DIGITAL ACCESS TO SCHOLARSHIP AT HARVARD

A Quantum-Quantum Metropolis Algorithm

(Article begins on next page)

A Quantum-Quantum Metropolis Algorithm

Man-Hong Yung[1, *] and Alán Aspuru-Guzik[1, †]

[1]*Department of Chemistry and Chemical Biology, Harvard University, Cambridge MA, USA*
(Dated: November 8, 2010)

Recently, the idea of classical Metropolis sampling through Markov chains has been generalized for quantum Hamiltonians. However, the underlying Markov chain of this algorithm is still classical in nature. Due to Szegedy's method, the Markov chains of classical Hamiltonians can achieve a quadratic quantum speedup in the eigenvalue gap of the corresponding transition matrix. A natural question to ask is whether Szegedy's quantum speedup is merely a consequence of employing classical Hamiltonians, where the eigenstates simply coincide with the computational basis, making cloning of the classical information possible. We solve this problem by introducing a quantum version of the method of Markov-chain quantization combined with the quantum simulated annealing (QSA) procedure, and describe explicitly a novel quantum Metropolis algorithm, which exhibits a quadratic quantum speedup in the eigenvalue gap of the corresponding Metropolis Markov chain for any quantum Hamiltonian. This result provides a complete generalization of the classical Metropolis method to the quantum domain.

Interacting many-body problems, classical or quantum mechanical, generally require an exponentially large amount of computing resources to find the (exact) solutions, as the system size increases. Nonetheless, ingenious classical methods such as Markov-chain Monte Carlo (MCMC) or quantum Monte Carlo (QMC) have been invented and proven to be highly successful in many applications. These methods, however, have certain limitations. For example, the running time of MCMC scales as $O(1/\delta)$ [1], where δ is the gap of the transition matrix. For problems such as spin glasses where δ is small, MCMC becomes computationally inefficient. QMC methods, on the other hand, suffer from the negative sign problem [2]. Despite many efforts have been made for improvement [3], this limitation is still one of the biggest challenges in QMC [4].

On the other hand, one of the most important goals in the field of quantum computation, as proposed by Feynman, is to look for new methods or algorithms that can solve these many-body problems more efficiently. This is referred to as *quantum simulation*, which can be implemented either by dedicated quantum simulators [5], or with universal quantum computers [6]. For the former case, high precision experimental techniques are required for faithful simulation, which are therefore closer to engineering problems. For the latter case, sophisticated quantum algorithms are needed. To this end, there are two main approaches, (a) bottom-up approaches [7–12]: to look for entirely new algorithms based on the special properties of quantum computers, and (b) top-down approaches [13–19]: to improve the existing classical algorithms by combining with elementary quantum algorithms. This work belongs to the latter class.

Some years ago, Szegedy [14] described a method to quantize classical Markov chains. The key result is that a quadratic speedup $O(1/\sqrt{\delta})$ in the gap δ of the transition matrix is possible. This was later adapted to some new algorithms [16, 17] that can prepare thermal states of classical Hamiltonians, based on the idea of quantum simulated annealing (QSA).

In fact, Markov chains that correspond to classical Hamiltonians are relatively easy to construct, as all the eigenstates are in the computational basis $|x\rangle$. Preparation of the thermal states of quantum Hamiltonians [12, 13, 18], however, is a much more challenging problem. It is because, classically, one needs 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.

The key to overcome this difficulty with a quantum computer is the use of quantum phase estimation algorithm [8]; 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 [13] were able to extend the Metropolis algorithm to the quantum domain, avoiding the negative sign problem in QMC. However, one limitation for their results is that the Metropolis step involve too many energy non-local transitions, making it conceivably inefficient. Recently, Temme *et al.*, [18] addressed this problem by introducing random local unitary operations in the update rule. In both cases, however, the underlying Markov chain is still classical in nature, which means that the scaling of the running time is still $O(1/\delta)$.

