Mind the Gap: Achieving a Super-Grover Quantum Speedup by Jumping to the End

Alexander M. Dalzell
dalzel@amazon.com
AWS Center for Quantum Computing
Pasadena, California, USA
California Institute of Technology
Pasadena, California, USA

Nicola Pancotti
nicola.pancotti@gmail.com
AWS Center for Quantum Computing
Pasadena, California, USA
California Institute of Technology
Pasadena, California, USA

Earl T. Campbell
Riverlane
Cambridge, UK
Dept. of Physics and Astronomy, University of Sheffield
Sheffield, UK

Fernando G.S.L. Brandão
AWS Center for Quantum Computing
Pasadena, California, USA
IQIM, California Institute of Technology
Pasadena, California, USA

ABSTRACT

We present a quantum algorithm that has rigorous runtime guarantees for several families of binary optimization problems, including Quadratic Unconstrained Binary Optimization (QUBO), Ising spin glasses (p-spin model), and k-local constraint satisfaction problems (k-CSP). We show that either (a) the algorithm finds the optimal solution in time $O^*(2^{(0.5-c)n})$ for an n-independent constant c, a 2^{cn} advantage over Grover's algorithm; or (b) there are sufficiently many low-cost solutions such that classical random guessing produces a $(1-\eta)$ approximation to the optimal cost value in subexponential time for arbitrarily small choice of η. Additionally, we show that for a large fraction of random instances from the k-spin model and for any fully satisfiable or slightly frustrated k-CSP formula, statement (a) is the case. The algorithm and its analysis are largely inspired by Hastings' short-path algorithm.

CCS CONCEPTS

• **Theory of computation** → *Discrete optimization*; **Quantum computation theory**.

KEYWORDS

quantum algorithms, quantum optimization, super-Grover speedup

ACM Reference Format:
Alexander M. Dalzell, Nicola Pancotti, Earl T. Campbell, and Fernando G.S.L. Brandão. 2023. Mind the Gap: Achieving a Super-Grover Quantum Speedup by Jumping to the End. In *Proceedings of the 55th Annual ACM Symposium on Theory of Computing (STOC '23), June 20–23, 2023, Orlando, FL, USA.* ACM, New York, NY, USA, 14 pages. https://doi.org/10.1145/3564246.3585203

STOC '23, June 20–23, 2023, Orlando, FL, USA

ACM ISBN 978-1-4503-9913-5/23/06.
https://doi.org/10.1145/3564246.3585203

1 OVERVIEW

1.1 Motivation and the Search for Super-Grover Speedups

It is hoped that quantum computers will outperform classical computers at solving combinatorial optimization problems, both in theory and, eventually, in practice. A primary motivation for this expectation is the existence of Grover's quantum algorithm [29] for unstructured search: since classical algorithms for hard optimization problems often resort to some form of brute-force search, Grover's algorithm potentially offers a quadratic speedup without exploiting any particular structure in the problem.

However, the hope to leverage Grover's algorithm for a practical advantage over state-of-the-art classical algorithms faces two pitfalls. First, it is rare that an exhaustive search through all possible solutions is the best classical algorithm for a specific combinatorial optimization problem. For example, the most well-known constraint satisfaction problem, 3-SAT[1], admits a $2^{0.39n}$-time classical algorithm for instances with n binary variables [30, 36], a nearly cubic speedup over a simple exhaustive search. These algorithms are still exponential time and they typically retain elements of exhaustive search. However, they exploit structure to reduce the search space and to search more efficiently.

Second, practical implementation of quantum algorithms with asymptotic quadratic speedup on actual quantum devices will suffer constant-factor slowdowns when compared to state-of-the-art classical hardware due to slower clock speeds, error-correction overheads, and general lack of parallelizability. When these factors are considered, realistic assessments of resources needed for quantum advantage using a quadratic speedup are pessimistic, suggesting that the breakeven point where quantum overtakes classical would occur only after many days, or in some cases, many years of runtime [11, 18, 50].

The outlook for practical advantage dramatically improves as the power of the polynomial speedup becomes greater than

[1] An instance of the 3-SAT problem is the question of whether a Boolean formula in conjunctive normal form (CNF)—i.e. where each constraint is the conjunction of at most three of the n binary variables—has a solution that satisfies all constraints. Elsewhere in the paper, we refer to this problem as 3-CNF-SAT.

quadratic. Ref. [11] found that an algorithm with quartic speedup (i.e. the quantum runtime scales as $T^{1/4}$ where T is the classical runtime) would offer a much more viable path to actual quantum advantage. This motivates the question of whether the quadratic Grover speedup can be surpassed for combinatorial optimization.

In some cases, a super-quadratic speedup over exhaustive search can be realized by combining the quantum techniques of amplitude amplification [15] or quantum walk search [12, 41, 56]—both of which can be understood as generalizing Grover's algorithm—with classical techniques for exploiting problem structure such as backtracking [8, 38, 42, 45], branch-and-bound [46], nested search [19], dynamic programming [7], and Markov chain Monte Carlo methods [44, 54, 58]. Establishing that these classical techniques can be employed while retaining the quadratic quantum speedup is typically nontrivial; however, these ingredients alone can only *restore* the quadratic speedup compared to classical state-of-the-art methods, not surpass it, since the fundamentally quantum part of the algorithm generates only a Grover-like speedup.

Progress toward proving a genuine *super-Grover* speedup for combinatorial optimization is limited. The quantum adiabatic algorithm (QAA) [5, 27] is one example of an algorithm that could conceivably generate super-Grover speedups for some problems. However, its runtime is notoriously hard to study due to dependence on the minimal spectral gap of certain non-commuting Hamiltonians. Additionally, some prior work has argued that the QAA can in some cases exhibit runtime that scales super-exponentially—even worse than exhaustive search—when applied to combinatorial optimization problems [6, 57], although for random instances it is expected to scale merely exponentially [35, 39, 60]. Many other works (see, e.g., Ref. [61]), often with a focus on near-term algorithms, aim only to solve combinatorial optimization problems *approximately* rather than exactly (and typically in only polynomial time, rather than exponential time); these algorithms have the potential to be practically useful, but rigorous guarantees of quantum advantage are difficult to establish.

In this work, we focus on *exact* binary optimization problems where the goal is to find the assignment z^* that minimizes some efficiently computable, classical cost function $H(z)$, with $z \in \{+1, -1\}^n$. We desire not only the best solution z^*, but also a high degree of confidence that no better solution exists; thus, we generally expect that any algorithm that solves this problem, quantum or classical, requires exponential runtime. For exponential-time algorithms, we follow the convention of writing $O^*(2^{an})$ to mean that there is an upper bound on the runtime of the form $h(n)2^{an}$ that holds for sufficiently large n, where h is a polynomial function. Exhaustive enumeration runs in (classical) time $O^*(2^n)$. Meanwhile, Grover's algorithm runs in (quantum) time $O^*(2^{0.5n})$. We seek a "super-Grover speedup," that is, a quantum algorithm that runs in time $O^*(2^{(0.5-c)n})$ for some n-independent constant c, where the additional advantage over Grover's algorithm is generated by a fundamentally quantum mechanism.

Obtaining speedups over Grover's algorithm was also the goal of a series of works by Hastings [32–34] on an algorithm called the *short-path algorithm*. Whereas Grover's algorithm begins in the equal superposition state $|+\rangle \equiv 2^{-n/2} \sum_j |j\rangle$ and measures the classical function $H(z)$ whilst performing amplitude amplification to boost the probability of obtaining an optimal-cost measurement outcome, the short-path algorithm instead uses quantum phase estimation to measure $H + V$, where in this context H denotes the diagonal operator corresponding to the cost function and V is an off-diagonal non-commuting perturbation term. Here again, amplitude amplification is used to boost the probability of ending in the ground state of $H + V$. After the ground state is prepared, a computational basis measurement is performed to find the optimal solution. It was shown that for certain choice of V and under certain assumptions on the spectral density of H at low energies (i.e. low cost values), a super-Grover speedup could be accomplished. However, no concrete scenario was given where the spectral density assumption was proved to hold. In follow-up work [33], Hastings showed that for a specific family of cost functions with 2-local Ising-like terms, the spectral density assumption could be dropped, but in this case the super-Grover speedup constant c was not n-independent, decaying like $1/\log(n)$ or faster (depending on the amount of frustration in the cost function).

1.2 A Simple Quantum Algorithm for Exact Combinatorial Optimization

Our main contribution is a simple algorithm inspired by Hastings' short-path algorithm [32] for which we can prove a super-Grover speedup for a few specific families of cost functions, including Ising spin glasses and k-local constraint satisfaction problems. The algorithm can be run on other families of cost functions and may very well have a super-Grover speedup more generally, but we present rigorous guarantees only for these specific cases.

