right-hand side is equal to $(m_{ij}m_{ji})^{1/2}$. On the other hand, $\langle i|U_X^\dagger U_Y|i\rangle = |\alpha_{\tilde{i}\tilde{i}}|^2 + \sum_k|\alpha_{k\tilde{i}}|^2(1-z_{ik})$, which, as we shall see, can be interpreted as the probability of not undergoing a transition. Now, using Eq. **2**, we obtain the following decomposition:

$$\langle j|U_X^\dagger U_Y|i\rangle = \langle j|D_\pi^{1/2}|j\rangle\langle j|M|i\rangle\langle i|D_\pi^{-1/2}|i\rangle, \quad [11]$$

where $D_\pi \equiv \sum_{j=0}^{N-1}\pi_j|j\rangle\langle j|$ is a diagonal matrix, and $M \equiv \sum_{i,j}m_{ij}|j\rangle\langle i|$, with $m_{ii} \equiv \langle i|U_X^\dagger U_Y|i\rangle$ and $m_{ij} \equiv |\alpha_{\tilde{j}\tilde{i}}|^2 z_{ij}$ for $j \neq i$, is the transition matrix of the Markov chain. Within the subspace $\{|i\rangle\}$, Eq. **11** implies that $U_X^\dagger U_Y$ and M are similar matrices, which means that they have the same set of eigenvalues λ_k. The stationary distribution $\pi_i = e^{-\beta E_i}/\mathcal{Z}$ of M is the same as the Gibbs distribution for the quantum system to be simulated. This property is relevant to the construction of Szegedy's operator W, which is defined in the following section.

Construction of the Generalized Szegedy Operator

Now, we have shown in Eq. **11** that the product $U_X^\dagger U_Y$ is a similarity transform of a Markov matrix M within the subspace $\{|i\rangle\}$. We will need to take one more step to see how this property can lead us to be able to make good use of it. Following Szegedy (30), the results above allow us to construct a unitary operator (see Eq. **6**)

$$W \equiv (2\Lambda_2 - \mathcal{I})(2\Lambda_1 - \mathcal{I}), \quad [12]$$

where $\mathcal{I}$ is the identity operator, and the two projectors are defined by $\Lambda_1 \equiv \sum_{i=0}^{N-1}|i\rangle\langle i|$ and $\Lambda_2 \equiv U_X^\dagger U_Y \Lambda_1 U_Y^\dagger U_X$. Note that the structure of W is similar to those reflectors in other Grover-type algorithms. It therefore comes as no surprise that one can achieve a quadratic quantum speedup. However, the quantum speedup achieved here is quite different from that in the Grover-search problem, where the quantum speedup refers to the gain in the reduction of computational steps compared with the problem size. Here, the quantum speedup refers to the gain in the amplification of the Markov gap from δ to $\sqrt{\delta}$, which has no direct relationship with the problem size.

Table 1. Comparison of various Markov-chain-based algorithms for thermal-state preparation (see text for definitions)

Methods	Hamiltonian	Input	Output	Quantum speedup
Q Metropolis I (33)	quantum	any ρ	ρ_{th}	no
Q Metropolis II (34)	quantum	any ρ	ρ_{th}	no
Q Markov chain I (31)	classical	$\|+\rangle^{\otimes n}$	CETS I	quadratic
Q Markov chain II (32)	classical	$\|+\rangle^{\otimes n}$	CETS I	quadratic
Q^2MA (this work)	quantum	$\|\phi_0\rangle$	CETS II	quadratic

Spectral Properties of the Generalized Szegedy Operator. The Szegedy operator W will be employed to perform the projection of the CETS (see Eq. **1**) in the phase estimation algorithm. In the following, we will show that the "ground state" of W is indeed the CETS. The spectral properties of W can be seen in the following way: Define $|\alpha_k\rangle \equiv \sum_{i=0}^{N-1} a_{ki}|i\rangle$ to be the eigenvectors of $\Lambda_1 U_X^\dagger U_Y \Lambda_1$; the eigenvalue equation can be written as $\Lambda_1 U_X^\dagger U_Y |\alpha_k\rangle = \lambda_k |\alpha_k\rangle$. On the other hand, using the fact (see *SI Text*) that $\Lambda_1 U_X^\dagger U_Y \Lambda_1 = \Lambda_1 U_Y^\dagger U_X \Lambda_1$, we have $\Lambda_2 |\alpha_k\rangle = \lambda_k U_X^\dagger U_Y |\alpha_k\rangle$. These two equations suggest that, if we start with vectors within Λ_1, W can be block-diagonalized into subspace of 2×2 matrices w_k spanned by the basis $\{|\alpha_k\rangle, U_X^\dagger U_Y |\alpha_k\rangle\}$. Explicitly (see Eq. **7**),

$$w_k = \begin{bmatrix} \cos(2\theta_k) & -\sin(2\theta_k) \\ \sin(2\theta_k) & \cos(2\theta_k) \end{bmatrix}, \qquad [13]$$

where $\cos\theta_k \equiv \lambda_k$. Note that the eigenvalues of w_k is $e^{\pm i\theta_k}$. In the case of $k = 0$, where $\lambda_0 = 1$ (or $\theta_0 = 0$), $w_0 = \mathcal{I}$ is simply an identity matrix.