In this letter, we propose a new quantum Metropolis algorithm for an arbitrary quantum Hamiltonian H at a given temperature T ($\beta \equiv 1/k_B T$), with a quadratic speedup $O(1/\sqrt{\delta})$. We call this algorithm the quantum-quantum Metropolis algorithm (Q^{2}MA), as it shows a quantum speedup for the Markov chain of a quantum system being simulated. The previous version of the quantum metropolis algorithm [18] is restricted by the no-

TABLE I: Comparison of various Markov-chain based algorithms for thermal state preparation

Methods	Hamiltonian[a]	Input[b]	Output[c]	Q. Speedup[d]
Q. Metropolis I[e]	Quantum	Any ρ	ρ_{th}	No
Q. Metropolis II[f]	Quantum	Any ρ	ρ_{th}	No
Q. Markov chain I[g]	Classical	$\|+\rangle^{\otimes n}$	CETS I	Quadratic
Q. Markov chain II[h]	Classical	$\|+\rangle^{\otimes n}$	CETS I	Quadratic
Q^{2}MA (This work)	Quantum	$\|\alpha^0\rangle$	CETS II	Quadratic

[a]Quantum Hamiltonians include classical Hamiltonians.
[b]Here, ρ is a density matrix, $|+\rangle \equiv (|0\rangle + |1\rangle)/\sqrt{2}$, and $|\alpha^0\rangle$ is defined in Eq. (6).
[c]Here ρ_{th} is the thermal density matrix (cf. Eq. (1)), CETS II is defined in Eq. (19), and CETS I (see also Ref. [19]) is similar to CETS II, but the $\{|i\rangle\}$ is replaced by the computational basis. Both of them are equivalent to ρ_{th}.
[d]We consider only the quantum speedup with respect to the gap δ of the transition matrix of the Markov chain.
[e]In Ref. [13], the Metropolis rule is implemented by a controlled-swap with a "duplicated environment".
[f]In Ref. [18], improvement of Ref. [13], a rejection rule for quantum states is imposed to implement the Metropolis rule.
[g]In Ref. [16], combining Szegedy's Markov-chain quantization with quantum simulated annealing (QSA).
[h]In Ref. [17], improvement of Ref. [16] using Grover's fixed-point search.

cloning theorem insofar that the information of a eigenstate cannot be retrieved after the proposed move in the Metropolis step. We relax this restriction by adopting a dual representation where the basis states consists of pairs of eigenstates related by time-reversal operation. For the cases where the physical system being simulated is time-reversal invariant, the pair of the eigenstates forming the basis vector become identical. A comparison of various Markov-chain based methods is summarized in Table I.

The goal of the Q^{2}MA is to prepare the coherent encoding of the thermal state (CETS) $|\alpha_0\rangle$ (cf. Eq. (19)), which, after tracing out the ancilla qubits, is equivalent to

$$\rho_{th} = \frac{1}{Z} \sum_i e^{-\beta E_i} |\varphi_i\rangle \langle \varphi_i| \quad , \tag{1}$$

where $Z = Tr\left[e^{-\beta H}\right]$ is the partition function, and $|\varphi_i\rangle$ is the eigenstate of a quantum Hamiltonian H, associated with the eigenvalue E_i. Before going into the technical details of this work, we briefly summarize the important features of Markov chains and Metropolis sampling, in a way applicable to both classical and quantum systems.

Brief review of the Metropolis method — In the standard Metropolis scheme, the Gibbs distribution $e^{-\beta E_i}/Z$ of certain eigenstates is generated through a Markov chain M, in which the matrix element m_{ij} refers to the transition probability from the eigenstate $|\varphi_i\rangle$ to $|\varphi_j\rangle$. The equilibrium (stationary) distribution

$$\pi_i \equiv e^{-\beta E_i}/Z \tag{2}$$

satisfies the detailed balance condition,

$$\pi_i m_{ij} = \pi_j m_{ji} \quad . \tag{3}$$

A solution which can satisfy the detailed balance condition is $m_{ji} = s_{ij}z_{ji}$, where $s_{ij} = s_{ji}$ is any symmetrical transition probability,

$$z_{ij} = \min\{1, e^{-\beta(E_j - E_i)}\} \tag{4}$$

is sometimes called the Metropolis filter. In the practical implementation of the Metropolis method, one starts with some initial configuration, and then apply a random transition (e.g. single spin-flip) and compare the energy between the new eigenstate and the old one. If the new eigenstate has a lower energy, accept the move. Otherwise, accept the move only with a probability distribution given by the ratio of the corresponding Boltzmann factors $e^{-\beta(E_j - E_i)}$. This is called the Metropolis method.