The algorithm can be understood as an implementation of the QAA for combinatorial optimization problems with two crucial modifications, as we now explain. For simplicity, suppose here that z^* is *unique*, i.e. $H(z^*) < H(z)$ for all $z \neq z^*$, and suppose that the optimal value $E^* = H(z^*)$ is known ahead of time (these assumptions can be dropped, as explained later). By convention, we assume that H is offset such that the average cost of a uniformly random input is zero, and thus $E^* < 0$. Let $X = \sum_{i=1}^n X_i$ be the transverse-field operator, where X_i denotes the Pauli-X operator on qubit i. Then, the QAA evolves by the Hamiltonian

$$H_b = -\frac{X}{n} + b\,\frac{H}{|E^*|} \qquad \text{(for the QAA)} \tag{1}$$

while the parameter b is slowly tuned from $b = 0$, where the ground state of H_b is $|+\rangle \equiv |+\rangle^{\otimes n}$, to $b = \infty$, where the ground state of H_b is $|z^*\rangle$. The first modification is that $\frac{H}{|E^*|}$ is replaced by $g(\frac{H}{|E^*|})$ for a certain piecewise linear function $g : [-1, \infty) \to [-1, 0]$:

$$H_b = -\frac{X}{n} + b\, g\left(\frac{H}{|E^*|}\right) \qquad \text{(for our algorithm)} \tag{2}$$

This modification allows us to control the spectral properties of the cost function to enable proof of our claims. The second modification is, rather than evolve continuously through values of b for which the spectral gap of H_b is small and unknown, the algorithm simply *jumps*, first from $b = 0$ to an n-independent value of $b > 0$ where the gap is guaranteed to be large, and then from that value of b all the way to the end of the algorithm ($b = \infty$). These jumps are accomplished with quantum phase estimation along with amplitude amplification to boost the success probability of projecting onto

the ground state of H_b. The first jump is small in the sense that the success probability is nearly one (and little or no amplitude amplification is required). The second jump is large in the sense that the success probability is exponentially small. We will be able to show that (for some specific families of cost functions) the success probability of the second jump is larger than $2^{-(1-2c)n}$, and hence, after amplitude amplification, the runtime of the algorithm is less than $O^*(2^{(0.5-c)n})$, where c is a constant independent of n.

Both modifications described above, as well as the analysis of the algorithm, are inspired by (and in some cases closely follow) Hastings' short-path algorithm [32]. However, the algorithms are not identical, and in some sense they are dual to each other: where our algorithm makes a small jump and then a large jump, the short-path algorithm instead makes a large jump and then a small jump. Moreover, where our algorithm makes the modification $H/|E^*| \to g\,(H/|E^*|)$ for a piecewise linear function g (see Eqs. (1) and (2)), the short-path algorithm makes the modification $-X/n \to g\,(-X/n)$ for the function $g(x) = x^K$, with K an odd integer.

By taking our approach, we can prove more concrete results than was possible to show for the original short-path algorithm. This is true for subtle technical reasons. Firstly, as in Ref. [32], our proof utilizes the log-Sobolev inequality, which relates the expectation value of the X/n operator for a given state $|\phi\rangle$ to the entropy of the distribution over measurement outcomes that arises when $|\phi\rangle$ is measured in the computational basis. This step goes through more cleanly for us because the X/n operator appears directly in our Hamiltonian (see Eq. (2)), whereas $(X/n)^K$ appeared in Hastings' Hamiltonian, requiring some work to argue that the expectation of $(X/n)^K$ could be approximated by the Kth power of the expectation of X/n. Secondly, by switching from Hastings' strategy of a large-then-small jump to our strategy of a small-then-large jump, (roughly speaking) Hastings' spectral density assumption switches from a statement about the number of states at low energy to a statement about the number of states at high energy, which is much easier to show using straightforward tail bounds. This switch does come at a cost: we lose Hastings' perturbation theory argument that proves the exponential advantage over Grover, assuming the spectral density assumption. We are able to replace this part of the argument in our case with a proof that utilizes an approximate ground state projector.

1.3 Overview of Provable Statements

Consider the following families of optimization problems over inputs in the set $\{+1, -1\}^n$.

- MAX-Ek-LIN2: $H(z) = p(z_1, \ldots, z_n)$, where p is a polynomial consisting only of monomials of degree k.
- Quadratic Unconstrained Binary Optimization (QUBO): $H(z) = p(z_1, \ldots, z_n)$, where p is a polynomial consisting only of monomials of degree-1 or degree-2.
- MAX-k-CSP with limited frustration: $H(z) = \sum_{j=1}^m C_j$ where each C_j is a k-constraint, defined by the criteria that (i) it is a function of at most k of the n bits of the input z, and (ii) it takes the value -1 on s_j ("satisfying") assignments to those bits and value $s_j/(2^k - s_j)$ on the other $2^k - s_j$ ("unsatisfying") assignments, such that the average across all 2^k assignments is zero. Let E^* be the optimal value of H. The magnitude of the super-Grover speedup will depend on $|E^*|/m$ and is maximal when $|E^*|/m = 1$, i.e. when the instance is frustration free (all constraints are simultaneously satisfiable).

We also study a random ensemble of MAX-Ek-LIN2 instances known as the "p-spin" model [24], which here we refer to as the "k-spin" model so that the symbol k consistently represents the locality of the terms of H throughout the paper. An instance of the k-spin model is given by

$$H(z_1, \ldots, z_n) = \sqrt{\frac{k!}{n^{k-1}}} \sum_{1 \le i_1 < \ldots < i_k \le n} J_{i_1,\ldots,i_k} z_{i_1} z_{i_2} \ldots z_{i_k}\,, \qquad (3)$$

where the weights $J_{i_1,\ldots,i_k}$ are each chosen independently at random from a standard Gaussian distribution with mean 0 and variance 1. The k-spin model is used in physics to model spin glasses. When $k = 2$ it is identical to the Sherrington-Kirkpatrick (SK) model [53].

Before we state our main results, we define our usage of big-O, big-Ω notation, which appears in various places throughout the paper. Let h be a function of n, k, and $\frac{m}{|E^*|}$. Then a quantity is said to be $O(h(\cdot))$ (resp. $\Omega(h(\cdot))$) if it is upper bounded (resp. lower bounded) by some constant times $h(\cdot)$. We typically think of k as a constant as n grows, but we include the k-dependence in our expressions (rather than absorbing it into big-O) to communicate how the size of the speedup depends on k.

Our main results are the following:

- MAX-k-CSP admits a super-Grover speedup when $|E^*|/m$ is independent of n. More precisely, there exists a constant $c = \Omega\left(\frac{|E^*|^3}{k^3 2^{3k} m^3}\right)$ such that MAX-k-CSP admits a quantum algorithm with runtime $O^*(2^{(0.5-c)n})$. The algorithm can thus solve the question of whether or not a k-CSP instance is fully satisfiable in $O^*(2^{(0.5-\Omega(1/k^3 2^{3k}))n})$ time.
- For each instance of MAX-Ek-LIN2 and QUBO, either there is a quantum algorithm with super-Grover speedup, or there is a classical algorithm that can achieve an arbitrarily good approximation ratio in sub-exponential time. More precisely, for any $\eta \in [0, 1]$ and any $\gamma \in [0, 1]$, either (a) the quantum algorithm has runtime $2^{0.5(1-\Omega(\gamma\eta/k))n}$, or (b) the classical algorithm that repeatedly samples assignments uniformly at random produces a bit string y with cost $H(y) \le (1-\eta)E^*$ within time $O^*(2^{\gamma n})$ (with high probability). Thus, if there is no n-independent choice of η, γ for which (a) holds, then for arbitrarily small η, the runtime of the classical algorithm is better than $O^*(2^{\gamma n})$ for arbitrarily small γ (i.e. sub-exponential time).
- For the k-spin model, there is a choice of $\eta = \Omega(1)$ and $\gamma = \Omega(1/k^2)$ such that case (a) is satisfied in the previous bullet for all but at most a $e^{-\Omega(n/k)}$ fraction of instances.

These results are stated in more formal language later in Sec. 3.4. Note that for random MAX-k-CNF-SAT, the random ensemble of MAX-k-CSP instances where each clause is the conjunction of k distinct randomly chosen variables (or their negations), the amount of frustration increases with the number of clauses. If $m \le \alpha_c n$ for some critical value α_c (which grows like 2^k), then most instances are fully satisfiable [2], which means that $|E^*|/m = 1$. If $m = \alpha n$ for $\alpha > \alpha_c$, then most instances are not fully satisfiable

("frustrated"), but for typical instances the ratio $|E^*|/m$ is larger than some n-independent number [1]. Thus, the super-Grover speedup persists despite the frustration, although the constant c indicating the size of the 2^{cn} advantage over Grover is reduced compared to the frustration-free case. If $m/n \to \infty$ as $n \to \infty$, then $|E^*|/m \to 0$ for typical instances in that limit (in particular, it is expected to decay as $\sqrt{n/m}$) [20], and our method fails to prove a super-Grover speedup.

1.4 Comparison to Classical Algorithms and Significance

In Tab. 1, we present several problems for which we can make concrete statements about the runtime of our algorithm, and we compare with the corresponding best known classical runtime. While our algorithm has a super-*Grover* speedup for the 3-CNF-SAT problem (i.e. the question of whether there is a fully satisfying assignment for a 3-CSP instance where each clause is in conjunctive normal form), it does not have a super-*quadratic* speedup—in fact, it has no speedup at all—owing to the fact that there are also classical algorithms with a significant speedup over exhaustive enumeration. At large k, our algorithm does give a speedup for k-CNF-SAT, but the speedup is sub-quadratic. However, it is important to note that these classical algorithms for k-CNF-SAT are the product of decades of incremental improvements: for $k = 3$, Monien and Speckenmeyer found a $O^*(2^{0.70n})$-time algorithm in 1985 [43] and the coefficient was subsequently reduced to 0.57, 0.45, 0.42, and 0.39 in Refs. [48], [47], [52], and [36], respectively, with further infinitesimal improvements in Refs. [51] and [30]. For the SK model, our algorithm also fails to deliver a speedup due to the existence of a classical branch-and-bound algorithm [46] which gives a significant provable advantage over exhaustive search. It is worth noting that the classical algorithms for k-CNF-SAT in Ref. [30] and for the SK model in Ref. [46] each admit a quadratic quantum speedup by applying amplitude amplification or quantum walk search techniques. The resulting quantum algorithm is the best-known quantum algorithm in both cases.

