2016 IEEE 57th Annual Symposium on Foundations of Computer Science

Simulated Quantum Annealing Can Be Exponentially Faster than Classical Simulated Annealing

Elizabeth Crosson
Institute for Quantum Information and Matter
California Institute of Technology
Pasadena, CA 91125, USA
crosson@caltech.edu

Aram W. Harrow
Center for Theoretical Physics
Massachusetts Institute of Technology
Cambridge, MA 02139, USA
aram@mit.edu

***Abstract*—Can quantum computers solve optimization problems much more quickly than classical computers? One major piece of evidence for this proposition has been the fact that Quantum Annealing (QA) finds the minimum of some cost functions exponentially more quickly than classical Simulated Annealing (SA).**

One such cost function is the simple "Hamming weight with a spike" function in which the input is an n-bit string and the objective function is simply the Hamming weight, plus a tall thin barrier centered around Hamming weight $n/4$. While the global minimum of this cost function can be found by inspection, it is also a plausible toy model of the sort of local minima that arise in real-world optimization problems. It was shown by Farhi, Goldstone and Gutmann [1] that for this example SA takes exponential time and QA takes polynomial time, and the same result was generalized by Reichardt [2] to include barriers with width n^ζ and height n^α for $\zeta + \alpha \leq 1/2$. This advantage could be explained in terms of quantum-mechanical "tunneling."

Our work considers a classical algorithm known as Simulated Quantum Annealing (SQA) which relates certain quantum systems to classical Markov chains. By proving that these chains mix rapidly, we show that SQA runs in polynomial time on the Hamming weight with spike problem in much of the parameter regime where QA achieves an exponential advantage over SA. While our analysis only covers this toy model, it can be seen as evidence against the prospect of exponential quantum speedup using tunneling.

Our technical contributions include extending the canonical path method for analyzing Markov chains to cover the case when not all vertices can be connected by low-congestion paths. We also develop methods for taking advantage of warm starts and for relating the quantum state in QA to the probability distribution in SQA. These techniques may be of use in future studies of SQA or of rapidly mixing Markov chains in general.

I. Introduction

Classical algorithms are often useful but not provably so, with justifications for their success coming from a combination of empirical and heuristic evidence. For example, the simplex algorithm for linear programming was successful for decades before being proven to run in polynomial time, and for a long time was the most practical LP solver even while the ellipsoid algorithm was the only provably poly-time solver. Another example is MCMC (Markov chain Monte Carlo) which is used for applications in statistics, simulation, optimization and elsewhere, but almost never in regimes that are covered by formal proofs of correctness.

With quantum algorithms, there has been necessarily a greater emphasis on provable correctness. The present state of quantum computing technology does not yet allow us to test large-scale quantum algorithms empirically, nor can we usually empirically determine whether a proposed quantum algorithm outperforms all classical algorithms on worst-case inputs. Nevertheless, heuristic quantum algorithms are likely to be important for practical problems, just as they have been throughout the history of classical computing.

A particularly compelling heuristic proposal for optimization problems is quantum annealing (QA), also known as quantum adiabatic optimization [3], [4] (in this work we use the term "quantum annealing" to mean adiabatic optimization in thermal equilibrium at a low but non-zero temperature, though in some other contexts QA may be taken to include non-equilibrium thermal effects). The idea of QA is to interpolate between a static problem-independent Hamiltonian such as $-\sum_i \sigma_x^i$ for which we can efficiently prepare the ground state, and a final Hamiltonian whose ground state yields the desired answer. If we want to minimize a function $f : \{0,1\}^n \to \mathbb{R}$ then we can take this final Hamiltonian to be proportional to $\mathrm{diag}(f)$. This can be thought of as a quantum version of classical simulated annealing (SA) with the diagonal terms playing the role of bias and the off-diagonal terms causing hopping. Like SA its performance is hard to make provable general statements about, but it is a promising general-purpose heuristic, and rigorous statements about its performance

0272-5428/16 $31.00 © 2016 IEEE
DOI 10.1109/FOCS.2016.81

are known for many illustrative cases.

Intriguingly, QA has been shown to have an exponential asymptotic advantage over simulated annealing for certain cost functions [1], and here we focus on one such class of examples called the "Hamming weight with a spike" which has a cost function given by

$$f(z) := \begin{cases} |z| + n^{\alpha} : \frac{n}{4} - \frac{n^{\zeta}}{2} < |z| < \frac{n}{4} + \frac{n^{\zeta}}{2} \\ |z| \qquad\quad : o.w. \end{cases}, \quad (1)$$

where $|z|$ is the Hamming weight of the string z, and $\alpha > 0$, $\zeta \geq 0$ are independent of n. The global minimum of f is the string with $|z| = 0$, but the spike term creates a local minimum at $|z| = \lceil n/4 + n^{\zeta}/2 \rceil$. The spike presents a problem for a simulated annealing algorithm which only proposes moves that flip k bits for $k \leq n^{\zeta}$. First recall the definition of SA: starting with a point $x \in \{0,1\}^n$, repeatedly choose a random nearby point y (say within the Hamming ball of radius k) and move to y with probability $\min(1, e^{(f(x)-f(y))/T})$, where T is a temperature parameter that is gradually lowered. Following [1], consider first a SA algorithm that flips one bit at a time, i.e. with $k = 1$. Since SA begins at high temperature the initial state is overwhelmingly likely to have Hamming weight near $n/2$, and as the temperature of the system is lowered the random walk will move to strings of lower Hamming weight until reaching the local minimum at $\lceil n/4 + n^{\zeta}/2 \rceil$. This will happen for $T = O(1)$, so at this point the probability of accepting a move onto the spike is $e^{-\Omega(n^{\zeta})}$, and so classical SA requires exponential time to find the global minimum with high probability. This argument applies to flipping any $k < n^{\zeta}$ bits at once. Now suppose that the SA algorithm flips n^c bits at a time for some $c \leq 1$. Once the Hamming weight is $\approx n/4$, flipping n^c random bits will change the Hamming weight by a random variable with expectation $\approx \frac{1}{2}n^c$ and standard deviation $\approx \frac{\sqrt{3}}{2}n^{c/2}$. This has probability $e^{-\Omega(n^c)}$ probability of being negative, so with high probability the SA algorithm will not even attempt to move past the spike. In contrast, for any $\alpha + \zeta < 1/2$ it can be shown that QA finds the global minimum with high probability in time $\mathcal{O}(n)$ [2], showing that an exponential separation in the perfomance of SA and QA is possible.

While the spike is clearly a toy problem and can be solved efficiently by classical algorithms that exploit its structure, an important aspect of both QA and SA is that a single, general implementation of these algorithms is meant to be useful for solving a large variety of different problems without knowledge of their structure. Moreover, the spike arguably demonstrates a general advantage of QA over SA in tunneling through thin, high barriers in the energy landscape.