The performance of the Metropolis method depends on the properties of the underlying transition matrix M of the Markov chain, especially, the spectrum of the eigenvalues λ_k, which are all positive and bounded by 1, and the largest eigenvalue is always 1. For convenience, we order them as

$$\lambda_0 = 1 > \lambda_1 \geq ... \geq \lambda_{N-1} > 0 \quad . \tag{5}$$

The convergence time of a Markov chain is limited by the eigenvalue gap $\delta \equiv 1 - \lambda_1$ of the transition matrix M by $O(1/\delta)$ [1]. The purposes of this work are: (1) to show that the running time can be improved to $O(1/\sqrt{\delta})$, and (2) to fully extend the Metropolis sampling algorithm into the quantum domain.

Generalization of the Markov-chain quantization — The original Markov-chain quantization method by Szegedy [14] is applicable to classical Hamiltonians only. To extend it to quantum Hamiltonians, we start with n qubits prepared 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 CNOT gate on a set of n ancilla qubits initialized in the state $|000...0\rangle$, one can *formally* express the entire quantum state as

$$|\alpha^0\rangle \equiv \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} |\varphi_i\rangle |\tilde{\varphi}_i\rangle \quad , \tag{6}$$

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

$$|\tilde{\varphi}_i\rangle \equiv \sum_{x=0}^{N-1} \langle \varphi_i|x\rangle |x\rangle \tag{7}$$

is the *time-reversal* counterpart of $|\varphi_i\rangle$, which is the eigenstate of the corresponding time-reversal Hamiltonian $\tilde{H} \equiv H^*$ with the same eigen-energy E_i (i.e.,

$\tilde{H}\,|\tilde{\varphi}_i\rangle = E_i\,|\tilde{\varphi}_i\rangle$). 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 [18]. In the following, for the purpose of demonstration, we shall assume that the PEA can be applied perfectly, in the sense that each eigenstate can be uniquely identified by a unique eigenvalue. We leave our discussion on the effects of degeneracy on this algorithm in the Appendix. It turns out that the degeneracy of the eigen-energies of a quantum Hamiltonian alone is not necessarily a problem.

Now, let us include into our Hilbert space an extra qubit initialized in $|0\rangle$ and define a more compact notation,

$$|i\rangle \equiv |\varphi_i\rangle\,|\tilde{\varphi}_i\rangle\,|0\rangle \quad . \tag{8}$$

The information of a Markov chain can be encoded in a pair of unitary operators U_X and U_Y (see Appendix for their detailed construction),

$$U_X\,|i\rangle = \sum_k \left(\sigma_{ik}\,|\varphi_i\rangle\,|\varphi_k\rangle\,|0\rangle + \gamma_{ik}\,|\varphi_i\rangle\,|\varphi_k\rangle\,|1\rangle\right), \tag{9}$$

$$U_Y\,|j\rangle = \sum_m \left(\sigma_{jm}\,|\varphi_m\rangle\,|\varphi_j\rangle\,|0\rangle + \gamma_{jm}\,|\varphi_j\rangle\,|\varphi_m\rangle\,|1\rangle\right), \tag{10}$$

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 which plays the same role as the spin-flip in the classical Metropolis method, and z_{ik} is the Metropolis filter defined in Eq. (4). Note that U_Y is related to U_X by a controlled-SWAP operation.