From Eq. **11** and the properties of the Markov matrix (see details in *SI Text*), the $k = 0$ state is the CETS in Eq. **1**. Recall that $|i\rangle \equiv |\varphi_i\rangle|\tilde{\varphi}_i\rangle|0\rangle$, the state $|\alpha_0\rangle$ becomes the Gibbs thermal state $\rho_{\text{th}} = e^{-\beta H}/\text{Tr}(e^{-\beta H})$ when the other qubits are traced out.

Quadratic Quantum Speedup. One of the most important features about W is that the minimum eigenvalue gap $\Delta_{\min} \equiv |2\theta_1|$ of W is larger than two times the square root of the gap $\delta \equiv 1 - \lambda_1$ of the transition matrix M (using $2\theta \geq |1 - e^{2i\theta}| = 2\sqrt{1 - \cos^2\theta}$):

$$\Delta_{\min} \geq 2\sqrt{\delta}, \qquad [14]$$

which is the mathematical origin of the quadratic speedup of Szegedy's algorithm. Recall that, for classical Markov-chain algorithms, the performance scales as $O(1/\delta)$ (6). As we shall see next, when we apply W in the phase estimation algorithm, the performance scales as $O(1/\sqrt{\delta})$.

This section completes our discussion on the necessary tools needed for the following discussion, and we are ready to go through the method of QSA (31).

Quantum Simulated Annealing

Our quantum simulation algorithm goes as follows: Given a quantum Hamiltonian H and any finite temperature T, the goal is to obtain the corresponding CETS of the form in Eq. **1**, from the initial state defined in Eq. **8**, which can be readily prepared from the state $|000...0\rangle$, and which can be considered as the infinite-temperature ($\beta = 0$) state. To achieve this goal, we can use the method of QSA (31). For completeness, we outline the basic strategy and summarize the related results in the following paragraph. Technical details are included in the *SI Text*.

Let us suppose that we aim to obtain the CETS corresponding to the temperature β. The strategy of the QSA method is to prepare a sequence, $j = 0,1,2,..,d$, of $d+1$ of coherent thermal states, $|\alpha_0^j\rangle = \sum_{i=0}^{N-1} (e^{-\beta_j E_i}/\mathcal{Z}_j)^{1/2}|i\rangle$, at a time; the temperature $\beta_j \equiv (j/d)\beta$ of the coherent thermal state is lowered in each step. The basic idea for implementing this procedure is that, for a sufficiently small change of β, $\Delta\beta \equiv \beta/d$, one can show that the CETS $|\alpha_0^j\rangle$ has a good overlap with another CETS $|\alpha_0^{j+1}\rangle$ at a lower temperature. Explicitly, the fidelity between the two states is bounded by the following expression: $|\langle\alpha_0^{j+1}|\alpha_0^j\rangle|^2 \geq 1 - \epsilon_0$, where ϵ_0 is bounded by $O(\Delta\beta^2\|H\|^2)$ (*SI Text*).

The next step in the QSA is to consider a projective measurement operator $\Pi_{k+1} = \sum_{k=0}^{N-1} |\alpha_k^{j+1}\rangle\langle\alpha_k^{j+1}|$ constructed in the eigenbasis of $|\alpha_k^{j+1}\rangle$. When Π_{k+1} is applied to the CETS $|\alpha_0^j\rangle$, it will result in the lower-temperature state $|\alpha_0^{j+1}\rangle$ with a high probability. The same procedure is then repeated for the other CETS at lower temperatures. For the entire process,

$$|\alpha_0^0\rangle \xrightarrow{\Pi_1} |\alpha_0^1\rangle \xrightarrow{\Pi_2} \cdots \xrightarrow{\Pi_d} |\alpha_0^d\rangle. \qquad [15]$$

The total error accumulated (see details in *SI Text*) $\epsilon = d\epsilon_0 < O(\beta^2\|H\|^2/d)$ can be suppressed by refining the step size, which is analogous to the quantum Zeno effect applied to a slowly varying basis.

Now, using the machinery we have developed, the operator* W_{j+1} can be used to construct such a projective measurement, through the phase estimation algorithm (32) (see also *SI Text*). Alternatively, one may employ the method in ref. 31, which is equivalent to some projective measurements, where W_{j+1} is applied multiple times randomly. In any case, the number of controlled W_{j+1} is at most $O(1/\sqrt{\delta})$ for each step of annealing, which is a quadratic speedup relative to classical Markov chains.

Comparison with Other Related Quantum Algorithms

To summarize, we have described a quantum Metropolis algorithm that extends Szegedy's method of classical Markov-chain quantization to the quantum domain and provides a quadratic quantum speedup $O(1/\sqrt{\delta})$ in the gap δ of the transition matrix M. This extension is achieved by adopting a dual representation where the set of basis states consists of pairs of eigenstates $|\varphi_i\rangle|\tilde{\varphi}_i\rangle$ related by the time-reversal operation. The algorithm requires twice the amount of system qubits, and the number of ancilla qubits has a logarithmic dependence on temperature $O[\log(1/T)]$. Details for comparison are summarized in the *SI Text*.