On the other hand, the standard formulation of QA uses a stoquastic Hamiltonian (i.e. a local Hamiltonian with non-positive off-diagonal matrix elements in the computational basis), and computational models based on ground states or thermal states of such systems are believed to be less powerful than universal quantum computation. In addition to complexity theoretic evidence [5], [6], suggestive evidence for this belief is also provided by the quantum-to-classical mapping of Suzuki et al. [7], [8], which allows for properties of low-energy states of a stoquastic Hamiltonian to be estimated using classical Markov chain Monte Carlo methods. These algorithms are known as quantum Monte Carlo (QMC) methods, and despite the name, are algorithms for classical computers. While QMC for stoquastic Hamiltonians is always a well-defined algorithm, its performance depends on the rate at which a Markov chain converges to its stationary distribution. This can range from polynomial to exponential time, and few general conditions are known in which it is provably polynomial-time. A few cases where the simulation can be made provably efficient are adiabatic evolution with frustration-free stoquastic Hamiltonians with a unique ground state [9] and ferromagnetic transverse Ising models in a large range of temperatures [10], but while these have some physics significance, they do not translate into nontrivial cost functions for QA.

When QMC is applied to QA Hamiltonians the result is an algorithm called simulated quantum annealing (SQA). Although there are examples for which standard versions of SQA take exponentially longer than the quantum evolution being simulated [11], [12], the general challenge from SQA to QA remains: for any purported speedup of QA we should see whether it can also be achieved by SQA. Moreover, since SQA is a Markov chain based algorithm on a domain that can be interpreted as a classical spin system, and since SQA is designed to sample from the output of a quantum optimization procedure, SQA can be considered as yet-another physics-inspired classical optimization method in its own right, which can naturally be compared with QA and with SA.

The main result of this paper is that the standard version of SQA, which does not use any structure of the problem, finds the global minimum of the cost function (1) in polynomial-time when $\alpha + \zeta < 1/2$.

Theorem 1. *Simulated quantum annealing based on the path-integral Monte Carlo method efficiently samples the output distribution of QA for the spike cost function*

(1) *when $\alpha + \zeta < 1/2$. The running time using single-qubit worldline updates is $\tilde{\mathcal{O}}(n^7)$, and the running time using single-site spin flips is $\tilde{\mathcal{O}}(n^{17})$. (Worldline and single-site spin flips are defined in Section II-B.)*

Thus SQA obtains an exponential speedup over SA for this particular problem. This result suggests that the benefits of tunneling through energy barriers with adiabatic evolution should not be thought of as an exclusively quantum advantage, since it can also be achieved by a general-purpose classical optimization algorithm.

Previous Work

There have been many past studies comparing the performance of SA and SQA using numerics [13], [14], [15] and more recently using physical methods such as the instanton approximation to tunneling [16], [17]. Studies comparing QA to SA and SQA have also emerged since [18] found the success probabilities of SQA are highly correlated with the results of QA performed on D-Wave quantum hardware with hundreds of qubits, while the distribution of success probabilities for SA on the same set of instances bears little resemblance to that of QA and SQA. More recently, the performance of QA, SQA, and SA was empirically compared on an ensemble of spin glass instances with were designed to have tall, thin barriers [19], as a step towards understanding the kinds of instances for which QA has an advantage over SA. In that work QA and SQA were found to have roughly the same scaling with system size for that particular ensemble of instances, though it was also pointed out that the large constant overhead in SQA made it less competitive in the sense of wall-clock times using modern classical hardware.

Another classical algorithm which has been compared with QA is spin vector dynamics (SVD), which approximates the state of the system by a collection of classical spin vectors, which are updated through either a nonlinear system of differential equations [20] or with Monte Carlo updates [21]. Although numerical evidence suggests that SVD can efficiently find the ground state of spike cost functions [22], a significant drawback of this method is that it uses an ad hoc dynamics and makes an uncontrolled approximation by neglecting the entanglement in the system, and so unlike SQA it is not guaranteed to systematically converge to a faithful simulation of the quantum system in the limit of increasing computational effort.

Without access to quantum hardware, comparison of SQA and QA is either limited to small system sizes where QA Hamiltonians can be exactly diagonalized ($\lesssim 50$ qubits), or to models for which analytical solutions of the quantum system are known (such as the spike problem we study here). We remark that the spike and related objective functions have the subject of recent analytic work [23], [24], and that there have also been numerical studies of SQA [25], [26], with findings that are consistent with our main result.

Proof Outline

Our proof of the efficient convergence of SQA on the spike problem involves bounding the mixing time of the underlying Markov chain, and there is an interesting parallel between a method which was used to lower bound the QA spectral gap when $\alpha + \zeta < 1/2$ [2]. There, a lower bound on the quantum gap can be found using a variational method with a trial wave function equal to the ground state of the system when no spike term is present (i.e. QA for the spikeless Hamming weight cost function $\tilde{f}(z) = |z|$). Similarly, we compare the spectral gap λ of the SQA Markov chain for the spike system with the spectral gap $\tilde{\lambda}$ of the spikeless system (throughout the subsequent sections we use tildes to distinguish quantities belonging to the spikeless system). Without a spike term, the quantum Hamiltonian $\tilde{H}$ is a tensor product operator with no interactions between the qubits. This trivial system translates in SQA to a collection of n non-interacting 1D classical ferromagnetic Ising models in a uniform magnetic field (which will become clear when the SQA Markov chain is described in detail in Section II-B), and upper bounding the mixing time for this system is relatively straightforward.

Let π and $\tilde{\pi}$ be the stationary distribution of the SQA Markov chain with and without the spike. These stationary distributions are close in a sense, $\|\pi - \tilde{\pi}\|_1 < \mathrm{poly}(n^{-1})$, but on the other hand there are exponentially many points $x \in \Omega$ for which the ratio $\pi(x)/\tilde{\pi}(x)$ is exponentially small. A review of existing comparison techniques concludes that none is quite suited to the present problem; indeed the review [27] states that there have been "relatively few successes in comparing chains with very different stationary distributions". To overcome this we introduce a comparison method which involves partitioning the state space into "good" and "bad" sets of vertices, $\Omega = \Omega_G \cup \Omega_B$. Beginning with a set of canonical paths yielding a bound $\tilde{\rho}$ on the congestion of the easy-to-analyze chain, we show that the paths which lie entirely within Ω_G can be used to construct an upper bound on the congestion ρ of the difficult-to-analyze chain, albeit within the set Ω_G of measure less than 1.

There are two main ingredients to the comparison method. First we show that if two chains have stationary distributions and transition probabilities that are similar on most of the points, then we can convert a known gap for one chain into a set of canonical paths for *most* of the state space of the other.