Satisfying the detailed balance conditions — The detailed balance condition (Eq. (3)) can be recovered by the product of $U_X^\dagger$ and U_Y. From Eq. (9) and (10), for $j \neq i$, we have

$$\langle j|\,U_X^\dagger U_Y\,|i\rangle = |\alpha_{j\tilde{i}}|^2 (z_{ij}z_{ji})^{1/2} \quad , \tag{11}$$

where we used $\langle\varphi_i|\tilde{\varphi}_j\rangle = \langle\varphi_j|\tilde{\varphi}_i\rangle$. On the other hand,

$$\langle i|U_X^\dagger U_Y|i\rangle = |\alpha_{i\tilde{i}}|^2 + \sum_k |\alpha_{k\tilde{i}}|^2 (1-z_{ik}), \tag{12}$$

which, as we shall see, can be interpreted as the probability of not undergoing a transition.

Construction of the operator W — Now, using Eq. (3), (11) and (12), we obtain the following decomposition:

$$\langle j|\,U_X^\dagger U_Y\,|i\rangle = \langle j|\,D_\pi^{1/2}\,|j\rangle\,\langle j|\,M\,|i\rangle\,\langle i|\,D_\pi^{-1/2}\,|i\rangle \quad , \tag{13}$$

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 $m_{ij} \equiv |\alpha_{j\tilde{i}}|^2 z_{ij}$ for $j \neq i$ is the transition matrix of the Markov chain. Within the subspace $\{|i\rangle\}$, Eq. (13) implies that $U_X^\dagger U_Y$ and M are similar matrices, which means that

they have the same set of eigenvalues λ_k (see Eq. (5)). Following Szegedy [14], this property allows us to construct an operator

$$W \equiv (2\Lambda_2 - I)(2\Lambda_1 - I) \quad , \tag{14}$$

where

$$\Lambda_1 \equiv \sum_{i=0}^{N-1} |i\rangle\,\langle i| \quad \text{and} \quad \Lambda_2 \equiv U_X^\dagger U_Y \Lambda_1 U_Y^\dagger U_X. \tag{15}$$

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

$$\Lambda_1 U_X^\dagger U_Y\,|\alpha_k\rangle = \lambda_k\,|\alpha_k\rangle \quad . \tag{16}$$

On the other hand, using the fact that $\Lambda_1 U_X^\dagger U_Y \Lambda_1 = \Lambda_1 U_Y^\dagger U_X \Lambda_1$, we have,

$$\Lambda_2\,|\alpha_k\rangle = \lambda_k U_X^\dagger U_Y\,|\alpha_k\rangle \quad . \tag{17}$$

Eq. (16) and (17) 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,

$$w_k = \begin{bmatrix} \cos(2\theta_k) & -\sin(2\theta_k) \\ \sin(2\theta_k) & \cos(2\theta_k) \end{bmatrix} \quad , \tag{18}$$

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 $\theta_0 = 0$), $w_0 = I$ is simply an identity. From Eq. (13) and the properties of the Markov matrix (see Appendix), the $k = 0$ state is the coherent encoding of the thermal state (CETS) of Eq. (1),

$$|\alpha_0\rangle = \sum_{i=0}^{N-1} \sqrt{\pi_i}\,|i\rangle \quad . \tag{19}$$

Recall that $|i\rangle \equiv |\varphi_i\rangle\,|\tilde{\varphi}_i\rangle\,|0\rangle$, the state $|\alpha_0\rangle$ becomes the Gibbs thermal state $\rho_{th} = e^{-\beta H}/Tr[e^{-\beta H}]$ when the other qubits are traced out.

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

$$\Delta_{min} \geq 2\sqrt{\delta} \quad , \tag{20}$$

which is the origin of the quadratic speedup of Szegedy's algorithm. This completes our discussion on the necessary tools needed for the following discussion.

Quantum simulated annealing (QSA) — Given a quantum Hamiltonian H and any finite temperature T, our goal is to obtain the corresponding coherent thermal state

of the form in Eq. (19), from the initial state defined in Eq. (6), which can be readily prepared from the "all-zero" state $|000...0\rangle$, and be considered as the infinite-temperature state. To achieve this goal, we can use the method of quantum simulated annealing (QSA) [16]. For completeness, we outline the basic strategy and summarize the related results in the following paragraph.