A comparison of various Markov-chain-based quantum algorithms is summarized in Table 1. To facilitate reading, we explain several definitions and terminologies below: (*i*) the term "classical Hamiltonian" (e.g., Ising spin model) refers to the cases where the eigenstates are the same as the basis states in the computational basis; the term "Quantum Hamiltonians" refer to the cases where the eigenstates generally contain superposition states of the states in the computational basis. Therefore, the eigenstates of the classical Hamiltonians are by definition known, whereas those for quantum Hamiltonians are not. (*ii*) The symbol ρ refers to a density matrix, $|+\rangle \equiv (|0\rangle + |1\rangle)/\sqrt{2}$, and the quantum state $|\phi_0\rangle$ as defined by Eq. **8**). (*iii*) The thermal density matrix is $\rho_{\text{th}} = e^{-\beta H}/\text{Tr}(e^{-\beta H})$, CETS II is defined in Eq. **1**), and

*The W operator defined in [12] for $|\alpha_k^{j+1}\rangle$.

CETS I (see also ref. 23) is similar to CETS II, but the basis states $\{|i\rangle\}$ are replaced by the computational basis. (*iv*) By "quantum speedup", we consider only the quantum speedup with respect to the gap δ of the transition matrix of the Markov chain. It is not known whether other types of quantum speedup exist.

Conclusion

In conclusion, we presented a quantum algorithm that completes the generalization of the classical Metropolis method to the quantum domain. Similar to ref. 34, the advantages of this quantum algorithm over classical algorithms can be exponential because there is no need to explicitly solve for the eigenvalues and eigenvectors using classical algorithms for the quantum Hamiltonians being simulated. Finally, as the application of the Metropolis method to quantum Hamiltonians can be considered as a special case of quantum maps (operations), it may be possible that the results presented here could be generalized to allow quantum speedup for a much broader class of quantum maps.

ACKNOWLEDGMENTS. We thank S. Boixo, J. McClean, D. Poulin, R. D. Somma, K. Temme, J. D. Whitfield, and P. Wocjan for insightful discussions, and we are grateful to the following funding sources: Croucher Foundation (M.-H.Y); Defense Advanced Research Planning Agency under the Young Faculty Award N66001-09-1-2101-DOD35CAP, The Camille and Henry Dreyfus Foundation, and the Sloan Foundation (A.A.-G).

1. Brush SG (1967) History of the Lenz-Ising model. *Rev Mod Phys* 39:883–893.
2. Nishimori H (2001) *Statistical Physics of Spin Glasses and Information Processing: An Introduction* (Oxford Univ Press, Oxford).
3. De las Cuevas G, Dür W, Van den Nest M, Martin-Delgado MA (2011) Quantum algorithms for classical lattice models. *New J Phys* 13:093021.
4. Kohn W (1999) Nobel Lecture: Electronic structure of matter—wave functions and density functionals. *Rev Mod Phys* 71:1253–1266.
5. Thijssen JM (1999) *Computational Physics* (Cambridge Univ Press, Cambridge, UK).
6. Aldous DJ (1982) Some inequalities for reversible Markov chains. *J London Math Soc* s2-25:564–576.
7. Mezard M, Parisi G, Virasoro MA (1987) *Spin Glass Theory and Beyond* (World Scientific, Teaneck, NJ).
8. Troyer M, Wiese U-J (2005) Computational complexity and fundamental limitations to fermionic quantum Monte Carlo simulations. *Phys Rev Lett* 94:170201–170204.
9. Kalos M, Pederiva F (2000) Exact Monte Carlo method for continuum fermion systems. *Phys Rev Lett* 85:3547–3551.
10. Foulkes W, Mitas L, Needs R, Rajagopal G (2001) Quantum Monte Carlo simulations of solids. *Rev Mod Phys* 73:33–83.
11. Feynman RP (1982) Simulating physics with computers. *Int J Theor Phys* 21:467–488.
12. Buluta I, Nori F (2009) Quantum simulators. *Science* 326:108–111.
13. Kassal I, Whitfield JD, Perdomo-Ortiz A, Yung M-H, Aspuru-Guzik A (2011) Simulating chemistry using quantum computers. *Annu Rev Phys Chem* 62:185–207.
14. Greiner M, Mandel O, Esslinger T, Hänsch TW, Bloch I (2002) Quantum phase transition from a superfluid to a Mott insulator in a gas of ultracold atoms. *Nature* 415:39–44.
15. Lloyd S (1996) Universal quantum simulators. *Science* 273:1073–1078.
16. Lidar D, Wang H (1999) Calculating the thermal rate constant with exponential speedup on a quantum computer. *Phys Rev E Stat Nonlin Soft Matter Phys* 59:2429–2438.
17. Zalka C (1998) Simulating quantum systems on a quantum computer. *Proc R Soc Lond A* 454:313–322.
18. Kassal I, Jordan SP, Love PJ, Mohseni M, Aspuru-Guzik A (2008) Polynomial-time quantum algorithm for the simulation of chemical dynamics. *Proc Natl Acad Sci USA* 105:18681–18686.
19. Ortiz G, Gubernatis J, Knill E, Laflamme R (2001) Quantum algorithms for fermionic simulations. *Phys Rev A* 64:022319–022322.
20. Poulin D, Wocjan P (2009) Preparing Ground States of Quantum Many-Body Systems on a Quantum Computer. *Phys Rev Lett* 102:130503–130506.
21. Poulin D, Wocjan P (2009) Sampling from the Thermal Quantum Gibbs State and Evaluating Partition Functions with a Quantum Computer. *Phys Rev Lett* 103:220502–220505.
22. Bilgin E, Boixo S (2010) Preparing Thermal States of Quantum Systems by Dimension Reduction. *Phys Rev Lett* 105:170405–170408.
23. Yung M-H, Nagaj D, Whitfield J, Aspuru-Guzik A (2010) Simulation of classical thermal states on a quantum computer: A transfer-matrix approach. *Phys Rev A* 82:060302(R)–060305.
24. Abrams D, Lloyd S (1999) Quantum algorithm providing exponential speed increase for finding eigenvalues and eigenvectors. *Phys Rev Lett* 83:5162–5165.
25. Aspuru-Guzik A, Dutoi AD, Love PJ, Head-Gordon M (2005) Simulated quantum computation of molecular energies. *Science* 309:1704–1707.
26. Wang HF, Kais S, Aspuru-Guzik A, Hoffmann MR (2008) Quantum algorithm for obtaining the energy spectrum of molecular systems. *Phys Chem Chem Phys* 10:5388–5393.
27. Veis L, Pittner J (2010) Quantum computing applied to calculations of molecular energies: CH2 benchmark. *J Chem Phys* 133:194106–194115.
28. Whitfield J, Biamonte J, Aspuru-Guzik A (2011) Simulation of electronic structure Hamiltonians using quantum computers. *Mol Phys* 109:735–750.
29. Li Z, et al. (2011) Solving Quantum ground-state problems with nuclear magnetic resonance. *Sci Rep* 1:88.
30. Szegedy M (2004) Quantum speed-up of Markov chain based algorithms. *45th Annual IEEE Symposium on Foundations of Computer Science* pp 32–41.
31. Somma R, Boixo S, Barnum H, Knill E (2008) Quantum simulations of classical annealing processes. *Phys Rev Lett* 101:130504–130507.
32. Wocjan P, Abeyesinghe A (2008) Speedup via quantum sampling. *Phys Rev A* 78:042336–042343.
33. Terhal B, DiVincenzo D (2000) Problem of equilibration and the computation of correlation functions on a quantum computer. *Phys Rev A* 61:22301–22322.
34. Temme K, Osborne TJ, Vollbrecht KG, Poulin D, Verstraete F (2011) Quantum Metropolis sampling. *Nature* 471:87–90.
35. Wootters WK, Zurek WH (1982) A single quantum cannot be cloned. *Nature* 299:802–803.