Table 1: Summary of concrete combinatorial optimization problems where we can prove an upper bound on the runtime of our algorithm, in comparison to the best known classical algorithm for the same problem. The number displayed is the coefficient of n in the exponential, i.e. if the algorithm runs in time $O^*(2^{an})$ then a appears in the table. The problem k-CNF-SAT refers to the question of whether or not a Boolean formula in conjunctive normal form with k-local constraints is fully satisfiable. The k-spin model refers to the random ensemble of MAX-Ek-LIN2 instances where every k-local term appears with a random Gaussian weight, defined in Eq. (3). The Sherrington-Kirkpatrick (SK) model corresponds to the k-spin model with $k = 2$.

Problem	Our quantum algo	Best classical algo	
3-CNF-SAT	$0.5 - (5.2 \times 10^{-7})$	0.39	[30]
k-CNF-SAT	$0.5 - \Omega(2^{-3k}k^{-3})$	$1 - \Omega(k^{-1})$	[30]
SK model	$0.5 - (2.7 \times 10^{-5})$	0.45	[46]
k-spin	$0.5 - \Omega(k^{-3})$	1	

For the k-spin model with $k \geq 3$, the branch-and-bound technique of Ref. [46] does not obviously generalize, and we do not know of a classical algorithm that has been proved to run in time $O^*(2^{(1-c)n})$ for an n-independent value c. In contrast, our algorithm runs in time $O^*(2^{(0.5-\Omega(1/k^3))n})$, a potential super-quadratic speedup. However, we believe that it is plausible that there does exist a $O^*(2^{(1-c)n})$-time classical algorithm for k-spin: a potential candidate is classical Metropolis sampling at high temperature. Evidence that this algorithm would be effective comes from prior work on the *spherical* k-spin model, a continuous variable analogue of the "Ising" k-spin model we study here. In particular, Ref. [28] showed that for the spherical k-spin model, the Langevin dynamics are rapidly mixing when the temperature is above some n-independent threshold, allowing efficient classical sampling from the Gibbs distribution. Since the k-spin model is normalized such that typical instances have an extensive optimal cost value $|E^*| = \Omega(n)$, the Boltzmann factor $e^{-\beta E^*}$ for the optimal assignment z^* at constant inverse temperature β is exponentially large in n. This fact suggests that there exists an n-independent choice of c for which only $2^{(1-c)n}$ Gibbs samples need to be drawn to find z^* with high probability. It would be interesting to extend these results from the spherical to the Ising k-spin model and formally verify that it leads to a speedup over exhaustive enumeration.

In all cases, it is apparent that the provable advantage of our algorithm over $O^*(2^{0.5n})$ is very small. Our goal has been to prove that there exists *some* constant improvement over Grover's algorithm, and we have not dedicated much effort to optimizing the proofs to maximize the constant. We are certain that the size of the provable speedup could be improved, and furthermore in Sec. 5, we give numerical evidence that the speedup over Grover is much more substantial than the proofs imply.

We emphasize that a key reason this speedup is interesting despite its small numerical size is that the speedup mechanism has no immediate classical analogue with comparable runtime guarantees upon which our algorithm simply applies a general technique like amplitude amplification. We connect the speedup mechanism to an observation about 1-norm vs. 2-norm localization, a feature shared by Hastings' short-path algorithm [33]. Namely, it is possible for a wavefunction to be localized on a single basis state when using the 2-norm, yet de-localized across many basis states when using the 1-norm, a situation with no classical analogue. See Sec. 4.1 for a more detailed discussion. With further innovations, future algorithms may be able to leverage this phenomenon for more substantial speedups.

2 ALGORITHM

2.1 Enacting Jumps From One Ground State to Another

The algorithm we present in this work is conceptually simple; it consists essentially of just two steps, each of which is a jump from the ground state of one Hamiltonian to the ground state of another Hamiltonian. The time it takes to perform each jump is related to the overlap of the two ground states and the spectral gap of the two Hamiltonians. This is the same primitive step that was used

in Hastings' short-path algorithm [32], as well as various other quantum algorithms before it (e.g. [13]).

Proposition 1 (jump from $K_1 \to K_2$, simplified). *Given two n-qubit Hamiltonians K_1 and K_2, let $|\psi_1\rangle$ be the (unique) ground state of K_1 and Π_2 be the projector onto the (possibly degenerate) ground space of K_2. Let Δ_1 and Δ_2 denote the spectral gap above the ground space for K_1 and K_2. Then there is a unitary U for which $U|\psi_1\rangle \propto \Pi_2|\psi_1\rangle$ that is enacted up to error δ by a quantum circuit consisting of $[\min(\Delta_1,\Delta_2)]^{-1}\|\Pi_2|\psi_1\rangle\|^{-1}\,\mathrm{poly}(n,\log(\delta^{-1}))$ gates, where $\|\cdot\|$ denotes the standard Euclidean norm for a vector.*

Additionally, if K_1 or K_2 is a classical Hamiltonian (i.e. diagonal in either the computational basis or the Hadamard basis, where diagonal entries can be efficiently classically computed), then the number of gates does not depend on the corresponding gap parameter Δ_1 or Δ_2, respectively.

To actually construct the unitary, we need to have knowledge of a lower bound on $\|\Pi_2|\psi_1\rangle\|$, upper bounds on the ground state energy of K_1 and K_2, and lower bounds on the excited energy of K_1 and K_2. A more complete version of the proposition that considers these factors appears in [23, App. B], along with its proof. There the unitary U is constructed through two steps. First, it uses phase estimation to produce unitary operators R_1 and R_2 that reflect about the state $|\psi_1\rangle$ and about the ground space of K_2, respectively. The gate cost of approximating R_j to error δ is $O(\Delta_j^{-1}\log(\delta^{-1}))$ calls to a so-called "block-encoding" of the Hamiltonian K_j, which typically requires just $\mathrm{poly}(n)$ gates. The exception to this statement is the case where K_j is a classical Hamiltonian. In this case, the terms of K_j commute and the energy can be measured exactly, allowing the reflection operator to be implemented exactly in $\mathrm{poly}(n)$ gates regardless of how small Δ_j is. Second, the unitary U performs fixed-point amplitude amplification [59] to produce the state $\frac{\Pi_2|\psi_1\rangle}{\|\Pi_2|\psi_1\rangle\|}$ using the reflection operators R_1 and R_2 $O(\|\Pi_2|\psi_1\rangle\|^{-1}\log(\delta^{-1}))$ times each.

2.2 Specification of Algorithm

The inputs to the algorithm are as follows:

(a) A classical cost function H on n-bit binary assignments $z \in \{+1,-1\}^n$. This may be specified, for example, by giving the coefficients of $H(z)$ when it is expanded as a polynomial in $z_1,\ldots,z_n$, where $z_i \in \{+1,-1\}$ denotes the ith bit of z. By convention we offset H so that it has no constant term, i.e. $\sum_z H(z) = 0$. In any case, we assume that for any z, $H(z)$ can be evaluated classically in $\mathrm{poly}(n)$ time.
(b) The value $E^* = \min_z H(z)$.
(c) A value for η satisfying $0 \le \eta < 1$.
(d) A value for b satisfying $0 \le b < 1$.

Note that here we assume the optimal value E^* of the cost function is known ahead of time, which may not always be the case. In [23, App. C], we discuss how this assumption can be dropped at the expense of only polynomial overheads.

We define the piecewise-linear function $g_\eta : [-1,\infty) \to [-1,0]$:

$$g_\eta(x) = \min\left(0, \frac{x+1-\eta}{\eta}\right) \tag{4}$$

We also define the Hamiltonian H_b, parameterized by $b > 0$, as

$$\begin{aligned} H_b &= -\frac{X}{n} + b\, g_\eta\left(\frac{H}{|E^*|}\right) \\ |\psi_b\rangle &= \text{ground state of } H_b \\ E_b &= \text{ground state energy of } H_b\,, \end{aligned} \tag{5}$$

where here (in a slight abuse of notation) H denotes the diagonal $2^n \times 2^n$ Hermitian operator for which $\langle z|H|z\rangle = H(z)$ and X denotes the transverse field $\sum_{i=1}^n X_i$, with X_i the Pauli-X operator on qubit i. Let $|\psi_b\rangle$ denote the ground state of H_b and E_b the ground-state energy. For illustration purposes, the spectrum of H_b for an example $n = 20$, $\eta = 0.5$ instance is shown in Fig. 1. Note that H_b is a *stoquastic* Hamiltonian, that is, when written in the computational basis, all of the off-diagonal entries of H_b are non-positive. As a consequence of being stoquastic, the ground state $|\psi_b\rangle$ can be taken to have non-negative real entries in the computational basis [17], and we assume this convention throughout.