The strategy of the QSA method is to prepare a sequence, $j = 0, 1, 2, .., d$, of $d+1$ of coherent thermal states,

$$|\alpha_0^j\rangle = \sum_{i=0}^{N-1} \left(e^{-\beta_j E_i}/Z_j\right)^{1/2} |i\rangle \quad , \qquad (21)$$

at a time; the temperature $\beta_j \equiv (j/d)\beta$ of the coherent thermal state is lowered in each step. The basic idea is that, for sufficiently small $\Delta\beta \equiv \beta/d$, one can show that $|\alpha_0^j\rangle$ has a good overlap with $|\alpha_0^{j+1}\rangle$, i.e.,

$$|\langle\alpha_0^{j+1}|\alpha_0^j\rangle|^2 \geq 1 - \epsilon_0 \quad , \qquad (22)$$

where ϵ_0 is bounded by $O(\Delta\beta^2 \|H\|^2)$, then a projective measurement

$$\Pi_{k+1} = \sum_{k=0}^{N-1} |\alpha_k^{j+1}\rangle\langle\alpha_k^{j+1}| \qquad (23)$$

on the eigenbasis of $|\alpha_k^{j+1}\rangle$ to $|\alpha_0^j\rangle$ will result in the lower temperature state $|\alpha_0^{j+1}\rangle$ with a high probability. For the whole process,

$$|\alpha_0^0\rangle \xrightarrow{\Pi_1} |\alpha_0^1\rangle \xrightarrow{\Pi_2} ... \xrightarrow{\Pi_d} |\alpha_0^d\rangle \quad , \qquad (24)$$

the total error (see Appendix)

$$\epsilon = d\epsilon_0 < O(\beta^2 \|H\|^2 /d) \qquad (25)$$

is suppressed by refining the step size, analogous to the quantum Zeno effect.

Now, using the machinery we have developed, the operator W_{j+1} (the W operator defined in Eq. (14) for $|\alpha_k^{j+1}\rangle$) can be used to construct such a projective measurement, through the phase estimation algorithm [17] (see also Appendix). Alternatively, one may perform an artificial way of introducing decoherence [16], where W_{j+1} is applied multiple times randomly. In any case, the number of controlled-W_{j+1} is at most $O(1/\sqrt{\delta})$, which is a quadratic speedup relative to classical Markov chains. This completes our description on the Q^{2}MA.

Conclusion — To summarize, we have described a new quantum Metropolis algorithm which 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. The restriction encountered by the previous version of the quantum Metropolis algorithm is mostly due to the no-cloning theorem, where the required information, such as the associated eigenvalue, of an eigenstate cannot be retrieved after the proposed move in the Metropolis step. We relax this restriction by adopting a dual representation where the set of basis states consists of pairs of eigenstates related by the time-reversal operation.

This result completes the generalization of the classical Metropolis method to the quantum domain. Morevoer, the advantages of this quantum algorithm over classical algorithms could be exponential, as there is no need to explicitly solve for the eigenvalues and eigenvectors in the classical ways 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.