Theorem 2 (Most-paths comparison)**.** *Let (π, P) and $(\tilde{\pi}, \tilde{P})$ be reversible Markov chains with the same state space graph (Ω, E). Let $a = \max_{x\in\Omega} \pi(x)/\tilde{\pi}(x)$ and define $\Omega_\theta := \{x \in \Omega : \pi(x) < \theta\tilde{\pi}(x)\}$. If there is a set of canonical paths for $(\tilde{\pi}, \tilde{P})$ achieving congestion $\tilde{\rho}$ and satisfying $3a^2\tilde{\rho}\tilde{\pi}(\Omega_\theta) < 1$, then there is a subset $\Omega_G \subset \Omega$ with $\pi(\Omega_G) \geq 1 - 3a^2\tilde{\rho}\,\theta\pi(\Omega_\theta)$, and a canonical flow for (π, P) that connects every $x, y \in \Omega_G$ with paths contained in Ω_G for which the congestion ρ of any edge in Ω_G satisfies*

$$\rho \leq 16\,\theta \max_{x,y\in\Omega_G} \left[\frac{\tilde{P}(x,y)}{P(x,y)}\right] a^2\tilde{\rho}. \tag{2}$$

The proof of Theorem 2 is omitted from this extended abstract and can be found in our full version [28]. The intuition behind it is that low-congestion paths for $(\tilde{\pi}, \tilde{P})$ (which we have assumed exist) rarely overload edges of (π, P). This does not yet imply a high-probability subset for which all pairs have good paths between them, but it does imply a high-probability subset that is well-connected to most other points. By routing paths through this high-probability subset we can find a good set of paths for almost all other points.

There is a caveat here, which is that this bound on the congestion applies to the transitions of the Markov chain P on the subset Ω_G. If we assume that walkers leaving Ω_G are deleted then, since P is not restricted to Ω_G, these transitions form a substochastic "leaky" random walk on the set Ω_G, with a quasi-stationary distribution equal to π within this subset. (The term "quasi-stationary" refers to the fact that in the infinite time limit repeated applications of $P|_{\Omega_G}$ will converge to zero, but there may be a long intermediate time when we are close to π.)

Thus it is necessary to show that the chain mixes before it leaves the good set Ω_G. One way to guarantee this is to use a "warm start," meaning a starting sample from a distribution that is close to the quasi-stationary distribution. Our analysis of SQA relies on the adiabatic path used by the quantum algorithm to fulfill the warm-start condition.

Theorem 3. *Let (π, Ω, P) be a reversible Markov chain and suppose $\Omega = \Omega_G \cup \Omega_B$ is a partition. Let P_G be the substochastic transition matrix $P_G(x, y) := P(x, y)1_{x\in\Omega_G}1_{y\in\Omega_G}$. Suppose there is a set of canonical paths connecting every pair of points $x, y \in \Omega_G$, and the congestion of the walk P on this set of paths is ρ. If μ is a warm start with $\mu(x) \leq M\pi(x)$ for all $x \in \Omega_G$ then the distribution obtained by starting from μ and applying t steps of the random walk satisfies*

$$\|\mu P_G^t - \pi\|_1 \leq Mt\pi(\Omega_B) + \pi_{\min}^{-1} e^{-t/\rho} \tag{3}$$

The proof of Theorem 3 is also omitted and can be found in our full version [28]. The proof simply uses a union bound to show that a not-too-long random walk with a warm start is unlikely to ever encounter a bad point. The warm start here is crucial, since otherwise we could simply start within the bad set, and we need to bound the length of the random walk, since if the mixing time is τ then after time $\tau/\pi(\Omega_B)$, we would expect to hit the bad set.

The SQA state space can be interpreted as a path (worldline) representation of the original quantum system, and the bad states which constitute Ω_B will be those for which the paths spend too much "time" on the location of the spike (i.e. on strings with Hamming weight between $n/4-n^\zeta/2$ and $n/4+n^\zeta/2$). States that spend too much time on the spike are those for which $\pi(x)/\tilde{\pi}(x)$ is exponentially small, and naturally those are the ones we will need to exclude. In Section III we show that the mean spike time is proportional to the square of the ground state amplitude on the spike, while the m-th moment of the spike time distribution can also be bounded using the properties of the corresponding quantum system. Finally, we use the derived upper bound on the m-th moment of the spike time distribution to upper bound the probability of large deviations from the mean spike time, which yields an upper bound on $\pi(\Omega_B)$ that suffices to complete the proof.

Discussion

Our proof does not bound the convergence time for SQA $\alpha + \zeta > 1/2$, although QA does work for some values of (α, ζ) in this range, such as when $\zeta = 0, \alpha = O(1)$ [23] or $\alpha+2\zeta < 1$ [24]. We conjecture that SQA will be efficient for these values as well (which is supported by numerical evidence), though this will require extensions of the present techniques.

More generally, we believe that our most-paths comparison methods should have wider applicability. For example, consider a collection of classical particles with weak repulsive interactions. If the particles were non-interacting the thermal distribution of the particles would be easy to sample from, and if the interactions are weak enough, then they do not shift the typical

probabilities (or energies) by very much. While some configurations will have exponentially lower probability in the interacting case (if many particles are very close to each other), these configurations should be overall very unlikely. In this setting our framework should imply that the repulsive interactions do not significantly worsen the mixing time.

While relatively few rigorous facts are known about the general performance of SQA, it remains in practice a successful and widely used class of algorithms. This strikes us as an area where theorists should work to catch up with current practice.

II. Background

A. Quantum annealing

QA associates a cost function $f : \{0,1\}^n \to \mathbb{R}$ with a Hamiltonian that is diagonal in the computational basis,

$$H_f := \sum_{z\in\{0,1\}^n} f(z)|z\rangle\langle z| \quad , \tag{4}$$

so that the ground state of H_f is a computational basis state corresponding to the bit string that minimizes f. To prepare the ground state of H_f the system is initialized in the ground state of a uniform transverse field, which can be easily prepared,

$$H_0 := -\sum_{i=1}^{n} \sigma_i^x \quad , \quad |\psi_{\text{init}}\rangle := \frac{1}{\sqrt{2^n}} \sum_{z\in\{0,1\}} |z\rangle, \tag{5}$$

and then linearly interpolates between H_0 and H_f,

$$H := H(s) = (1-s)\, H_0 + sH_f \ , \tag{6}$$

where the adiabatic parameter s sweeps through the interval $0 \le s \le 1$. The total run time $t_{\max}$ of the algorithm depends on how quickly the adiabatic parameter is adjusted, which defines a time-dependent Hamiltonian $H(t) := H(s = t/T)$. At zero temperature the system evolves according to the Schrödinger equation, $\frac{d}{dt}|\psi(t)\rangle = -iH(t)\psi(t)$, and the adiabatic theorem ensures that the state $\psi(T)$ at the end of the evolution has a high overlap with the ground state of H_f as long as $T \ge \text{poly}(n, \Delta^{-1})$, where $\Delta = \min_s E_1(s) - E_0(s)$ is the minimum gap between the two lowest eigenvalues of $H(s)$ during the evolution.