Let Π^* denote the projector onto the (potentially degenerate) groundspace of H (spanned by computational basis states). The algorithm begins by preparing the initial state $|+\rangle \equiv |+\rangle^{\otimes n}$, the ground state of the Hamiltonian $-X/n$ (which is diagonal in the Hadamard basis). Next, the algorithm prepares $|\psi_b\rangle$, the (unique) ground state of H_b by performing a jump from Hamiltonian $-X/n$ to Hamiltonian H_b, using the unitary described in Prop. 1. Finally, the algorithm prepares $\frac{\Pi^*|\psi_b\rangle}{\|\Pi^*|\psi_b\rangle\|}$ using a second jump, from Hamiltonian H_b to the (classical) Hamiltonian $H/|E^*|$. The state $\Pi^*|\psi_b\rangle$ is a superposition of optimal solutions to the cost function H; one of these solutions can be retrieved by measurement in the computational basis. Pseudocode for the algorithm appears in Algorithm 1, where a "jump" from Hamiltonian $K_1 \to K_2$ refers to the procedure in Prop. 1.

Algorithm 1: Pseudocode for main algorithm

Input: H, E^*, η, b, which together define H_b in Eq. (5)
Output: an optimal assignment z^* for H

1 Prepare $|+\rangle \equiv |+\rangle^{\otimes n} \equiv \sum_{i=1}^{2^n}|i\rangle$, the ground state of $-\frac{X}{n}$
2 Prepare $|\psi_b\rangle$ up to exponentially small error with jump $-\frac{X}{n} \to H_b$
3 Prepare $\frac{\Pi^*|\psi_b\rangle}{\|\Pi^*|\psi_b\rangle\|}$ up to exponentially small error with jump $H_b \to \frac{H}{|E^*|}$
4 Measure in the computational basis to produce $|z^*\rangle$

2.3 Condition for Success and Overall Runtime

Note that implementing the jumps in steps 2 and 3 of the algorithm requires an upper bound on E_b, a lower bound on the excited energy of H_b, and lower bounds on the quantities $|\langle+|\psi_b\rangle|$ and $\|\Pi^*|\psi_b\rangle\|$, where here $|+\rangle \equiv |+\rangle^{\otimes n}$ (see [23, App. B] for a formal presentation of jump implementation). The bounds on $|\langle+|\psi_b\rangle|$ and $\|\Pi^*|\psi_b\rangle\|$ can be "guessed" in the sense that one can try one value, see if the algorithm succeeds (it is easy to check if the output y of the algorithm satisfies $H(y) = E^*$), and if not, repeat with a guess that is smaller by a fixed constant factor. This procedure contributes at

most poly(n) overhead compared to if these quantities were known ahead of time. On the other hand, bounds on the eigenenergies of H_b must be shown separately, and are not guaranteed to hold for every possible cost function H and choice of parameters b, η. Accordingly, we define a condition under which success can be proved (and later we prove that the condition holds in specific cases).

Condition 1 (large-excited-energy condition). *We say the Hamiltonian H_b, as defined in Eq. (5), satisfies the* large-excited-energy condition *if the ground-state of H_b is non-degenerate and all excited states have energy greater than $-1 + 1/n$.*

Observe that when $b = 0$, the Hamiltonian H_b is equal to $-X/n$ and the large-excited-energy condition is satisfied since all excited states have energy at least $-1 + 2/n$. In the situations for which we prove rigorous bounds, the large-excited-energy condition will continue to hold as b is increased, up until some n-independent threshold. For example, in Fig. 1, we plot the numerically computed eigenvalues of H_b as a function of b for an $n = 20$ instance drawn from the 3-spin ensemble, with $\eta = 0.5$. For this instance, the large-excited-energy condition persists past $b = 0.8$.

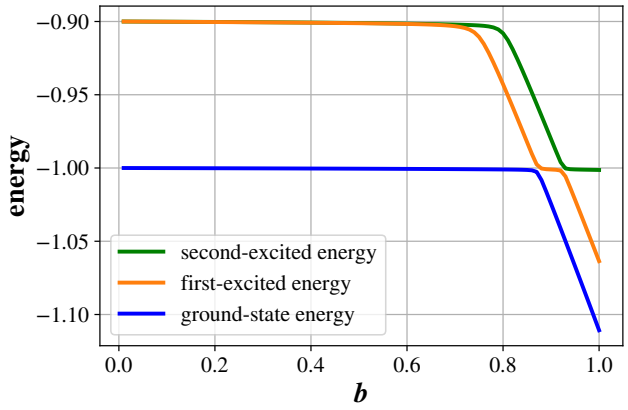

Figure 1: Plot of the lowest three eigenvalues of H_b as a function of b, for an $n = 20$ instance randomly chosen from the 3-spin ensemble, with $\eta = 0.5$. Eigenvalues were computed numerically using exact diagonalization. The two key features are that the spectral gap remains large and the ground state energy barely shifts from -1 until a relatively large value of b, namely $b \approx 0.85$.

Theorem 1 (runtime). *Fix a cost function H and parameters η and b, which defines H_b through Eq. (5). If H_b has the large excited-energy condition (Cond. 1), then with probability at least $1 - \exp(-\Omega(n))$, the algorithm outputs an optimal solution z^* of H and runs in time at most*

$$\text{poly}(n) \left(|\langle +|\psi_b\rangle|^{-1} + \|\Pi^* |\psi_b\rangle\|^{-1} \right), \tag{6}$$

where $|\psi_b\rangle$ is the ground state of H_b and Π^* is the projector onto the ground space of H.

Proof. Note that since $g_\eta(H/|E^*|)$ is a negative semidefinite operator, we have $\langle +| H_b |+\rangle \leq -1$ and thus, by the variational principle, the ground state energy E_b of H_b satisfies $E_b \leq -1$. The large-excited-energy condition then implies that $\Delta = 1/n$ is a lower bound on the spectral gap of H_b. We now refer to the steps described in the pseudocode of Algorithm 1. The only nontrivial steps are steps 2 and 3. Step 2 performs the jump $-X/n \to H_b$. Note that $-X/n$ is a classical Hamiltonian as it is diagonal in the Hadamard basis with efficiently computable entries. Thus, by Prop. 1, we may choose $\delta = e^{-\Omega(n)}$ and assert that step 2 prepares $|\psi_b\rangle$ up to error δ (in standard Euclidean vector norm) and runs in time poly$(n)\Delta^{-1}|\langle +|\psi_b\rangle|^{-1}$. Similarly, step 3 performs the jump $H_b \to H/|E^*|$. Noting that $H/|E^*|$ is also a classical Hamiltonian, and that Π^* is its ground-space projector, Prop. 1 implies that it prepares $\Pi^* |\psi_b\rangle / \|\Pi^* |\psi_b\rangle\|$ up to error at most 2δ and runs in time poly$(n)\Delta^{-1}\|\Pi^* |\psi_b\rangle\|^{-1}$. This is true even if the spectral gap of $H/|E^*|$ is exponentially small. A subsequent computational basis measurement produces an optimal solution z^* with probability at least $1 - 2\delta = 1 - \exp(-\Omega(n))$. As Δ^{-1} is poly(n), the theorem statement follows. □

Theorem 1 shows that the large-excited-energy condition is sufficient for algorithmic success and a bound on its runtime. However, the large-excited-energy condition is not a necessary condition: a similar statement would follow for a relaxed version of the condition, where the excited energy is at least $-1+1/\text{poly}(n)$. Additionally, the algorithm could still succeed even if the excited energy falls beneath -1, as long as there is a sizable gap between the ground and excited energy and a good approximation to ground/excited energy. We focus on the large-excited-energy condition because we will be able to prove that it holds under certain circumstances.

3 PROVING SPEEDUP OVER GROVER

The runtime statement in Theorem 1 implies that a super-Grover speedup can be shown given sufficient control of the spectrum of H_b (and in particular the first-excited energy) in combination with an upper bound on the quantity $|\langle +|\psi_b\rangle|^{-1} + \|\Pi^* |\psi_b\rangle\|^{-1}$, which is essentially equivalent to a lower bound on the overlaps $\langle +|\psi_b\rangle$ and $\|\Pi^* |\psi_b\rangle\|$. These tasks are accomplished separately, and in this section, we illustrate all the technical steps involved. Most of the proofs appear in [23, App. A]. In Sec. 3.1 we define additional conditions needed to organize our technical results. In Sec. 3.2, we show how, when these conditions are met, the runtime enjoys a super-Grover speedup. In Sec. 3.3, we show how a tail bound on the spectral density of the cost function implies that all of the conditions are met. We also discuss when such a tail bound is guaranteed to hold, and what can be said in the case there is no such tail bound.

3.1 Additional Conditions and Properties

In addition to the large-excited-energy condition, we define the small-ground-energy-shift condition and the α-subdepolarizing property, which will be needed to bound the runtime of the algorithm.

Condition 2 (small-ground-energy-shift condition). *We say that the Hamiltonian H_b, as defined in Eq. (5), satisfies the small-ground-energy-shift condition if the ground-state energy E_b of H_b satisfies $-1 - 1/n^3 \leq E_b \leq -1$.*

To get a qualitative sense of the idea behind Cond. 2, observe the remarkable flatness of the ground state energy of the 3-spin instance depicted in Fig. 1: E_b stays very close to -1 until relatively large values of b.

We also define a property of cost functions that we call the α-subdepolarizing property, which is important for establishing a lower bound on the overlap that determines the algorithm's runtime. To define α-subdepolarizing, we introduce the notation $y \sim x$ to denote that bit string $y \in \{+1, -1\}^n$ is generated from x by flipping a single bit chosen uniformly at random. First we define α-depolarizing before generalizing to α-subdepolarizing.