———————

* Electronic address: mhyung@chemistry.harvard.edu
† Electronic address: aspuru@chemistry.harvard.edu
[1] D. Aldous, J. London Math. Soc. **25** (3), 564 (1982).
[2] M. Troyer and U.-J. Wiese, Phys. Rev. Lett. **94**, 170201 (2005).
[3] M. H. Kalos and Francesco Pederiva, Phys. Rev. Lett. **85**, 3547 (2000).
[4] W. M. C. Foulkes, L. Mitas, R. J. Needs and G. Rajagopal, Rev. Mod. Phys. **73**, 33 (2001).
[5] I. Buluta and F. Nori, Science **306**, 5949 (2009).
[6] I. Kassal, J. D. Whitfield, A. Perdomo-Ortiz, M.-H. Yung, A. Aspuru-Guzik, arXiv:1007.2648.
[7] C Zalka, Proc. R. Soc. Lond. A **454**, 313 (1998).
[8] D. Abrams and S. Lloyd, Phys. Rev. Lett. **83**, 5162 (1999).
[9] D. Lidar and H. Wang, Phys. Rev. E **59**, 2429 (1999).
[10] G. Ortiz, J. Gubernatis, E. Knill, and R. Laflamme, Phys. Rev. A **64**, 022319 (2001).
[11] A. Aspuru-Guzik, A. D. Dutoi, P. J. Love, and M. Head-Gordon, Science **309**, 1704 (2005).
[12] D. Poulin and P. Wocjan, Phys. Rev. Lett. **102**, 130503 (2009); D. Poulin and P. Wocjan, Phys. Rev. Lett. **103**, 220502 (2009).
[13] B. M. Terhal and D. P. DiVincenzo, Phys. Rev. A **61**, 022301 (2000).
[14] M. Szegedy, *Proceedings of the 45th IEEE Symposium on Foundations of Computer Science (2004)*, p. 32.
[15] P. C. Richter, Phys. Rev. A **76**, 042306 (2007).
[16] R. Somma, S. Boixo, H. Barnum, and E. Knill, Phys. Rev. Lett. **101**, 130504 (2008).
[17] P. Wocjan and A. Abeyesinghe, Phys. Rev. A **78**, 042336

(2008).

[18] K. Temme, T.J. Osborne, K.G. Vollbrecht, D. Poulin, and F. Verstraete, arXiv:0911.3635.

[19] Man-Hong Yung, Daniel Nagaj, James D. Whitfield, Alán Aspuru-Guzik, arXiv:1005.0020.

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

$$U_X \left| i \right\rangle = \sum_k \left(\sigma_{ik} \left| \varphi_i \right\rangle \left| \varphi_k \right\rangle \left| 0 \right\rangle + \gamma_{ik} \left| \varphi_i \right\rangle \left| \varphi_k \right\rangle \left| 1 \right\rangle \right) \quad , \quad (26)$$

where

$$\sigma_{ik} \equiv \alpha_{k\tilde{i}} \sqrt{z_{ik}} \quad , \quad \gamma_{ik} \equiv \alpha_{k\tilde{i}} \sqrt{1 - z_{ik}} \quad , \quad (27)$$

and

$$\alpha_{k\tilde{i}} \equiv \left\langle \varphi_k \right| K \left| \tilde{\varphi}_i \right\rangle \quad . \quad (28)$$

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. (4). Note that U_Y is related to U_X by a controlled-SWAP. To ensure Eq. (11) is positive, we assume that K is symmetrical in the computational basis:

$$\left\langle x' \right| K \left| x \right\rangle = \left\langle x \right| K \left| x' \right\rangle \quad . \quad (29)$$

For example, K can be the SWAP operation.

We start with the following n-qubit state

$$\left(\frac{1}{\sqrt{2}} \right)^n \left(\left| 0 \right\rangle + \left| 1 \right\rangle \right)^{\otimes n} \quad , \quad (30)$$

which is equivalent to the "all-input" state

$$\frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} \left| x \right\rangle \quad , \quad (31)$$

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

$$\left| 000...0 \right\rangle \quad , \quad (32)$$

an apply a bit-by-bit CNOT 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),

$$\left| x \right\rangle \left| 000...0 \right\rangle \rightarrow \left| x \right\rangle \left| x \right\rangle \quad . \quad (33)$$

The resulting state is

$$\frac{1}{\sqrt{N}} \sum_{x=0}^{N-1} \left| x \right\rangle \left| x \right\rangle \quad . \quad (34)$$