More generally (and realistically) we can take the state of the system to be not the ground state but a thermal state with inverse temperature $\beta < \infty$. The equilibrium thermal state of the system evolves with the adiabatic parameter,

$$\sigma(s) := \frac{e^{-\beta H(s)}}{\mathcal{Z}(s)} \quad , \quad \mathcal{Z}(s) := \operatorname{tr} e^{-\beta H(s)}. \tag{7}$$

As described in more detail in Section III, the thermal equilibrium state will have a high overlap with the ground state at sufficiently low temperatures, and so the distribution $\Pi_s(z) := \langle z|\sigma(s)|z\rangle$ can be sampled to determine the minimum of f.

B. Simulated quantum annealing

Stoquastic Hamiltonians such as (6) are amenable to a variety of classical Markov chain based simulation algorithms (at least in principle), which are collectively known as quantum Monte Carlo (QMC) methods (the term "stoquastic" is a combination of "quantum" + "stochastic" in the sense of stochastic matrices [5]). Any QMC method applied to the QA Hamiltonian (6) defines a version of SQA. Here we consider a version based on the path-integral representation of the thermal state (7). See [10] for a full derivation.

The state space of the SQA Markov chain is the set of trajectories $(x_1, ..., x_L)$, where $x_i \in \{0,1\}^n$ and L is a polynomial in n which is $\Theta(n^2\beta^{3/2})$ in order for certain standard approximations to hold with high precision. The stationary distribution of the SQA Markov chain is

$$\pi(x_1, ..., x_L) = \frac{1}{Z} e^{-\frac{\beta s}{L}\sum_{i=1}^{L} f(x_i)} \prod_{j=1}^{n} \phi(\bar{x}_j) \tag{8}$$

where $\bar{x}_j := (x_{j,1}, ..., x_{j,L})$ is called "the worldline of the j-th qubit", where $\omega := \beta(1-s)/L$, and $\phi(\bar{x}_j) := \tanh(\omega)^{|\{k:x_{j,k}\neq x_{j,k+1}\}|}$ counts the number of consecutive bits which disagree in that worldline. The quantum distribution $\Pi(x) = \langle x|\sigma|x\rangle$ arising from (7) can be expressed as a marginal of π,

$$\Pi(x) = \sum_{x_2,\ldots,x_L} \pi(x, x_2, ..., x_L). \tag{9}$$

From this point SQA proceeds by discretizing the adiabatic path and using the Markov chain Monte Carlo method to sample from π at various values of the adiabatic parameter $s_1, s_2, \ldots, s_{\max}$, with $s_1 \approx 0$ and $s_{\max} \approx 1$. We will analyze two discrete-time Markov chains with stationary distribution (8). The first chain consists of **single-site Metropolis updates**. If $x, x' \in \Omega$ differ by a single bit, the transition probability from x to x' is

$$P_M(x, x') = \frac{1}{2nL} \min\left\{1, \frac{\pi(x')}{\pi(x)}\right\}, \tag{10}$$

and otherwise the transition probability is zero.

In practical implementations of SQA it is important to speed up the equilibration using non-local cluster updates, therefore in addition to the single-site Metropolis updates defined above we will analyze the popular

single-qubit heat-bath worldline updates, which are a form of generalized heat-bath updates [29].

Definition. *The heat-bath worldline update* $(\bar{x}_1, \ldots \bar{x}_n) \to (\bar{x}'_1, \ldots, \bar{x}'_n)$ *proceeds as follows:*

1) *Select a site* $i \in \{1, \ldots, n\}$ *uniformly at random.*
2) *Set* $\bar{x}'_j = \bar{x}_j$ *for all* $j \neq i$.
3) *Choose* $\bar{x}'_i$ *from the conditional distribution* $\pi(\bar{x}'_i | \bar{x}'_1, \ldots, \bar{x}'_{i-1}, \bar{x}'_{i+1}, \ldots, \bar{x}'_n)$.

As with all generalized heat-bath updates these transitions define a Markov chain which is reversible with respect to π. An algorithm that efficiently implements these transitions is given in [30] and without using any structure of the cost function (1) it runs in $\tilde{\mathcal{O}}(n^2\beta^2)$ elementary steps with high probability.

C. Mixing times of Markov chains

At each value of the adiabatic parameter, the run-time of SQA will be determined by the mixing time of either of the Markov chains described above. This quantity can be defined in terms of the total variation distance from π to the distribution $P^t(x, \cdot)$ obtained by running the chain for t steps starting from x,

$$d_x(t) := \max_{A \subseteq \Omega} |P^t(x, A) - \pi(A)|, \tag{11}$$

with the mixing time $\tau(\epsilon)$ being the worst-case time needed to be within variation distance ϵ of the stationary distribution,

$$t_{\text{mix}}(\epsilon) := \max_{x \in \Omega} \min_t \{t : d_x(t') \leq \epsilon \ \ \forall t \geq t'\}. \tag{12}$$

A standard way to bound the mixing time is to relate it to the spectral gap λ of the transition matrix P [31]. For all $x \in \Omega$,

$$\|P^t(x, \cdot) - \pi\|_1 \leq \pi_{\min}^{-1} e^{-\lambda t}. \tag{13}$$

which implies $t_{\text{mix}}(\epsilon) \leq \lambda^{-1} \log\left(\frac{1}{\epsilon \pi_{\min}}\right)$. Finally, in Section III we will use the method of canonical paths to bound the the spectral gap of a chain in terms of its congestion. For any set of paths $\{\gamma_{xy}\}$ that connects every pair of points in the state space, define the congestion $\rho(e)$ through the edge e to be

$$\rho := \frac{1}{Q(e)} \sum_{\substack{x,y \in \Omega \\ e \in \gamma_{xy}}} \pi(x)\pi(y)|\tilde{\gamma}_{xy}|, \tag{14}$$

where $Q(x, y) = \pi(x)P(x, y)$. The maximum congestion through any edge, $\rho := \max_e \rho(e)$, is related to the spectral gap by $\lambda \geq 1/\rho$ [31].

III. Efficient convergence of SQA for the spike cost function

In this section we apply Theorem 2 with the easy-to-analyze chain $(\tilde{\pi}, \tilde{P})$ taken to be the SQA Markov chain with heat-bath worldline updates for the system without the spike, and (π, P) equal to the corresponding chain for the spike system.

The subset Ω_B will be shown to satisfy $\tilde{\pi}(\Omega_B) \leq \mathcal{O}(n^{-c})$ for a constant c that we will choose so that we can prove a walker beginning in Ω_G is likely to mix before it hits a point in Ω_B.

Congestion of the spikeless chain. Recall that Ω_B is defined in terms of a set of canonical paths $\{\gamma\}$ on Ω with congestion $\tilde{\rho}$ for the spikeless chain, together with a subset Ω_θ of points which are excluded from paths in $\{\gamma\}$ to obtain a new set of paths with congestion $\rho \leq \mathcal{O}(\theta\tilde{\rho})$ for the chain with the spike, within the subset Ω_G. The spikeless distribution $\tilde{\pi}$ corresponds to a collection of n non-interacting 1D ferromagnetic Ising models of length L in the presence of 1-local fields that bias the distribution towards configurations of lower Hamming weight. The spin-spin coupling is such that each broken bond in x lowers $\tilde{\pi}(x)$ by a factor of $\Theta(\tanh(\omega))$.