Supporting Information

Yung and Aspuru-Guzik 10.1073/pnas.1111758109

SI Text

Appendix: Construction of the Unitary Operators U_X and U_Y. Here, we show how one may construct the unitary operator U_X defined in Eq. **9**, namely

$$U_x|i\rangle = \sum_k (\sigma_{ik}|\varphi_i\rangle|\varphi_k\rangle|0\rangle + \gamma_{ik}|\varphi_i\rangle|\rangle\varphi_k\rangle|1\rangle), \qquad [S1]$$

where

$$\sigma_{ik} \equiv \alpha_{k\tilde{i}}\sqrt{z_{ik}}, \qquad \gamma_{ik} \equiv \alpha_{k\tilde{i}}\sqrt{1-z_{ik}}, \qquad [S2]$$

and

$$\alpha_{k\tilde{i}} \equiv \langle\varphi_k|K|\tilde{\varphi}_i\rangle. \qquad [S3]$$

Here, K is an unitary operator which plays the same role as the spin-flip in the classical Metropolis method, and z_{ik} is the Metropolis filter defined in Eq. **3**. Note that U_Y is related to U_X by a controlled SWAP. To ensure $\langle j|U_X^\dagger U_Y|i\rangle$ is positive, we assume that K is symmetrical in the computational basis:

$$\langle x'|K|x\rangle = \langle x|K|x'\rangle. \qquad [S4]$$

For example, K can be the SWAP operation. We start with the following n-quantum bit (qubit) state

$$\left(\frac{1}{\sqrt{2}}\right)^n (|0\rangle + |1\rangle)^{\otimes n}, \qquad [S5]$$

which is equivalent to the "all-input" state

$$\frac{1}{\sqrt{N}}\sum_{x=o}^{N-1}|x\rangle, \qquad [S6]$$

where $N \equiv 2^n$. Suppose now we include a set of n ancilla qubits initialized in the "all-down" state

$$|000\ldots0\rangle, \qquad [S7]$$

and apply a bit-by-bit controlled-NOT operation, which is equivalent to a copy of the value of x to the register qubits (which is not the same as quantum state cloning),

$$|x\rangle|000\ldots0\rangle \to |x\rangle|x\rangle. \qquad [S8]$$

The resulting state is

$$\frac{1}{\sqrt{N}}\sum_{x=0}^{N-1}|x\rangle|x\rangle. \qquad [S9]$$

Given any Hamiltonian H, if we formally insert the completeness relation

$$I = \sum_{i=0}^{N-1}|\varphi_i\rangle\langle\varphi_i|, \qquad [S10]$$

expanded in the eigenvector basis $\{|\varphi_i\}$, to the state in Eq. **9**, we then get the state in Eq. **8**,

$$|\alpha^0\rangle \equiv \frac{1}{\sqrt{N}}\sum_{i=0}^{N-1}|\varphi_i\rangle|\tilde{\varphi}_i\rangle, \qquad [S11]$$

without solving the eigenvalue equation, where

$$|\tilde{\varphi}_i\rangle \equiv \sum_{x=0}^{N-1}\langle\varphi_i|x\rangle|x\rangle \qquad [S12]$$

is the time-reversal counterpart of $|\varphi_i\rangle$. This state can be considered as the infinite-temperature state, and is also the starting point for the quantum simulated annealing (QSA).