Definition 1 (α-depolarizing). *We say a cost function H is α-depolarizing if for every bit string x*

$$\mathop{\mathbb{E}}_{y\sim x} H(y) = (1-\alpha)H(x) \tag{7}$$

We call the property α-depolarizing because it states that flipping a single bit at random brings the energy toward zero by exactly a fixed factor $1-\alpha$, on average. We can immediately note that all MAX-Ek-LIN2 instances are α-depolarizing, due to the fact that every term has the same degree.

Proposition 2. *Any MAX-Ek-LIN2 instance has the α-depolarizing property with $\alpha = 2k/n$.*

Proof. If one of the n bits is flipped at random, then the sign of any degree-k monomial will flip with probability k/n. Thus the expectation value of the monomial is brought toward zero by a factor $1 - 2k/n$, and the monomial is $2k/n$-depolarizing. Moreover, the sum of cost functions all of which possess the α-depolarizing property is also α-depolarizing, by linearity of Eq. (7). This extends the property to all MAX-Ek-LIN2 instances. □

Although Hastings [32] did not use the same terminology, it was precisely the α-depolarizing property that led to an upper bound on the runtime of the short-path algorithm that suggested the possibility of super-Grover speedup. We now define the α-subdepolarizing property.

Definition 2 (α-subdepolarizing). *Consider a pair (H, g), where H is a cost function with optimal value $E^* < 0$ and $g : [-1, \infty) \to [-1, 0]$ is a monotonic non-decreasing, concave function that is twice-differentiable at every point where it is nonzero. Let $f(x) := -g(-x)$, so that f is monotonically non-decreasing and convex. We say that (H, g) is α-subdepolarizing if for any set of constants $0 < c_1, c_2, \ldots, c_T < 1$,*

$$\mathop{\mathbb{E}}_{y\sim x} \prod_{t=1}^{T} f\left(\frac{c_t H(y)}{E^*}\right) \geq \prod_{t=1}^{T} f\left(\frac{c_t(1-\alpha)H(x)}{E^*}\right). \tag{8}$$

The definition appears complex, but it attempts to capture the same idea as α-depolarizing, with minor relaxations that allow for MAX-k-CSP cost functions to be included. First, note that if H is α-depolarizing, then (H, g) is α-subdepolarizing for any function g satisfying the criteria in Def. 2, which includes g_η from Eq. (4) for any η (see [23, App. A] for a more complete justification). Second, note that any MAX-k-CSP instance has the property for the function g_η for any η, as stated in the following proposition, which is proved in [23, App. A].

Proposition 3. *Suppose H is a MAX-k-CSP instance with m terms and optimal value E^*. Then, for any η, (H, g_η) is α-subdepolarizing with*

$$\alpha = \frac{m}{|E^*|}\frac{k2^k}{(1-\eta)n} \tag{9}$$

In particular, if H is frustration free, i.e. fully satisfiable, then $|E^*| = m$ and $\alpha = k2^k/(1-\eta)n$.

3.2 Bounding the Runtime With an Approximate Ground-State Projector

Per Theorem 1, the runtime of the algorithm (assuming the large-excited-energy condition) depends on the quantity $|\langle +|\psi_b\rangle|^{-1} + \|\Pi^* |\psi_b\rangle\|^{-1}$. We wish to upper bound this quantity. In particular, we want to show that it is at most $2^{(0.5-c)n}$ for some constant c, implying a super-Grover speedup.

Let z^* be any optimal bit string and note that $\|\Pi^* |\psi_b\rangle\| \geq \langle z^*|\psi_b\rangle$ (recall we take the convention that all of the entries of $|\psi_b\rangle$ are positive in the computational basis). Then we have

$$\begin{aligned}\langle +|\psi_b\rangle^{-1} + \|\Pi^* |\psi_b\rangle\|^{-1} &\leq \langle +|\psi_b\rangle^{-1} + \left\langle \psi_b|z^*\right\rangle^{-1} \\ &\leq 2\left(\langle +|\psi_b\rangle \left\langle \psi_b|z^*\right\rangle\right)^{-1} \\ &= 2\left\langle +\right| \Pi_b \left|z^*\right\rangle^{-1} \end{aligned} \tag{10}$$

where $\Pi_b = |\psi_b\rangle\langle\psi_b|$ is the ground-state projector for the Hamiltonian H_b. We will replace Π_b by an *approximate ground state projector*, a tool that has been used successfully in the completely different context of proving area laws for ground states of many-body Hamiltonians [3, 9, 10]. As in the context of area laws, our approximate ground state projector will be a degree-ℓ polynomial in H_b; however, where they used Chebyshev polynomials, we need only examine the simpler polynomial

$$P_\ell := \left(\frac{H_b}{E_b}\right)^\ell. \tag{11}$$

The operator P_ℓ approximates Π_b since $|\psi_b\rangle$ is an eigenstate with eigenvalue 1, and, assuming ℓ is sufficiently large, the other eigenvalues of P_ℓ will be close to zero. We show that $\ell = \Omega(n^2)$ is sufficiently large, assuming the large-excited-energy condition.

Lemma 2. *If H_b satisfies the large-excited-energy condition (Cond. 1), then for any z and any $L \geq (\mu + 1.5\ln(2))n^2$, the following equation holds either for $\ell = L$ or $\ell = L+1$ (or both):*

$$\langle +|\psi_b\rangle\langle\psi_b|z\rangle \geq \langle +| P_\ell |z\rangle - 2^{-n/2}e^{-\mu n}. \tag{12}$$

Proof. Consider the operator $-X/n$, which is the first term of H_b, as defined in Eq. (5). Its maximum eigenvalue is 1, associated with eigenvector $|-\rangle \equiv 2^{-n/2}(|0\rangle - |1\rangle)^{\otimes n}$, and its second-largest eigenvalue is $1 - 2/n$. Meanwhile, the second term $b\, g_\eta(H/|E^*|)$ is a negative semidefinite operator. Denoting the largest eigenvalue of H_b by E_b', and the associated eigenvector by $|\psi_b'\rangle$, we can say that $E_b' \leq 1$. Additionally, we can assert that all other eigenvalues of H_b are smaller than $1 - 2/n$. To see this, suppose for contradiction

that there were two eigenvectors $|p\rangle$ and $|q\rangle$ of H_b with eigenvalue greater than $1-2/n$. Then the state $|p\rangle - \frac{\langle -|p\rangle}{\langle -|q\rangle}|q\rangle$ is orthogonal to $|-\rangle$ and has average energy larger than $1-2/n$. However, this is impossible, since the average value of the $-X/n$ term can be at most $1-2/n$ (as the state is orthogonal to $|-\rangle$) and the average energy of the $b\, g_\eta(H/|E^*|)$ term can be at most 0.

The operator P_ℓ has the same eigenvectors as H_b, and for each eigenvalue λ of H_b, $(\lambda/E_b)^\ell$ is an eigenvalue of P_ℓ. Thus, in the limit of $\ell \to \infty$, P_ℓ approaches the projector $|\psi_b\rangle\langle\psi_b|$. By assumption, $\ell \geq L \geq \nu n^2$ with $\nu = \mu + 1.5\ln(2)$. Hence, by the large-excited-energy condition, all 2^n eigenvalues of P_ℓ have magnitude at most $(1-1/n)^{\nu n^2} \leq e^{-\nu n} \leq 2^{-3n/2} e^{-\mu n}$, except for the eigenvalue associated with $|\psi_b\rangle$, which is 1, and (if $E_b' > 1-1/n$) the eigenvalue associated with $|\psi_b'\rangle$ which is $(E_b'/E_b)^\ell$. Denote these 2^n-2 eigenvalues by λ_i and associated eigenvectors by $|\lambda_i\rangle$ for $i = 1, \ldots, 2^n-2$. Thus, the quantity $\langle +| P_\ell |z\rangle$ is equal to

$$\langle +|\psi_b\rangle\langle\psi_b|z\rangle + \left(\frac{E_b'}{E_b}\right)^\ell \langle +|\psi_b'\rangle\langle\psi_b'|z\rangle + \sum_{i=1}^{2^n-2} \lambda_i \langle +|\lambda_i\rangle\langle\lambda_i|z\rangle$$

which is upper bounded by

$$\langle +|\psi_b\rangle\langle\psi_b|z\rangle + \left(\frac{E_b'}{E_b}\right)^\ell \langle +|\psi_b'\rangle\langle\psi_b'|z\rangle + (2^n-2)2^{-1.5n} e^{-\mu n} \quad (13)$$

In the case where $E_b' > 1-1/n > 0$, we have $E_b'/E_b < 0$ (since $E_b < 0$) and hence $(E_b'/E_b)^\ell \langle +|\psi_b'\rangle\langle\psi_b'|z\rangle$ is non-positive for exactly one of the choices $\ell = L$ or $\ell = L+1$. For this choice of ℓ, we have

$$\langle +|\psi_b\rangle\langle\psi_b|z\rangle \geq \langle +| P_\ell |z\rangle - 2^{-n/2} e^{-\mu n}. \quad (14)$$

In the case where $E_b' \leq 1-1/n$, the second term of Eq. (13) can be combined with the third term to arrive at the same result in Eq. (14). □

Lemma 2, together with Eq. (10), reduces the task of upper bounding the runtime of the algorithm to lower bounding the quantity $\langle +| P_\ell |z^*\rangle$ for $\ell = \Omega(n^2)$. We produce a lower bound by expanding P_ℓ as a sum of 2^ℓ terms using its definition in Eq. (11) and the definition of H_b in Eq. (5). Assuming the small-ground-energy-shift condition (Cond. 2) and that $\ell < O(n^3)$, we can say that the magnitude of the denominator $|E_b|^\ell$ is at most a constant, as $(1+O(1/n^3))^{O(n^3)} = O(1)$. Each of these 2^ℓ terms contributes a positive amount to the sum; we lower bound the sum by lower bounding each individual term under the assumption that (H, g_η) has the α-subdepolarizing property (Def. 2). The result of this calculation is captured in Lemma 3, which is proved in [23, App. A].