First we will bound the congestion of heat-bath worldline updates (defined in Section II-B). Here it is convenient to represent states $x \in \Omega$ by their worldlines $x = (\bar{x}_1, \ldots, \bar{x}_n)$, where $\bar{x}_i := (x_{i,1}, \ldots, x_{i,L})$. For the spikeless system $(\tilde{\pi}, \tilde{P})$ spins in different worldlines do not interact and so the conditional distribution of the i-th worldline is equal to the marginal of the stationary distribution on that worldline,

$$\tilde{\pi}(\bar{x}'_i | \bar{x}'_1, \ldots, \bar{x}'_{i-1}, \bar{x}'_{i+1}, \ldots, \bar{x}'_n) = \sum_{\bar{x}_j : j \neq i} \tilde{\pi}(\bar{x}_1, \ldots, \bar{x}'_i, \ldots, \bar{x}_n).$$

Including a $1/2$ probability of the chain not moving at each step in order to make it irreducible, the probability of a transition $P\left((\bar{z}_1, \ldots, \bar{z}_i, \ldots, \bar{z}_n), (\bar{z}'_1, \ldots, \bar{z}'_i, \ldots, \bar{z}'_n)\right)$ that updates the i-th worldline ($\bar{z}_j = \bar{z}'_j$ for all $j \neq i$) is

$$\frac{1}{2n} \sum_{\bar{z}''_j : j \neq i} \tilde{\pi}(\bar{z}''_1, \ldots, \bar{z}'_i, \ldots, \bar{z}''_n) \tag{15}$$

The path γ_{xy} from $x = (\bar{x}_1, \ldots, \bar{x}_n)$ to $y = (\bar{y}_1, \ldots, \bar{y}_n)$ proceeds by updating the worldlines in order $\{1, \ldots, n\}$. The paths have length $|\gamma_{xy}| = n$. The k-th step of the path γ_{xy} will go through the edge (z, z') with $z = (\bar{y}_1, \ldots, \bar{y}_{k-1}, \bar{x}_k, \ldots, \bar{x}_n)$ and $z' = (\bar{y}_1, \ldots, \bar{y}_k, \bar{x}_{k+1}, \ldots, \bar{x}_n)$. To evaluate the sum

(14) we apply the standard encoding trick [32]. Define an injective function $\eta_{(z,z')}$ which maps the paths through (z, z') into the state space Ω,

$$\eta_{(z,z')}(\gamma_{xy}) = (\bar{x}_1, \ldots, \bar{x}_{k-1}, \bar{y}_k, \ldots, \bar{y}_n),$$

which is injective since z and $\eta_{(z,z')}(\gamma_{xy})$ provide sufficient data to uniquely determine γ_{xy}. Notice that $\tilde{\pi}(x)\tilde{\pi}(y) = \tilde{\pi}(z)\tilde{\pi}\left(\eta_{(z,z')}(\gamma_{xy})\right)$, and so the congestion $\tilde{\rho}(z, z')$ is

$$= \frac{1}{\tilde{\pi}(z)P(z,z')} \sum_{\gamma_{xy} \ni (z,z')} \tilde{\pi}(x)\tilde{\pi}(y)|\gamma_{xy}| \tag{16}$$

$$= \frac{n}{\tilde{P}(z,z')} \sum_{y_{xy} \ni (z,z')} \tilde{\pi}\left(\eta_{(z,z')}(\gamma_{xy})\right) \tag{17}$$

$$= \frac{n}{\tilde{P}(z,z')} \sum_{\substack{\bar{x}_1,\ldots,\bar{x}_{k-1} \\ \bar{y}_{k+1},\ldots,\bar{y}_n}} \tilde{\pi}(\bar{x}_1, ..., \bar{x}_{k-1}, \bar{y}_k, ..., \bar{y}_n) \tag{18}$$

$$= 2n^2, \tag{19}$$

where in going from (17) to (18) we do not sum over $\bar{y}_k$ because it is fixed by z'. Finally, since the edge (z, z') was arbitrary we have

$$\tilde{\rho} = \mathcal{O}(n^2). \tag{20}$$

Now we will analyze the single-site Metropolis chain for the spikeless system $(\tilde{\pi}, \tilde{P}_M)$. The path from $x = ((x_{1,1}, \ldots, x_{1,L}), \ldots, (x_{n,1}, \ldots, x_{n,L}))$ to $y = ((y_{1,1}, \ldots, y_{1,L}), \ldots, (y_{n,1}, \ldots, y_{n,L}))$ proceeds as follows: for each i in order from $\{1, \ldots, n\}$ update spin (i, j) for j going in order from $\{1, \ldots, L\}$. The paths have length $|\gamma_{xy}| = nL$. Edges (z, z') along such a path will create at most two new broken bonds (i.e. pairs of spins which disagree) along the direction $\{1, \ldots, L\}$ so $P_M(z, z')$ is $\Omega\left((nL)^{-1}\tanh(\omega)\right)$.

Just as for the heat-bath worldline updates, we apply an encoding function to injectively map the paths through an arbitrary edge into the state space. The only difference from the version above is that z and $\eta_{(z,z')}(\gamma_{xy})$ may in the worst-case each have two broken imaginary-time bonds which are not present in either x or y, and so

$$\tilde{\pi}(x)\tilde{\pi}(y) = \mathcal{O}\left(\coth(\omega)^2 \tilde{\pi}(z)\tilde{\pi}\left(\eta_{(z,z')}(\gamma_{xy})\right)\right). \tag{21}$$

Applying the same calculations in (16)-(18) but using (21) along with the the fact that $\coth(\omega) = \mathcal{O}(\omega_{\min}^{-1}) = \mathcal{O}(nL\beta^{-1})$, the congestion is

$$\tilde{\rho}_M = \mathcal{O}(L^5 n^5 \beta^{-3}) = \mathcal{O}(n^{15}\beta^{-9/2}). \tag{22}$$

Comparing (22) with (20) shows that there is a large polynomial overhead resulting from single-site updates, which arises from the strong interactions that occur in the imaginary-time direction when the transverse field term of the QA Hamiltonian is small.