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

$$I = \sum_{i=0}^{N-1} \left| \varphi_i \right\rangle \left\langle \varphi_i \right| \quad , \quad (35)$$

expanded in the eigenvector basis $\{ \left| \varphi_i \right\rangle \}$, to the state in Eq. (34), then we get the state in Eq. (6),

$$\left| \alpha^0 \right\rangle \equiv \frac{1}{\sqrt{N}} \sum_{i=0}^{N-1} \left| \varphi_i \right\rangle \left| \tilde{\varphi}_i \right\rangle \quad , \quad (36)$$

without solving the eigenvalue equation, where

$$\left| \tilde{\varphi}_i \right\rangle \equiv \sum_{x=0}^{N-1} \left\langle \varphi_i | x \right\rangle \left| x \right\rangle \quad (37)$$

is the time-reversal counterpart of $\left| \varphi_i \right\rangle$. 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

$$\left| \varphi_i \right\rangle \left| \tilde{\varphi}_i \right\rangle \quad , \quad (38)$$

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

$$K \left| \tilde{\varphi}_i \right\rangle = \sum_k \alpha_{k\tilde{i}} \left| \varphi_k \right\rangle \quad . \quad (39)$$

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

$$\left| \varphi_i \right\rangle \left| \varphi_k \right\rangle \left| 0 \right\rangle \rightarrow \left| \varphi_i \right\rangle \left| \varphi_k \right\rangle \left(\sqrt{z_{ik}} \left| 0 \right\rangle + \sqrt{1 - z_{ik}} \left| 1 \right\rangle \right) , \quad (40)$$

where

$$z_{ij} \equiv \left\{ 1, e^{-\beta(E_j - E_i)} \right\} \quad . \quad (41)$$

This 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 usual apply random spin-flips to the spins. This feature can be incorporated in our Q^2MA. On the other hand, implementation of these spin-flips is also necessary for systems with time-reversal symmetry where the eigenstates $\left| \tilde{\varphi}_i \right\rangle = \left| \varphi_i \right\rangle$ contain only real coefficients, in the computational basis.

To be specific, we consider a system of n spin-1/2 particles, we include an extra ancilla qubits initialized as $|000...0\rangle$. Then, in the first step of U_X, we perform a transformation such that

$$|000...0\rangle \rightarrow \frac{1}{\sqrt{n}} \sum_{\lambda=1}^{n} |\lambda\rangle . \qquad (42)$$

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

$$\langle j_0| U_X^\dagger U_Y |i_0\rangle = \frac{1}{n} \sum_\lambda |\langle \varphi_j| K_\lambda |\tilde{\varphi}_i\rangle|^2 \sqrt{z_{ij} z_{ji}} , \quad (43)$$

and Eq. (10) changes in a similar way.

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:

$$H_{\text{Ising}} = J \sum_{\langle i,j \rangle} \sigma_i^z \sigma_j^z . \qquad (44)$$

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: (A) the implementation of the projector

$$\Lambda_1 \equiv \sum_i |i\rangle \langle i| \qquad (45)$$

in Eq. (15). (B) 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. [17]); this is essentially the same as applying the phase estimation algorithm (PEA) multiple times. Explicitly, we consider two states $|\varphi_i\rangle |\varphi_j\rangle$, we wish to determine whether $E_i = E_j$. We have to perform PEA w.r.t the operator

$$U = e^{-iHt} \otimes e^{+iHt} , \qquad (46)$$

which produces only the differences of two eigenvalues.

Suppose we normalize all our eigen-energy 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

$$\varepsilon < \frac{\Delta^2}{|E_i - E_j|^2} . \qquad (47)$$

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

$$\varepsilon \rightarrow \varepsilon^k , \qquad (48)$$

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 the point (B) above. Let us call this new projector Λ_1' which has resolution up to Δ.

Consider the transformation described in Eq. (9). Let us called the contribution from those σ_{ik} where $|E_i - E_k| < \Delta$ a "leakage". It is because 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 $|E_i - E_k| < \Delta$,

$$\langle \varphi_k| K |\tilde{\varphi}_i\rangle = 0 \quad \text{(ideal)}. \qquad (49)$$

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

$$\eta \equiv \sum_k{}' |\langle \varphi_k| K |\tilde{\varphi}_i\rangle|^2 \ll 1 , \qquad (50)$$

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 eigen-energy 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 eigen-states 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.,

$$[K, \tilde{H}] \neq 0 , \qquad (51)$$

then the change in energy

$$\Omega \equiv |\langle \tilde{\varphi}_i| K\tilde{H}K |\tilde{\varphi}_i\rangle - \langle \tilde{\varphi}_i| \tilde{H} |\tilde{\varphi}_i\rangle| \qquad (52)$$

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 $\eta \ll 1$, and the correction to Eq. (16) and (17) 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 $|\alpha_{k\tilde{i}}|^2 \equiv |\langle\varphi_k| K |\tilde{\varphi}_i\rangle|^2$ is smoothly distributed over a range of energy which 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. We leave a more quantitative analysis of this point for the future work.