Lemma 3. *Given positive parameters $\eta < 1, b < 1, \alpha < (1-b)/2$, and integer ℓ, suppose that (H, g_η) has the α-subdepolarizing property (Def. 2), that $3/\alpha^2 \leq \ell < n^3$, and that H_b satisfies the small-ground-energy shift condition (Cond. 2). Define the function $F : [0,1] \to [0,1]$ as follows.*

$$F(x) = 1 - x + x\ln(x). \quad (15)$$

Let $z \in \{+1,-1\}^n$ be any assignment for which $\mathcal{E} := H(z)/|E^*|$ satisfies $\mathcal{E} \leq -(1-\eta)$. Then (noting that $e^{-1}-2e^{-2} \approx 0.0972 = \Omega(1)$), we have

$$\langle +| P_\ell |z\rangle \geq 2^{-n/2} \exp\left(\frac{b}{\alpha}\frac{|\mathcal{E}|}{\eta} F\left(\frac{1-\eta}{|\mathcal{E}|}\right)\right)(e^{-1}-2e^{-2}). \quad (16)$$

These lemmas together imply that the following conditions are sufficient for a super-Grover speedup: (i) the cost function has the α-subdepolarizing property for $\alpha = O(1/n)$, as is the case for MAX-Ek-LIN2 and MAX-k-CSP with limited frustration, and (ii) b and η are constants independent of n chosen such that H_b satisfies the large-excited-energy and small-ground-energy-shift conditions. Generally speaking, the larger b and η are, the larger the speedup that can be shown. This is formally captured in the following Theorem.

Theorem 4 (Super-Grover speedup). *Let H be a cost function on n variables with $n \geq 4$. Fix parameters η, b and a. Suppose that H has the α-subdepolarizing property (Def. 2) with $\alpha = a/n$ and $1 \leq a < n(1-b)/2$. Suppose further that H_b satisfies the large-excited-energy condition (Cond. 1) and the small-ground-energy-shift condition (Cond. 2). Then, Algorithm 1 successfully produces an optimal solution with probability $1 - e^{-\Omega(n)}$ while running in time bounded above by*

$$\mathrm{poly}(n) 2^{(0.5-c)n} \quad (17)$$

where

$$c = \frac{bF(1-\eta)}{a\eta\ln(2)} \geq \frac{b\eta}{2a\ln(2)}. \quad (18)$$

Proof. Theorem 1, together with Eq. (10) and Lemma 2 imply that for $L > (\mu+1.5)n^2$, the runtime is upper bounded by

$$\mathrm{poly}(n)(\langle +| P_\ell |z^*\rangle - 2^{-n/2}e^{-\mu n})^{-1} \quad (19)$$

for either $\ell = L$ or $\ell = L+1$, where $|z^*\rangle$ is any optimal solution to H. We choose $L = 3.5n^2$ (i.e. $\mu = 2$) and note that, by the assumptions of the lemma, $3/\alpha^2 \leq \ell < n^3$ holds for both $\ell = L$ and $\ell = L+1$. Thus all the conditions of Lemma 3 hold. As $|\mathcal{E}| = 1$ in this case, we have that the runtime is upper bounded by

$$\mathrm{poly}(n) 2^{n/2} (e^{n\frac{bF(1-\eta)}{\eta a}}(e^{-1}-2e^{-2}) - e^{-2n})^{-1}. \quad (20)$$

The e^{-2n} term will be smaller than the first term by a constant factor whenever $n \geq 4$. We can thus absorb it, as well as the $e^{-1}-2e^{-2}$ factor, into the $\mathrm{poly}(n)$ expression. This proves the theorem. Note that $F(1-\eta)/\eta \geq \eta/2$ holds for $0 \leq \eta \leq 1$. □

Our results actually say something stronger: every bit string z lying in a deep cost valley, that is, those for which $H(z) \leq (1-\eta)E^*$, will have overlap with $|\psi_b\rangle$ that is $2^{c'n}$ larger than $2^{-n/2}$ for some n-independent constant c'. Thus, if slightly suboptimal solutions are acceptable, the probability that our algorithm will find one of these bit strings upon measurement of $|\psi_b\rangle$ is also boosted by an amount $2^{c'n}$ compared to measurement of $|+\rangle$.

3.3 Tail Bound on Spectral Density Implies Conditions are Met

For a family of cost functions with the α-subdepolarizing property with $\alpha = O(1/n)$, Theorem 4 reduces the task of showing super-Grover speedup to the task of proving that the large-excited-energy and small-ground-energy-shift conditions are met. In general, we do not show that the conditions are always satisfied. Rather, we show that they are satisfied whenever there is a tail bound on the cumulative number of assignments to H beneath a certain cost (i.e. integral of spectral desnity of H), as follows.

Lemma 5. *Let H be a cost function over assignments $\{+1,-1\}^n$, and let $C(E)$ denote the number of assignments to H with cost value beneath E, i.e.*

$$C(E) := |\{z : H(z) \le E\}| . \tag{21}$$

Suppose that for a certain choice of $\eta < 1$

$$C\left((1-\eta)E^*\right) \le 2^{(1-\gamma)n} . \tag{22}$$

Suppose further that $\gamma \ge (1+4\log_2(n))/n$ (which holds for any n-independent γ and sufficiently large n). Then the large-excited energy condition (Cond. 1) and the small-ground-energy-shift condition (Cond. 2) are met for the Hamiltonian H_b (as defined in Eq. (5)) whenever

$$b \le \frac{\ln(2)\gamma}{2+\ln(2)} \approx 0.257\gamma \tag{23}$$

Lemma 5 is proved in [23, App. A]; the proof utilizes the log-Sobolev inequality and tools from statistical mechanics.

Proof sketch of Lemma 5. The basic idea is as follows. For the large-excited-energy condition to be violated, there must exist a state $|\phi\rangle$ orthogonal to $|+\rangle$ for which $\langle\phi| H_b |\phi\rangle \le -1+1/n$. Thus, referring to the definition of H_b in Eq. (5), there must be a number $\mathcal{U}$ satisfying $-1 \le \mathcal{U} \le 0$ such that the following two relations hold simultaneously:

$$\mathcal{U} = \langle\phi| g_\eta(H/|E^*|) |\phi\rangle \tag{24}$$

$$\langle\phi| X/n |\phi\rangle \ge 1 - 1/n + b\,\mathcal{U} . \tag{25}$$

Let $p(z) = |\langle z|\phi\rangle|^2$ denote the probability of obtaining z when measuring $|\phi\rangle$ in the computational basis. Viewing p as a probability distribution over assignments, Eq. (24) says that the average cost for the function $g_\eta(H/|E^*|)$ of a sample from p is equal to $\mathcal{U}$. Meanwhile, the log-Sobolev inequality [32, 49] allows us to turn the lower bound on $\langle\phi| X/n |\phi\rangle$ in Eq. (25) into a lower bound on S, the (log base-2) entropy of p, which reads:

$$S \ge n\left(1 - \ln(2)^{-1}(-b\,\mathcal{U} + 1/n)\right) . \tag{26}$$

This turns the question into a statistical mechanics problem: is there a distribution p with entropy greater than $n(1-O(|\mathcal{U}|))$, yet average cost of $g_\eta(H/|E^*|)$ equal to $\mathcal{U}$? Here, the tail bound constrains what is possible: since g_η zeroes out the cost of any assignment z for which $H(z) > (1-\eta)E^*$, there are at most $2^{(1-\gamma)n}$ assignments z that have negative cost for $g_\eta(H/|E^*|)$. In order to have a distribution with average cost $\mathcal{U} < 0$, a large portion of the probability must be concentrated on this subset of $2^{(1-\gamma)n}$ assignments, constraining the entropy. Formally, for this step we use an elementary fact from statistical mechanics: the distribution that maximizes entropy for a fixed average energy (average cost) is a Gibbs distribution, where assignments with cost E are allocated probability proportional to $e^{-\beta E}$ for some value of β, which physically corresponds to the inverse temperature. Applying these tools, we are able to show that whenever $b \le 0.257\gamma$, it is not possible for Eq. (24) and Eq. (25) to simultaneously be true, implying that the large-excited-energy condition holds.