Ω_θ and the spike time distribution. The states in Ω_θ which will be excluded from the set of paths $\{\gamma\}$ are those which have $|x_i| \in I_S := (n/4 - n^\zeta/2, n/4 + n^\zeta/2)$ for too many i. Define $1_S : \{0,1\}^n \to \{0,1\}$ to be the indicator function for the spike i.e. $1_S(z) = 1$ if $z \in I_S$, and $1_S(z)$ is zero otherwise. The spike time for $x \in \Omega$ is defined to be $\mathrm{ST}(x) := \sum_{i=1}^{L} 1_S(x_i)$. Let $\epsilon := \frac{1}{2} - \alpha$, and define

$$\Omega_\theta = \left\{ x \in \Omega : \mathrm{ST}(x) \geq \frac{L}{n^{\frac{1}{2}(1-\epsilon)-\zeta}} \right\}. \tag{23}$$

Set $\beta = n^{\epsilon/2}$ so that $\pi(x)/\tilde{\pi}(x)$ is $\Omega(1)$ for every $x \notin \Omega_\theta$.which shows $\theta = \mathcal{O}(1)$ in the congestion bound in Theorem 2. The remainder of the section will be devoted to computing the m-th moment of the random variable $\mathrm{ST} \sim \tilde{\pi}$, with $m = c/\epsilon$, in order to show,

$$\Pr\left[\mathrm{ST} \geq \frac{L}{n^{\frac{1}{2}(1-\epsilon)-\zeta}}\right]_{\tilde{\pi}} \leq \mathcal{O}(n^{-c}), \tag{24}$$

which is equivalent to the statement $\tilde{\pi}(\Omega_\theta) \leq \mathcal{O}(n^{-c})$.

To calculate the moments $\langle \mathrm{ST}^m \rangle_{\tilde{\pi}}$ we will relate them to expectation values of the spikeless quantum system, and use the fact that the latter is exactly solvable because the qubits are non-interacting. Let $\{|k\rangle : k = 0, \ldots, n\}$ be a basis of states for the symmetric subspace which are labeled by Hamming weight, and let $S = \sum_{k \in I_S} |k\rangle\langle k|$. Since the observable S is diagonal in the computational basis we can include the term λS into the diagonal part of the Hamiltonian for the quantum-to-classical mapping and find that

$$\langle S \rangle_{\tilde{\sigma}} = L^{-1} \langle \mathrm{ST} \rangle_{\tilde{\pi}}. \tag{25}$$

Let $|\tilde{\psi}_1\rangle, \ldots, |\tilde{\psi}_n\rangle$ denote the excited eigenstates of $\tilde{H}$. Define $\Delta := 2\sqrt{(1-s)^2 + s^2}$ and observe that $|\tilde{\psi}_k\rangle$ is an eigenstate of $\tilde{H}$ with eigenvalue $k\Delta$, and that the degeneracy of the k-th energy level is $\binom{n}{k}$ so

$$\|\tilde{\sigma} - |\tilde{\psi}_0\rangle\langle\tilde{\psi}_0|\|_1 \leq \sum_{k=1}^{n} e^{-\beta\Delta k} \binom{n}{k}, \tag{26}$$

which is $\mathcal{O}(ne^{-n^\epsilon})$ and since $\epsilon > 0$ is a constant this error will be sub-leading, and this justifies replacing the thermal state in (25) with the ground state.

The ground state probability distribution for the spikeless system is a binomial distribution [2] on the Hamming weights, and so $\langle S \rangle_{\tilde{\sigma}}$ is asymptotically never larger than the central binomial coefficient times the width n^ζ of the spike term, which implies

$$\langle \mathrm{ST} \rangle_{\tilde{\pi}} = \mathcal{O}\left(Ln^{\zeta - 1/2}\right). \tag{27}$$

To obtain (24) we will use the moment inequality,

$$\Pr\left[\mathrm{ST} \geq b\right]_{\tilde{\pi}} \leq \frac{\langle \mathrm{ST}^m \rangle_{\tilde{\pi}}}{b^m}, \tag{28}$$

with $b = Ln^{-\frac{1}{2}(1-\epsilon)}$. Expanding the definition we have

$$\langle \mathrm{ST}^m \rangle_{\tilde{\pi}} = \sum_{t_1,\ldots,t_m}^{L} \langle 1_S(z_{t_1}) \ldots 1_S(z_{t_m}) \rangle_{\tilde{\pi}}. \tag{29}$$

To compute these m-point correlation functions we return to the quantum description,

$$\langle 1_S(z_{t_1}) \ldots 1_S(z_{t_m}) \rangle_{\tilde{\pi}} = \left\langle \prod_{i=1}^{m} e^{-(\tau_i - \tau_{i-1})H} S \right\rangle_{\tilde{\sigma}}$$

where $\tau_i := \beta t_i / L$. Once again we replace the low-temperature thermal state with the ground state and incur a sub-leading error as in (26). Since the ground state, the Hamiltonian, and the operator S are all bit-symmetric, the expectation can be evaluated in the symmetric subspace. Using the basis of symmetric energy eigenstates $\{|\tilde{\psi}_k\rangle\}$, $\langle \mathrm{ST}^m \rangle_{\tilde{\pi}}$ can be expressed as

$$\sum_{\substack{k_1,\ldots,k_m \\ t_1,\ldots,t_m}} \prod_{i=1}^{m} e^{-(\tau_{i+1} - \tau_i)\Delta k_i} \langle \psi_{k_{i+1}} | S | \psi_{k_i} \rangle \tag{30}$$

States with higher energy will contribute less to the sum over all times $t_1, \ldots, t_m$ in (30)because the exponentials decay more quickly. For $k_i > 0$, the sum over t_i can be truncated whenever $\tau_i - \tau_{i-1} \gg 1/k_i\Delta$.

Since the ground state wave function is a binomial distribution the mean spike time will only be large when the peak of the ground state is near the support of the spike I_S. In the range of the adiabatic parameter in which this occurs the excited spikeless eigenstates satisfy $\langle \tilde{\psi}_i | k \rangle \leq |\langle \tilde{\psi}_0 | k \rangle| \leq \mathcal{O}(n^{-1/4})$ for all $i = 1, \ldots, n$ and $k \in I_S$, because the ground state wave function is centered on the spike and the excited state wave functions have a greater spread, which can be seen from the explicit form of the spikeless eigenfunctions given in [23]. Now we define $g_i := t_i - t_{i-1}$ and relabel the sum of $t_1, \ldots, t_m = 0, \ldots, L$ by a sum over the g_i. For the purpose of obtaining an upper bound on the m-th moment we relax the constraint $\sum_i g_i = L$, and instead sum over the full range $g_i = 1, \ldots, L$ for each i. Using these facts we can upper bound (30) by

$$\langle \mathrm{ST}^m \rangle_{\tilde{\pi}} \leq n^{m(\zeta - 1/2)} \sum_{\substack{k_1,\ldots,k_m \\ g_1,\ldots,g_m}} e^{-\sum_{i=1}^{m} g_i \Delta k_i} \tag{31}$$