Appendix: Error analysis on the process of quantum simulated annealing

Here we perform an error estimation for the process of quantum simulated annealing (QSA). First, the procedure of quantum simulated annealing starts with the infinite-temperature ($\beta = 0$) state Eq. (6), and end up at some finite-temperature ($\beta \neq 0$) state Eq. (19). The inverse temperature β is divided into uniform intervals

$$\beta_j \equiv (j/d)\,\beta \quad , \tag{53}$$

where $j = 0, 1, 2, .., d$, of $d + 1$. The coherent thermal states $|\alpha_0^j\rangle$ corresponding to the intermediate temperatures β_j are prepared sequentially. The is made possible by the projective measurement which projects $|\alpha_0^j\rangle$ to $|\alpha_0^{j+1}\rangle$. The finally 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

$$|\alpha_0^{j+1}\rangle = \frac{1}{\sqrt{\langle\alpha_0^j|e^{-\Delta\beta H}|\alpha_0^j\rangle}} e^{-\Delta\beta H/2}|\alpha_0^j\rangle \quad , \tag{54}$$

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

$$|\langle\alpha_0^j|\alpha_0^{j+1}\rangle|^2 = \frac{|\langle\alpha_0^j|e^{-\Delta\beta H/2}|\alpha_0^j\rangle|^2}{\langle\alpha_0^j|e^{-\Delta\beta H}|\alpha_0^j\rangle} \quad , \tag{55}$$

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

$$|\langle\alpha_0^j|\alpha_0^{j+1}\rangle|^2 \approx 1 - O(\Delta\beta^2\langle H^2\rangle) \quad . \tag{56}$$

This is analogous to the 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 non-linear division of the β_j. Here we assume it bounded above,

$$\langle H^2\rangle \leq \langle H^2\rangle_0 \quad . \tag{57}$$

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

$$\epsilon = d \times O(\Delta\beta^2\langle H^2\rangle_0) = O(\beta^2\langle H^2\rangle_0/d) \quad , \tag{58}$$

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

$$d = O(\beta^2\langle H^2\rangle_0/\epsilon) \tag{59}$$

steps in the process of quantum simulated annealing.

Next, for each step, the projective measurement

$$\Pi_{k+1} = \sum_{k=0}^{N-1} |\alpha_k^{j+1}\rangle\langle\alpha_k^{j+1}| \tag{60}$$

can be achieved by the phase estimation algorithm. For this purpose, an improved version is described in Ref. [17] (Lemma 2). Here we summarize the result: to achieve an accuracy of ϵ_0, the number of application of the controlled-W gates is

$$O\left(\log\left(1/\epsilon_0\right)/\Delta_{\min}\right) \quad , \tag{61}$$

where $\Delta_{\min}$ is the minimun eigenvalue gap of W, see Eq. (20). Here we should put $\epsilon_0 = \epsilon/d$, which equals $\epsilon^2/\beta^2\langle H^2\rangle_0$ from Eq. (59). Recall that $1/\Delta_{\min}$ is of order $O(1/\sqrt{\delta})$ from Eq. (20), we conclude that the number of controlled-W gate required is

$$O\left(\frac{\beta^2\langle H^2\rangle_0}{\sqrt{\delta}\epsilon}\log\left(\frac{\beta^2\langle H^2\rangle_0}{\epsilon^2}\right)\right) \quad . \tag{62}$$

A quadratic speedup of $1/\sqrt{\delta}$ is achieved. This completes the error analysis.