Separately, the small-ground-energy-shift condition follows from a perturbation-theory argument. Viewing $b\,g_\eta(H/|E^*|)$ as a perturbation to $-X/n$ in the expression for H_b in Eq. (5), the magnitude of the first-order shift in energy is $|\langle+|\, b\, g_\eta(H/|E^*|)\, |+\rangle|$, which is smaller than $b2^{-\gamma n}$ (an exponentially small number), simply due to the tail bound, and the fact that only $2^{(1-\gamma)n}$ entries of $g_\eta(H/|E^*|)$ have nonzero cost. The full proof also bounds higher-order contributions. □

To get a sense of the interplay between the tail bound and the magnitude of the super-Grover speedup, suppose $H(z)$ comes from a family of cost functions with a unique optimal solution and for which exactly half of the eigenvalues are negative, so that $C(E^*) = 1$ and $C(0) = 2^{n-1}$. Thus, asymptotically speaking, the tail bound is satisfied with $\gamma = 1$ at $\eta = 0$, and with $\gamma = 0$ at $\eta = 1$. However, from Theorem 4, we see that the magnitude of the super-Grover speedup is zero if either $\eta = 0$ or if $\gamma = 0$ (since $\gamma = 0$ implies $b = 0$, by Lemma 5), so in both of these cases, these observations are not sufficient to show a super-Grover speedup. To show a super-Grover speedup, we need a nontrivial tail bound for nonzero η, γ to be true.

What happens if the tail bound condition is not satisfied for any choice of η, γ? Then, regardless how small we make η, there must be *many* assignments for which $H(z) \le (1-\eta)E^*$. There are so many assignments that a classical algorithm could produce an assignment achieving a $1-\eta$ approximation ratio by simple repetition in sub-exponential time for any constant value of η. Thus, instances with the α-subdepolarizing property can be partitioned into a set where our quantum algorithm has a super-Grover speedup, and a set that are unusually classically easy in a precise sense.

When can the tail bound be shown? For spin-glass-like cost functions consisting of random local terms, we often expect the spectral density to be roughly Gaussian, with the minimum cost assignment among all 2^n assignments lying $\Omega(\sqrt{n})$ standard deviations beneath 0. In this case, an exponential tail bound can be shown for any $\eta > 0$. As η increases, the corresponding bound γ decreases.

Proposition 4. *In each of the following situations, a tail bound of the form in Eq. (22) holds for any η and some choice of γ that depends on η but is independent of n.*

(1) If the cost function H is a MAX-k-CSP instance where each variable participates in no more than $2km/n$ of the m clauses[2], then for any η a tail bound holds with

$$\gamma = \left(\frac{|E^*|}{m}\right)^2 \frac{(1-\eta)^2}{2\ln(2)2^{2k}k^2} \tag{27}$$

(2) If the cost function H is randomly chosen from the k-spin ensemble of MAX-Ek-LIN2 instances, then for any η a tail bound holds with probability at least $1-2^{-\gamma n+1}$ over choice of instance, with

$$\gamma = \frac{(1-\eta)^2}{32\pi\ln(2)k^2} \tag{28}$$

Proof. Each item is proved separately in [23, App. A]. □

3.4 Formal Statement of Main Results

Theorem 6. *Let H be an instance of Quadratic Unconstrained Binary Optimization (QUBO) with $n \ge 12$ or MAX-Ek-LIN2 with $n \ge 6k$. Let $\gamma_0 = (1+4\log_2(n))/n$ (and note that $\gamma_0 \to 0$ as $n \to \infty$).*

[2] In the context of showing super-Grover speedup, this variable-participation restriction can be assumed for all MAX-k-CSP instances without loss of generality. See the proof of Theorem 7 for details.

For every $\gamma \in [\gamma_0, 1]$ and every $\eta \in [0, 1]$, one of the following must be true.

(a) There is a quantum algorithm running in time $O^*(2^{(0.5-c)n})$ that produces an optimal solution to H with probability at least $1 - e^{-\Omega(n)}$, where

$$c = \frac{1}{4(2+\ln(2))} \frac{\gamma\eta}{k} \approx 0.0928 \frac{\gamma\eta}{k} \tag{29}$$

with $k = 2$ in the case of QUBO.

(b) There is a classical algorithm which repeatedly samples assignments uniformly at random that outputs a solution z for which $H(z) \leq E^*(1-\eta)$ in expected time $O^*(2^{\gamma n})$.

Thus, if (b) is the case for arbitrarily small η and arbitrarily small γ, we can say that the classical algorithm can achieve an arbitrarily good approximation to the optimal cost value in sub-exponential time.

Proof. Rather than analyze QUBO directly, we employ a well-known reduction from QUBO to MAX-E2-LIN2: given a QUBO instance H on n variables, construct a MAX-E2-LIN2 instance H' on $n+1$ variables by introducing a binary variable z_0 and multiplying it with all degree-1 terms of H. Since all terms of H' have even degree, there is a $\mathbb{Z}_2$ symmetry where the value of $H'(z)$ is unchanged under a flip of all bits of z; thus, the spectrum of H' is the same as the spectrum of H, with each energy appearing twice. The ground state of H can be determined by computing a ground state z of H' and flipping the bits $z_1, \ldots, z_n$ if $z_0 = -1$. Since H and H' have the same spectrum, one will obey a tail bound of the form of Eq. (22) if and only if the other obeys the same tail bound. Thus all subsequent statements made about MAX-Ek-LIN2 will also apply to QUBO with the substitution $k = 2$.

Note that by Prop. 2, any MAX-Ek-LIN2 instance H is α-depolarizing (Def. 1) with $\alpha = 2k/n$. In [23, App. A] it is shown that this implies that for every η the pair (H, g_η) is α-subdepolarizing (Def. 2) with the same value of α. For a certain choice of η and γ, either H satisfies a tail bound of the form in Eq. (22) (case (a)), or it does not (case (b)).

In case (a), Lemma 5 implies that H_b satisfies the large-excited-energy and small-ground-energy-shift conditions when we choose $b = \gamma \ln(2)/(2+\ln(2))$. Note that if $\gamma < 1$ then $b < 0.3$. So long as $n \geq 6k$, the condition $1 \leq (\alpha n) \leq n(1-b)/2$ will hold, and we may apply Theorem 4 to prove the theorem statement.

In case (b), the fact that there are more than $2^{(1-\gamma)n}$ assignments with cost at most $(1-\eta)E^*$ among 2^n total assignments implies that the average number of samples the classical algorithm must make before finding one of these assignments is at most $2^{\gamma n}$. Assuming the algorithm terminates as soon as it finds one such assignment, and noting that each sample can be drawn in $\mathrm{poly}(n)$ time, the statement follows. □

Theorem 7. *There is a quantum algorithm which, for any instance of MAX-k-CSP, produces an optimal solution with probability at least $1 - e^{-\Omega(n)}$ while running in time $O^*(2^{(0.5-c)n})$, where*

$$c = 0.00722 \left(\frac{|E^*|}{m}\right)^3 \frac{1}{2^{3k}k^3} \tag{30}$$

with m denoting the number of clauses and E^* denoting the optimal value of the instance. In particular, the question of whether or not the instance is fully satisfiable can be answered with probability $1 - e^{-\Omega(n)}$ in time $O^*(2^{(0.5-c)n})$ setting $|E^*|/m = 1$ in the above expression for c. If additionally $k = 3$, we find that $c = 5.22 \times 10^{-7}$.

Proof. We may assume E^* is known. If it is not known, we may simply loop through all possible values that E^* could possibly take, which incurs only polynomial overhead by the following argument. Note that, in the general setting, there are at most $m \leq n^k$ clauses and for each clause C_t, the value of C_t is -1 when satisfied and $s_t/(2^k - s_t)$ when not satisfied, where $1 \leq s_t \leq 2^k - 1$ denotes the number of satisfying assignments to C_t. Thus, for every t and every assignment, the value of C_t is an integral multiple of $[(2^k-1)!]^{-1}$, and the number of distinct values E^* can possibly take is upper bounded by $(2^k-1)!m = \mathrm{poly}(n)$.

First, assume that each variable of H participates in at most $2km/n$ clauses. We fix a value of η to be specified later. By Prop. 3, the pair (H, g_η) has the α-subdepolarizing property with

$$\alpha = \frac{m}{|E^*|} \frac{k2^k}{(1-\eta)n}. \tag{31}$$

Furthermore, by Prop. 4, H obeys a tail bound of the form of Eq. (22) with

$$\gamma = \left(\frac{|E^*|}{m}\right)^2 \frac{(1-\eta)^2}{2\ln(2)2^{2k}k^2}. \tag{32}$$

By Lemma 5, this implies that the large-excited-energy and small-ground-energy-shift conditions are satisfied when we choose $b = \gamma\ln(2)/(2+\ln(2))$. By Theorem 4, and substituting our choices for b and γ above, the runtime is $O^*(2^{(0.5-c)n})$ with

$$c = \left[\frac{1}{2\ln(2)(2+\ln(2))} \frac{(1-\eta)^3 F(1-\eta)}{\eta}\right] \left(\frac{|E^*|}{m}\right)^3 \frac{1}{2^{3k}k^3}. \tag{33}$$

The expression in brackets achieves its maximum of 0.0145 at $\eta = 0.189$.