We will now organize the terms of (31) according to the number ℓ of excited energies $\tilde{E}_{k_i} > 0$ they contain. There are $\binom{m}{\ell}$ terms of (31) that contain ℓ eigenstates above the ground state, and for each ℓ we must sum over the $g_{a_1}, \ldots, g_{a_\ell}$ for which the corresponding $k_{a_1}, \ldots, k_{a_\ell}$ are non-zero. Now (31) becomes

$$\leq n^{m(\zeta - 1/2)} \sum_{\ell=1}^{m} \binom{m}{\ell} \frac{L^{m-\ell}}{(m-\ell)!} \sum_{\substack{g_{a_1},\ldots g_{a_\ell} \\ k_{a_1} \ldots k_{a_\ell}}} \prod_{i=1}^{\ell} e^{-g_{a_i} \Delta k_{a_i}}$$

where the factor of $L^{m-\ell}/(m-\ell)!$ results from performing the sum over the $m - \ell$ of the g_i which have $k_i = 0$. Now we sum over $g_{a_1}, \ldots, g_{a_\ell}$ using the fact that $\sum_{g=1}^{L} e^{-gk} \leq k^{-1}$,

$$\leq L^m n^{m(\zeta - 1/2)} \sum_{\ell=0}^{m} \binom{m}{\ell} (\beta\Delta)^{-\ell} \sum_{k_{a_1} \ldots k_{a_\ell}}^{n} \prod_{i=1}^{\ell} \frac{1}{k_i}.$$

Using $\binom{m}{\ell} \leq m^\ell$ and $\sum_{k=1}^{n} \leq \log(n) + 1$, at last this becomes

$$\langle \mathrm{ST}^m \rangle_{\tilde{\pi}} \leq L^m n^{m(\zeta - 1/2)} \sum_{\ell=0}^{m} \left(\frac{\log(n) + 1}{m\beta\Delta} \right)^{\ell} \tag{32}$$

Since m and Δ are constant and $\beta = n^{\epsilon/2}$ with fixed $\epsilon > 0$ the terms with inverse powers of β are sub-leading and so $\langle \mathrm{ST}^m \rangle_{\tilde{\pi}} \leq \mathcal{O}(L^m n^{m(\zeta - 1/2)})$. Finally, applying (28) with $m = c/\epsilon$ yields the desired result (24).

Adiabatic schedule. Here we show that a discretization of the adiabatic path with $1/\operatorname{poly}(n)$ step size is sufficient to fulfill the statement we need for the warm starts in Theorem 3. We will take the largest value of the adiabatic parameter to be $s_{\max} = 1 - n^{-1}$ so that $||\psi_0(1)\rangle - |\psi_0(s_{\max})\rangle|_1 \leq \operatorname{poly}(n^{-1})$, and the global minimum of the cost function can be obtained by sampling from Π at $s = s_{\max}$ with effectively the same probability at it would be obtained by sampling from the ground state probability distribution.

We will sample from π at several values of the adiabatic parameter $s_1, \ldots, s_{\max}$. Define $s_i := s_0 - i\Delta s$, $\omega_i := \beta(1 - s_i)/L$, $\Delta\omega := \beta\Delta s/L$, and let π^i be the stationary distribution (8) when the adiabatic parameter is s_i. At each stage we simulate the Markov chain (10) for sufficiently many steps to achieve a variational distance to the stationary distribution of $\exp(-n^{\Omega(1)})$. These errors then add up to a negligible amount. To choose a step size Δs satisfying the warm start condition $\pi_{i+1} \leq 2\pi_i$ we'll use a claim which is inspired by Lemma 5.1 of [9] but is a bit simpler in the classical case.

Lemma 4. *Let $E_1, E_2 : A \to \mathbb{R}$ be energy functions on a domain A and define $Z_i := \sum_{x \in A} e^{-E_i(x)}$ and $p_i(x) = e^{-E_i(x)}/Z_i$ for $i = 1, 2$. If $\max_x |E_1(x) - E_2(x)| \leq \delta$, then $|\log(Z_1/Z_2)| \leq \delta$ and $\max_x |\log(p_1(x)/p_2(x))| \leq 2\delta$.*

Proof. Applying the uniform bound $|E_2(x)-E_1(x)| \leq \delta$ for all $x \in A$ to the sum $Z_2 = \sum_{x\in A} e^{-E_1(x)}$ leads to $e^{-\delta}Z_1 \leq Z_2 \leq e^{\delta}Z_1$, therefore $|\log(Z_1/Z_2)| \leq \delta$. Using the triangle inequality $\max_x |\log(p_1(x)/p_2(x))|$ can be expressed as $|\log(Z_2/Z_1)| + \max_x |\log(e^{E_2(x)-E_1(x)})|$ which is then at most 2δ. Since $|\log(p_1(x)/p_2(x))| = |\log(p_2(x)/p_1(x))|$ the other case follows similarly.

Applying Lemma 4 with the form of the stationary distribution (8) we may take

$$\delta = \frac{\beta}{L}\Delta s f_{\max} + nL\log\left(\frac{\tanh(\omega_i - \Delta\omega)}{\tanh(\omega_i)}\right),$$

which is $\mathcal{O}\left(\beta\Delta s n\log(n)\right)$ and so taking Δs to be $\mathcal{O}\left((\beta n\log(n))^{-1}\right)$ fulfills the warm start condition.

Quasi-stationary mixing. Finally we will bound the overall run time of SQA applied to the spike cost function. First we need to show that taking c in (24) to be a sufficiently large constant will allow the leaky walk for the spike system sample from Ω_G according to the quasi-stationary distribution π for an expected time of n^q, for any desired constant q, before it is eventually likely to escape into Ω_B.

Inserting $\tilde{\rho}$ from (20) into Theorem 2 yields $\rho = \mathcal{O}(n^2)$. To apply Theorem 3 we next must consider $\pi_{\min}^{-1}$. From (8) we have $\log\pi_{\min}^{-1} \leq \mathcal{O}(nL\log(n))$ because there can be L pairs of bits which disagree in each of the n worldlines. However, according to (8) these disagreements follow a binomial distribution with mean $\beta(1-s)$, and so we may abort the algorithm with exponentially small probability if it ever encounters a configuration with $\Omega(\beta\log n)$ jumps in any worldline, which allows us to take $\log\pi_{\min}^{-1} = \mathcal{O}(n\beta\log(n)$. This implies a mixing time of $t_{\text{mix}} = \mathcal{O}(n^3\beta\log n)$ within Ω_G for the SQA spike chain with single qubit worldline updates at each value s_i of the adiabatic path.

Meanwhile, from (24) we have $\pi(\Omega_B) \leq \Theta(\tilde{\rho}\tilde{\pi}(\Omega_\theta)) = \Theta(n^{2-c})$. At each step s_i of the adiabatic path, after time $t \geq t_{\text{mix}}$ the leaky random walk mixes to within a distance $\mathcal{O}(t\pi(\Omega_B))$ so by Theorem 3 it suffices to take $c = 2 + q + \log(1/\delta)$ in order for the leaky walk to be with distance δ to the stationary distribution π for times $t_{\text{mix}} \leq t \leq \Theta(n^q)$.