We are now ready to consider the explicit procedure for constructing U_X defined Eq. **9**. Starting with the paired state

$$|\varphi_i\rangle|\tilde{\varphi}_i\rangle, \qquad [S13]$$

we apply the "kick" operator K to $|\tilde{\varphi}_i\rangle$, and write

$$K|\tilde{\varphi}_i\rangle = \sum_k \alpha_{k\tilde{i}}|\varphi_k\rangle. \qquad [S14]$$

Next, we implement the Metropolis filter by performing a controlled-rotation (based on the difference of the eigenvalues):

$$|\varphi_i\rangle|\varphi_k\rangle|0\rangle \to |\varphi_i\rangle|\varphi_k\rangle(\sqrt{z_{ik}}|0\rangle + \sqrt{1-z_{ik}}|1\rangle), \qquad [S15]$$

where

$$z_{ik} \equiv \{1, e^{-\beta(E_k - E_i)}\}. \qquad [S16]$$

This operation creates a state as described by Eq. **9**. Note that only the information about the difference of the eigenvalues is needed. There is no need to determine each eigenvalue individually.

Appendix: Generalization to Include Random "Kicks." In applying the classical Metropolis method, for example to Ising model, one usually applies random spin-flips to the spins. This feature can be incorporated in our quantum–quantum Metropolis algorithm (Q^2MA). On the other hand, implementation of these spin-flips is also necessary for systems with time-reversal symmetry where the eigenstates $|\tilde{\varphi}_i\rangle = |\varphi_i\rangle$ contain only real coefficients, in the computational basis.

To be specific, we consider a system of n spin-1/2 particles, and we include an extra ancilla qubits initialized as $|000\ldots0\rangle$. Then, in the first step of U_X, we perform a transformation such that

$$|000\ldots0\rangle \to \frac{1}{\sqrt{n}}\sum_{\lambda=1}^{n}|\lambda\rangle. \qquad [S17]$$

Then, conditioned on each value of λ, we apply a kick operator K_λ, where $K_\lambda^\dagger = K_\lambda$ (e.g., spin-flip operator), to the spin-λ. The net effect is that Eq. **9** becomes

$$\langle j_0|U_X^\dagger U_Y|i_0\rangle = \frac{1}{n}\sum_\lambda |\langle\varphi_j|K_\lambda|\tilde{\varphi}_i\rangle|^2\sqrt{z_{ij}z_{ji}}, \qquad [S18]$$

and Eq. **10** changes in a similar way.

Appendix: Stationary Distribution of the Szegedy Operator. Recall that in Eq. **13**, we block-diagonalize the Szegedy operator W by two results, namely,

$$\Lambda_1 U_X^\dagger U_Y |\alpha_k\rangle = \lambda_k |\alpha_k\rangle \quad \text{and} \quad \Lambda_2 |\alpha_k\rangle = \lambda_k U_X^\dagger U_Y |\alpha_k\rangle. \quad \textbf{[S19]}$$

The first equation is in fact the eigenvalue equation which defines $|\alpha_k\rangle$. The second equation requires the relation

$$\Lambda_1 U_X^\dagger U_Y \Lambda_1 = \Lambda_1 U_Y^\dagger U_X \Lambda_1, \quad \textbf{[S20]}$$

which comes from the first equation: Because the eigenvalues λ_k are real (and nonnegative), taking the inner product with $\langle\phi_k|$, we have

$$\lambda_k = \lambda_k^\dagger = \langle\alpha_k|\Lambda_1 U_X^\dagger U_Y \Lambda_1 |\alpha_k\rangle = \langle\alpha_k|\Lambda_1 U_Y^\dagger U_X \Lambda_1 |\alpha_k\rangle. \quad \textbf{[S21]}$$

Because this relationship is true for all k values, we write

$$\sum_k \langle\alpha_k|\Lambda_1 U_X^\dagger U_Y \Lambda_1 |\alpha_k\rangle - \langle\alpha_k|\Lambda_1 U_Y^\dagger U_X \Lambda_1 |\alpha_k\rangle = 0. \quad \textbf{[S22]}$$

Now, recall that the eigenvectors form a complete basis; i.e.,

$$\sum_k |\alpha_k\rangle\langle\alpha_k| = \sum_i |i\rangle\langle i| = \Lambda_1. \quad \textbf{[S23]}$$

We therefore have the equality $\Lambda_1 U_X^\dagger U_Y \Lambda_1 = \Lambda_1 U_Y^\dagger U_X \Lambda_1$.