Finally, since there are $\tilde{\mathcal{O}}(n\beta)$ steps of the adiabatic path, and each worldline update takes $\tilde{\mathcal{O}}(n^2\beta^2)$ time to implement, we obtain a total run time of $\tilde{\mathcal{O}}(n^6\beta^4)$, which implies the $\tilde{\mathcal{O}}(n^7)$ stated in Theorem 1. Repeating the analysis above for single-site Metropolis updates, we combine the congestion (22) with $\log\pi_{\min}^{-1} = \tilde{\mathcal{O}}(n\beta)$ and $\tilde{\mathcal{O}}(n\beta)$ steps of the adiabatic path to arrive bound the total run time by $\tilde{\mathcal{O}}(n^{17})$ as stated in Theorem 1.

ACKNOWLEDGMENTS

We thank Dave Bacon, Wim van Dam and Alistair Sinclair for helpful conversations. Elizabeth Crosson gratefully acknowledges funding provided by the Institute for Quantum Information and Matter, an NSF Physics Frontiers Center (NSF Grant PHY-1125565) with support of the Gordon and Betty Moore Foundation (GBMF-12500028), and is also grateful for support received while completing a portion of this work at the MIT Center for Theoretical Physics with funding from NSF grant number CCF-1111382. Aram Harrow was funded by NSF grants CCF-1111382 and CCF-1452616 and ARO contract W911NF-12-1-0486.

REFERENCES

[1] E. Farhi, J. Goldstone, and S. Gutmann, "Quantum adiabatic evolution algorithms versus simulated annealing," 2002.

[2] B. W. Reichardt, "The quantum adiabatic optimization algorithm and local minima," in *Proceedings of the thirty-sixth annual ACM symposium on Theory of computing*. ACM, 2004, pp. 502–510.

[3] T. Kadowaki and H. Nishimori, "Quantum annealing in the transverse Ising model," *Phys. Rev. E*, vol. 58, pp. 5355–5363, Nov 1998.

[4] E. Farhi, J. Goldstone, S. Gutmann, and M. Sipser, "Quantum computation by adiabatic evolution," 2000.

[5] S. Bravyi, D. P. DiVincenzo, R. I. Oliveira, and B. M. Terhal, "The complexity of stoquastic local Hamiltonian problems," *Quant. Inf. Comp.*, vol. 8, no. 5, pp. 0361–0385, 2006.

[6] S. Bravyi, A. J. Bessen, and B. M. Terhal, "Merlin-Arthur games and stoquastic complexity," 2006.

[7] M. Suzuki, S. Miyashita, and A. Kuroda, "Monte Carlo simulation of quantum spin systems," *Prog. Theor. Phys.*, vol. 58, no. 5, pp. 1377–1387, 1977.

[8] M. Suzuki, "Quantum statistical Monte Carlo methods and applications to spin systems," *Journal of Statistical Physics*, vol. 43, no. 5-6, pp. 883–909, 1986.

[9] S. Bravyi and B. M. Terhal, "Complexity of stoquastic frustration-free Hamiltonians," *SIAM J. Comput.*, vol. 39, no. 4, pp. 1462–1485, 2009.

[10] S. Bravyi, "Monte Carlo simulation of stoquastic Hamiltonians," 2014.

[11] M. B. Hastings, "Obstructions to classically simulating the quantum adiabatic algorithm," *Quantum Information & Computation*, vol. 13, no. 11-12, pp. 1038–1076, 2013.

[12] M. Jarret, S. P. Jordan, and B. Lackey, "Adiabatic optimization versus diffusion monte carlo," *arXiv preprint arXiv:1607.03389*, 2016.

[13] R. Martoňák, G. E. Santoro, and E. Tosatti, "Quantum annealing by the path-integral Monte Carlo method: The two-dimensional random Ising model," *Phys. Rev. B*, vol. 66, p. 094203, Sep 2002.

[14] D. Battaglia, G. E. Santoro, and E. Tosatti, "Optimization by quantum annealing: Lessons from hard 3-SAT cases," *Phys. Rev. E*, 2005.

[15] E. Inack and S. Pilati, "Simulated quantum annealing of double-well and multiwell potentials," *Physical Review E*, vol. 92, no. 5, p. 053304, 2015.

[16] S. V. Isakov, G. Mazzola, V. N. Smelyanskiy, Z. Jiang, S. Boixo, H. Neven, and M. Troyer, "Understanding quantum tunneling through quantum Monte Carlo simulations," 2015.

[17] Z. Jiang, V. N. Smelyanskiy, S. V. Isakov, S. Boixo, G. Mazzola, M. Troyer, and H. Neven, "Scaling analysis and instantons for thermally-assisted tunneling and quantum monte carlo simulations," *arXiv preprint arXiv:1603.01293*, 2016.

[18] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, "Quantum annealing with more than one hundred qubits," *Nature Phys.*, p. 218, 2014.

[19] V. S. Denchev, S. Boixo, S. V. Isakov, N. Ding, R. Babbush, V. Smelyanskiy, J. Martinis, and H. Neven, "What is the computational value of finite range tunneling?" 2015.

[20] J. A. Smolin and G. Smith, "Classical signature of quantum annealing," *arXiv preprint arXiv:1305.4904*, 2013.

[21] S. W. Shin, G. Smith, J. A. Smolin, and U. Vazirani, "How" quantum" is the d-wave machine?" *arXiv preprint arXiv:1401.7087*, 2014.

[22] S. Muthukrishnan, T. Albash, and D. A. Lidar, "Tunneling and speedup in quantum optimization for permutation-symmetric problems," *arXiv preprint arXiv:1511.03910*, 2015.

[23] L. Kong and E. Crosson, "The performance of the quantum adiabatic algorithm on spike Hamiltonians," 2015.

[24] L. T. Brady and W. van Dam, "Spectral gap analysis for efficient tunneling in quantum adiabatic optimization," 2016.

[25] E. Crosson and M. Deng, "Tunneling through high energy barriers in simulated quantum annealing," 2014.

[26] L. T. Brady and W. van Dam, "Quantum Monte Carlo simulations of tunneling in quantum adiabatic optimization," 2015.

[27] M. Dyer, L. A. Goldberg, M. Jerrum, and R. Martin, "Markov chain comparison," *Probability Surveys*, vol. 3, pp. 89–111, 2006.

[28] E. Crosson and A. W. Harrow, "Simulated quantum annealing can be exponentially faster than classical simulated annealing," *arXiv preprint arXiv:1601.03030*, 2016.

[29] M. Dyer, A. Sinclair, E. Vigoda, and D. Weitz, "Mixing in time and space for lattice spin systems: a combinatorial view," in *Randomization and approximation techniques in computer science*. Springer, 2002, pp. 149–163.

[30] E. Farhi, J. Goldstone, D. Gosset, S. Gutmann, H. B. Meyer, and P. Shor, "Quantum adiabatic algorithms, small gaps, and different paths," *arXiv preprint arXiv:0909.4766*, 2009.

[31] D. Levin, Y. Peres, and E. Wilmer, *Markov Chains and Mixing Times*. American Mathematical Soc., 2008.

[32] A. Sinclair, "Improved bounds for mixing rates of Markov chains and multicommodity flow," *Combinatorics, Probability and Computing*, vol. 1, no. 4, pp. 351–370, 1992.