The second point we need to address is to show that, in the $k = 0$ case ($\lambda_0 = 1$), $|\alpha_0\rangle$ contains the stationary distribution π_i of the Markov chain; i.e.,

$$|\alpha_0\rangle = \sum_i \sqrt{\pi_i}\,|i\rangle. \quad \textbf{[S24]}$$

We showed in Eq. **11** that

$$\Lambda_1 U_X^\dagger U_Y \Lambda_1 = D_\pi^{1/2} M D_\pi^{-1/2}, \quad \textbf{[S25]}$$

where $D_\pi \equiv \sum_{j=0}^{N-1} \pi_j |j\rangle\langle j|$ is a diagonal matrix, and $M \equiv \sum_{i,j} m_{ij} |j\rangle\langle i|$, with $m_{ii} \equiv \langle i|U_X^\dagger U_Y|i\rangle$ and $m_{ij} \equiv |\alpha_{\tilde{ji}}|^2 z_{ij}$ for $j \neq i$, is the transition matrix of the Markov chain. Note that we can also write

$$|\alpha_0\rangle = D_\pi^{1/2} \sum_i |i\rangle. \quad \textbf{[S26]}$$

Using the requirement of the conservation of probability sum of M, i.e.,

$$\sum_j m_{ij} = 1, \quad \textbf{[S27]}$$

which implies that

$$M \sum_i |i\rangle = \sum_i |i\rangle, \quad \textbf{[S28]}$$

we can indeed verify the eigenvalue equation

$$\Lambda_1 U_X^\dagger U_Y \Lambda_1 |\alpha_0\rangle = |\alpha_0\rangle \quad \textbf{[S29]}$$

is indeed true.

Appendix: Effects of Degeneracy and the Limitations of the Algorithm. The Q^{2}MA presented in the main text does not necessarily break down when the Hamiltonian H of the quantum system is highly degenerate in the space of the eigenstates. For the sake of the argument, consider the Ising model:

$$H_{\text{Ising}} = J \sum_{\langle ij \rangle} \sigma_i^z \sigma_j^z. \quad \textbf{[S30]}$$

Although the eigenstates are highly degenerate, the Q^{2}MA does not require the knowledge of the individual eigenvalues, but instead, it needs the difference between two eigenstates before and after the kick (spin-flip). In the case of the Ising model, the change in energy is $O(J)$, thus it is sufficient to ensure our resolution in resolving the energy change be smaller than $O(J)$.

To understand this point better, note that there are two places degeneracy would affect the agrument: (*i*) the implementation of the projector

$$\Lambda_1 \equiv \sum_i |i\rangle\langle i| \quad \textbf{[S31]}$$

defined in Szegedy's operator W; and (*ii*) the "leakage" to the degenerate subspace in Eq. **9**. These two points are related. The projector Λ_1 can be implemented via a filter method (e.g., see ref. 1); this is essentially the same as applying the phase estimation algorithm (PEA) multiple times. Explicitly, we consider two states $|\varphi_i\rangle|\varphi_j\rangle$, and we wish to determine whether $E_i = E_j$. We have to perform PEA with respect to the operator

$$U = e^{-iHt} \otimes e^{+iHt}, \quad \textbf{[S32]}$$

which produces only the differences of two eigenvalues.

Suppose we normalize all our eigenenergy such that $0 < E_k < 1$. We determine an energy window $\Delta = 2^{-a}$, for some integer a, the error ε for a single run of the PEA is bounded by

$$\varepsilon < \frac{\Delta^2}{|E_i - E_j|^2}. \quad \textbf{[S33]}$$

If we apply PEA k times with the same precision, then

$$\varepsilon \to \varepsilon^k, \quad \textbf{[S34]}$$

which means that the errors for those energy changes $|E_i - E_j|$ being greater than Δ, $|E_i - E_j| > \Delta$, would become exponentially small. The remaining problem is to deal with those energies change smaller than Δ; this issue is related to point (*ii*) above. Let us call this new projector Λ_1' which has resolution up to Δ.

Consider the transformation described in Eq. **9**. Let us call the contribution from those σ_{ik}, where $|E_i - E_k| < \Delta$, a leakage. If those (nearly) degenerate eigenstates do not contribute to the resulting state, then the operator $\Lambda_1' U_X^\dagger U_Y \Lambda_1'$ would not transform any basis state $|i\rangle$ outside of the paired basis. In other words, ideally we want, for those cases where $|E_i - E_k| < \Delta$,

$$\langle\varphi_k|K|\tilde{\varphi}_i\rangle = 0 \quad \text{(ideal)}. \quad \textbf{[S35]}$$

Practically, this condition cannot be satisfied because, typically, $K|\tilde{\varphi}_i\rangle$ should have a board or almost continuous spectrum. However, we can argue that the sum of these contributions,

$$\eta \equiv \sum_k{}' |\langle\varphi_k|K|\tilde{\varphi}_i\rangle|^2 \ll 1, \quad \textbf{[S36]}$$

could be made negligibly small for generic quantum systems where the Hamiltonian involves local interaction terms. Here, the summation is over those k where $|E_i - E_k| < \Delta$.

The reason is as follows: let us focus again on Eq. **9**. Before applying K, note that both $|\varphi_i\rangle$ and $|\tilde{\varphi}_i\rangle$ have the eigenenergy E_i. We may expect, in the worst case scenarios, $|\tilde{\varphi}_i\rangle$ would contain a significant weight (i.e., sharply peaked) in those energy eigenstates close to $|\varphi_i\rangle$. In this case, the leakage in the degenerate subspace would be very bad. However, when we apply a kick K, which does not preserve the symmetry of $\tilde{H}$, i.e.,

$$[K,\tilde{H}] \neq 0, \quad \textbf{[S37]}$$

then the change in energy

$$\Omega \equiv |\langle\tilde{\varphi}_i|K\tilde{H}K|\tilde{\varphi}_i\rangle - \langle\tilde{\varphi}_i|\tilde{H}|\tilde{\varphi}_i\rangle| \quad \textbf{[S38]}$$

is typically of order $O(J)$, where J is the typical size of the local terms. If we set $\Delta \ll J$, then we expect that

$$\eta \ll 1, \quad \textbf{[S39]}$$

and the correction to Eq. **13** is $O(\eta)$. The random kicks described in the previous section help spread out this effect (making the distribution more uniform).

In short, as long as the distribution

$$|\alpha_{k\tilde{i}}|^2 \equiv |\langle\varphi_k|K|\tilde{\varphi}_i\rangle|^2 \quad \textbf{[S40]}$$

is smoothly distributed over a range of energy that is much greater than the window Δ of the energy filter for Λ_1', we should expect that the contribution coming from the leakage can be made arbitrarily small, by decreasing Δ. A more quantitative analysis of this point is beyond the scope of this work.

Appendix: Error Analysis on the Process of Quantum Simulated Annealing. Here, we perform an error estimation for the process of QSA. First, the procedure of quantum simulated annealing starts with the infinite-temperature ($\beta = 0$) state (Eq. **8**) and ends up at some finite-temperature ($\beta \neq 0$) state (Eq. **1**). The inverse temperature β is divided into uniform intervals

$$\beta_j \equiv (j/d)\beta, \quad \textbf{[S41]}$$

where $j = 0,1,2,\ldots,d$, of $d+1$. The coherent thermal states $|\alpha_0^j\rangle$ corresponding to the intermediate temperatures β_j are prepared sequentially. This preparation is made possible by the projective measurement which projects $|\alpha_0^j\rangle$ to $|\alpha_0^{j+1}\rangle$. The final fidelity depends crucially on the overlap $|\langle\alpha_0^{j+1}|\alpha_0^j\rangle|^2$ between these states.

To estimate the overlap, note that

$$|\alpha_0^{j+1}\rangle = \frac{1}{\sqrt{\langle\alpha_0^j|e^{-\Delta\beta H}|\alpha_0^j\rangle}} e^{-\Delta\beta H/2}|\alpha_0^j\rangle, \quad \textbf{[S42]}$$

where $\Delta\beta \equiv \beta/d$. The overlap,

$$|\langle\alpha_0^j|\alpha_0^{j+1}\rangle|^2 = \frac{|\langle\alpha_0^j|e^{-\Delta\beta H/2}|\alpha_0^j\rangle|^2}{\langle\alpha_0^j|e^{-\Delta\beta H}|\alpha_0^j\rangle}, \quad \textbf{[S43]}$$

is second-order in $\Delta\beta$—i.e.,

$$|\langle\alpha_0^j|\alpha_0^{j+1}\rangle|^2 \approx 1 - O(\Delta\beta^2\langle H^2\rangle). \quad \textbf{[S44]}$$

This result is analogous to the quantum Zeno effect. In general, the energy fluctuation $\langle H^2\rangle$ is smaller for thermal states of lower temperatures. Therefore, a potential improvement could be made by nonlinear division of the β_j. Here, we assume it bounded above

$$\langle H^2\rangle \leq \langle H^2\rangle_0. \quad \textbf{[S45]}$$

For the whole process, the total error accumulates at each step to

$$\epsilon = d \times O(\Delta\beta^2\langle H^2\rangle_0) = O(\beta^2\langle H^2\rangle_0/d), \quad \textbf{[S46]}$$

as $\Delta\beta = \beta/d$. Hence, the total error can be made arbitrarily small by increasing d. In other words, to achieve any given accuracy ϵ, one must perform at least

$$d = O(\beta^2\langle H^2\rangle_0/\epsilon) \quad \textbf{[S47]}$$

steps in the process of quantum simulated annealing.

Next, for each step, the projective measurement

$$\Pi_{k+1} = \sum_{k=0}^{N-1} |\alpha_k^{j+1}\rangle\langle\alpha_k^{j+1}| \quad \textbf{[S48]}$$

can be achieved by the phase estimation algorithm. For this purpose, an improved version is described in ref. 1 (lemma 2). Here, we summarize the result: To achieve an accuracy of ϵ_0, the number of application of the controlled-W gates is

$$O[\log(1/\epsilon_0)/\Delta_{\min}], \quad \textbf{[S49]}$$

where $\Delta_{\min}$ is the minimum eigenvalue gap of W, see Eq. **14**. Here, we should put $\epsilon_0 = \epsilon/d$, which equals $\epsilon^2/\beta^2\langle H^2\rangle_0$ from Eq. **47**. Recall that $1/\Delta_{\min}$ is of order $O(1/\sqrt{\delta})$ from Eq. **14**, so we conclude that the number of controlled-W gate required is

$$O\left[\frac{\beta^2\langle H^2\rangle_0}{\sqrt{\delta}\epsilon}\log\left(\frac{\beta^2\langle H^2\rangle_0}{\epsilon^2}\right)\right]. \quad \textbf{[S50]}$$

A quadratic speedup of $1/\sqrt{\delta}$ is achieved.

Appendix: Spatial Resource Estimation. ***For Constructing U_X and U_Y.*** Now, let us consider the spatial resources required (in terms of the number of qubits) to construct U_X. The resource for constructing U_Y should be roughly the same. Suppose the Hamiltonian H to be simulated requires n qubits, which provides a Hilbert space size as large as $2^n \times 2^n$ for H. Then we need $2n+1$ qubits to define (see Eq. **9**) U_X. For the implementation, we will need more qubits; this comes from the step described in Eq. **15**. One way to implement it is to (*i*) apply phase estimation, which records the difference of the energies

$$\Delta_{ki} \equiv E_k - E_i \quad \textbf{[S51]}$$

into a register of r ancilla qubits, then (*ii*), conditioned on the value of the ancilla qubit, perform the rotation

$$|\Delta_{ki}\rangle|0\rangle \rightarrow |\Delta_{ki}\rangle(\sqrt{z_{ik}}|0\rangle + \sqrt{1-z_{ik}}|1\rangle). \quad \textbf{[S52]}$$

Then, we undo the phase estimation. Step (*ii*) requires some extra scratchpad qubits, but the number does not scale with the problem size. So we focus on step (*i*).

Suppose the Hamiltonian H is scaled such that $0 < \Delta_{ki} < 1$, and we approximate Δ_{ki} to an accuracy of 2^{-m}. It is well known that we need as many as

$$r = m + \lceil\log\left(2 + \frac{1}{2\varepsilon}\right)\rceil \quad \textbf{[S53]}$$

ancilla qubits to perform the phase estimation algorithm with an error ε. Here, m can be estimated in the following way: If we truncated the phase estimation algorithm to accuracy of $2^{-m}\|H\|$, then the Hamiltonian simulated will be different from the real one by $O(2^{-m}\|H\|)$. Denote this difference by ΔH. Then, following the same analysis in Eq. **42**, the error is

$$O(2^{-2m}\beta^2\|H\|^2), \qquad \textbf{[S54]}$$

which should be taken to be smaller then ε, which means that

$$m > \frac{1}{2}\log\left(\frac{\beta^2\|H\|^2}{2\varepsilon}\right). \qquad \textbf{[S55]}$$

In summary, apart from a scratchpad register of ancilla qubits, to implement U_X and U_Y for a Hamiltonian that is simulable by n qubits, we need

$$2n + 1 + r \qquad \textbf{[S56]}$$

qubits. Note that the inverse of $U_X^\dagger$ and $U_Y^\dagger$ requires the same amount of qubits.

For Constructing the Szegedy Operator W. The Szegedy operator W (see Eq. **12**) can be written as

$$W = U_X^\dagger U_Y(2\Lambda_1 - \mathcal{I})U_Y^\dagger U_X(2\Lambda_1 - \mathcal{I}), \qquad \textbf{[S57]}$$

which means that we have to consider the implementation of the projector $(2\Lambda_1 - \mathcal{I})$, which is described around Eq. **31**. As argued above, there is a leakage problem due to degeneracy of the eigenvalue spectrum of H, but for generic cases, the errors should be negligible. So we ignore them here. The implementation of Λ_1 depends on the phase estimation algorithm. So it requires, minimally, the same amount of qubits as U_X and U_Y. The result of applying Λ_1 to any state $|\psi\rangle$ would be of the form

$$\Lambda_1|\psi\rangle|0\rangle + (1 - \Lambda_1)|\psi\rangle|1\rangle. \qquad \textbf{[S58]}$$

To implement $(2\Lambda_1 - \mathcal{I})$, we can simply apply σ_z to the ancilla qubit and uncompute the phase estimation. Therefore, the total number of qubits required for W is essentially the same as that of U_X and U_Y.

For Performing Quantum Simulated Annealing. Once the Szegedy operator W is constructed, we can then apply the method of QSA to prepare thermal states at any temperature, in a way similar to that in refs. 1 and 2. In particular, lemme 2 of ref. 1 suggests that, to perform the projection operation in QSA for a Markov-chain with gap Δ, we need as many as

$$\log(1/\sqrt{\varepsilon}\Delta) \qquad \textbf{[S59]}$$

ancilla qubits. Of course, the ancilla qubits employed in constructing U_X and U_Y can be recycled here. We therefore see that the requirement of the ancilla qubits depends on the Markov-chain gap Δ of the problem. In general, the value of the gap cannot be predetermined. A practical method for the estimation is by including more and more ancilla qubits in the phase estimation algorithm until the value converges. This procedure produces constant overhead for the whole algorithm.

1. Wocjan P, Abeyesinghe A (2008) Speedup via quantum sampling. *Phys Rev A* 78:042336–042343.

2. Somma R, Boixo S, Barnum H, Knill E (2008) Quantum simulations of classical annealing processes. *Phys Rev Lett* 101:130504–130507.