Now, suppose some non-empty set of variables participates in more than $2km/n$ clauses. There can be at most $n/2$ such variables, and we can identify them in polynomial time. We may design a more complex quantum algorithm that enumerates over all $2^{n/2}$ possible settings of these $n/2$ variables, and for each one determines the resulting MAX-k-CSP formula with m clauses on the remaining $n/2$ variables. Each of these formulas will satisfy the constraint of having each variable participate in at most $2km/(n/2)$ clauses. Thus, conditioned on each particular setting of the first $n/2$ variables, the algorithm will find an optimal assignment in $O^*(2^{(0.5-c)(n/2)})$ time, or if there is no such optimal assignment consistent with that setting, it can be made to output a fail flag. By running the exhaustive search over the $2^{n/2}$ assignments coherently and performing amplitude amplification on the event that an optimal assignment is found, the overall runtime is $O^*(2^{0.5(n/2)}2^{(0.5-c)(n/2)}) = O^*(2^{(0.5-c/2)n})$. This is why the expression for c in the theorem statement is a factor of 2 smaller than what was found by numerical optimization beneath Eq. (33). □

Theorem 8. *There is a quantum algorithm which, for at least a fraction $1 - e^{-\Omega(n/k)}$ instances drawn from the k-spin ensemble, produces an optimal solution with probability at least $1 - e^{-\Omega(n)}$ while running in time $O^*(2^{(0.5-c)n})$, where*

$$c = (2.24 \times 10^{-4}) \frac{1}{k^3} \tag{34}$$

Proof. By Prop. 2, any MAX-Ek-LIN2 instance H is α-depolarizing (Def. 1) with $\alpha = 2k/n$. In [23, App. A], it is shown that this implies that for every η the pair (H, g_η) is α-subdepolarizing (Def. 2) with the same value of α. Furthermore, by Prop. 4, H obeys a tail bound of the form of Eq. (22) with

$$\gamma = \frac{(1-\eta)^2}{32\pi \ln(2) k^2} . \tag{35}$$

By Lemma 5, this implies that the large-excited-energy and small-ground-energy-shift conditions are satisfied when we choose $b = \gamma \ln(2)/(2 + \ln(2))$. By Theorem 4, and substituting our choices for b and γ above, the runtime is $O^*(2^{(0.5-c)n})$ with

$$c = \left[\frac{1}{64 \ln(2) \pi (2 + \ln(2))} \frac{(1-\eta)^2 F(1-\eta)}{\eta} \right] \frac{1}{k^3} . \tag{36}$$

The expression in brackets achieves its maximum of 2.24×10^{-4} at $\eta = 0.405$, proving the theorem. □

4 SPEEDUP MECHANISM

In this section, we describe an inherently quantum feature of our algorithm that might be regarded as the reason for the super-Grover speedup. Then we discuss how classical algorithms might try to replicate the behavior of our algorithm.

4.1 Mechanism for Speedup: Localization in 1-Norm vs. 2-Norm

In quantum mechanics, the Born rule dictates that probabilities of measurement outcomes are the square of amplitudes of the wavefunction. In other words, if we want to measure the probability that the wavefunction lies within a certain subset of the 2^n assignments, we should use the 2-norm. This contrasts with standard probability theory, where the 1-norm would be used.

In describing the mechanism for speedup for the short-path algorithm [33], Hastings pointed to precisely this fact. In the short-path algorithm, a state $|\phi\rangle$ was prepared that was simultaneously localized in the 2-norm and de-localized in the 1-norm, in the following sense. Define

$$w_1(z) = \frac{|\langle z|\phi\rangle|}{\sum_z |\langle z|\phi\rangle|} \qquad w_2(z) = \frac{|\langle z|\phi\rangle|^2}{\sum_z |\langle z|\phi\rangle|^2} , \tag{37}$$

both of which are probability distributions over inputs z. For the state $|\phi\rangle$ prepared by the algorithm, the distribution w_2 had nearly all its mass concentrated on the optimal solution z^*, while the distribution w_1 had an exponentially small 2^{-cn} fraction of its mass concentrated on z^*. In fact, if we break the assignments up into subsets of equal Hamming distance from z^*, the subset with the most w_1 mass was order-n bit flips away from z^*. This 1-norm delocalization is what gave rise to the super-Grover speedup during the large jump made by the algorithm, while the 2-norm localization allowed the small jump to succeed with high probability.

Our algorithm exploits a similar effect to produce a super-Grover speedup, although it does so in the Hadamard basis. That is, the same comments for $|\phi\rangle$ above apply to the state $|\psi_b\rangle$ if we redefine

$$w_1(u) = \frac{|\langle u|\psi_b\rangle|}{\sum_u |\langle u|\psi_b\rangle|} \qquad w_2(u) = \frac{|\langle u|\psi_b\rangle|^2}{\sum_u |\langle u|\psi_b\rangle|^2} , \tag{38}$$

where the states $|u\rangle$ are the 2^n n-fold tensor products of $|+\rangle$ and $|-\rangle$. The state $|\psi_b\rangle$ is has high overlap with $|\mathbf{+}\rangle$ (see, e.g., Fig. 2): thus, it is 2-norm localized in the Hadamard basis. Now, without loss of generality, we assume the optimal solution is the computational basis state $|0^n\rangle = 2^{-n/2}(|+\rangle + |-\rangle)^{\otimes n}$. Then, the overlap that determines the runtime of our algorithm is given by

$$\langle 0^n|\psi_b\rangle = 2^{-n/2} \sum_{u \in \{+,-\}^n} \langle u|\psi_b\rangle . \tag{39}$$

The statement that $\langle 0^n|\psi_b\rangle \geq 2^{-(0.5-c)n}$ is thus equivalent to the statement $\sum_u \langle u|\psi_b\rangle \geq 2^{cn}$. This implies the denominator of the definition of w_1 in Eq. (38) is at least 2^{cn}, but meanwhile $\langle \mathbf{+}|\psi_b\rangle \approx 1$, meaning $w_1(\mathbf{+}) \leq 2^{-cn}$. In other words, an exponentially small amount of 1-norm weight lies on the $|\mathbf{+}\rangle$ basis state.

Crucially, this dichotomy between the distribution of weight in the 1-norm and the 2-norm is only possible in quantum mechanics; it is not obvious how one would exploit this phenomenon with a classical algorithm. As a result, this mechanism has the potential to deliver super-Grover speedups, and, potentially, super-quadratic speedups compared to best-known classical algorithms.

4.2 Performance of Analogous Classical Markov Chain Methods

While a classical algorithm cannot exactly replicate the same 1-norm vs. 2-norm dichotomy that our quantum algorithm exploits, it is valuable to scrutinize whether a classical algorithm might nonetheless emulate our quantum algorithm. We identify two categories of classical algorithms based on sampling Markov processes that might, in a certain sense, be regarded as classical analogues of our algorithm. However, in both cases, they lack rigorous runtime guarantees similar to those we have shown about our algorithm.

The first candidate analgoue is the classical algorithm mentioned in Sec. 1.4 that repeatedly samples the high-temperature classical Gibbs distribution of the cost function (or some transformation applied to the cost function), a feat which can be accomplished via classical Metropolis sampling or simulated annealing. Showing that this classical algorithm runs in time $O^*(2^{(1-c)n})$ would require proving that the associated Markov chain remains rapidly mixing up to some sufficiently large inverse temperature β. As rapid-mixing proofs often leverage the log-Sobolev inequality and its relatives (e.g., [25, 28, 55]), perhaps it is unsurprising that the log-Sobolev inequality is also helpful for us in showing lower bounds on the spectral gap of the Hamiltonian H_b. However, despite a heuristic connection between the inverse temperature β in the Metropolis sampling algorithm and the inverse transverse-field strength parameter b in our algorithm, there is no direct relationship between the runtime of the two algorithms. Indeed, a clearer quantum analogue of simulated annealing is given by applying the "quantum simulated annealing" algorithm proposed in Ref. [54] (see also Refs. [4, 14, 40, 50, 58]), which, when combined with amplitude amplification as in Ref. [44], gives a direct quadratic (but no larger) quantum speedup of the classical algorithm described above.

The second candidate analogue of our quantum algorithm is formed by applying a Quantum Monte Carlo (QMC) approach to the Hamiltonian H_b (see, e.g., [16, 21, 22, 26, 37]). A reason to think that this might be successful is that our Hamiltonian H_b is

stoquastic and thus it does not have the sign problem [17]. Indeed, the core step of our algorithm is the application of quantum phase estimation to the stoquastic Hamiltonian H_b, a situation that was shown in Ref. [16] to admit a classical QMC simulation whenever there exists a "guiding state." In the case of H_b, there is no obvious state that satisfies the criteria of Ref. [16]: the state $|+\rangle$ falls short since it differs from $|\psi_b\rangle$ by an exponentially large factor on some of the amplitudes. More generally, showing that the QMC algorithm can sample (up to inverse exponential precision) the same distribution as computational basis measurements on the ground state of H_b requires proving that a certain Markov chain representing the imaginary time evolution $e^{-\beta H_b}$ remains rapidly mixing up to imaginary time β of order-n^2 (this is necessary to guarantee that the second largest eigenvalue of $e^{-\beta H_b}$ is exponentially smaller than the largest eigenvalue). For comparison, Ref. [22] showed rapid mixing for transverse-field Ising models up to imaginary time of order $(J(\Delta+2))^{-1}$, where J is the maximum interaction strength and Δ is the interaction-degree. This encompasses the SK model (i.e. Eq. (3) with $k=2$), where, after normalization by $|E^*|$ (order-n), we have $J^{-1} = O(n\sqrt{n})$ and $\Delta^{-1} = 1/n$, so this fails to prove a relevant statement for our situation. However, it remains an interesting question to examine the performance of QMC as a classical simulation of our algorithm. Toward that end, it is interesting to note that previous work (see fourth counterexample of Ref. [31] and Ref. [37]) has demonstrated examples where the classical QMC algorithm struggles to simulate its quantum counterpart due in part to a discrepancy between wavefunction localization in the 1-norm versus the 2-norm, the exact effect exploited by our quantum algorithm (albeit in a different basis).

5 NUMERICAL RESULTS

5.1 Example Instance

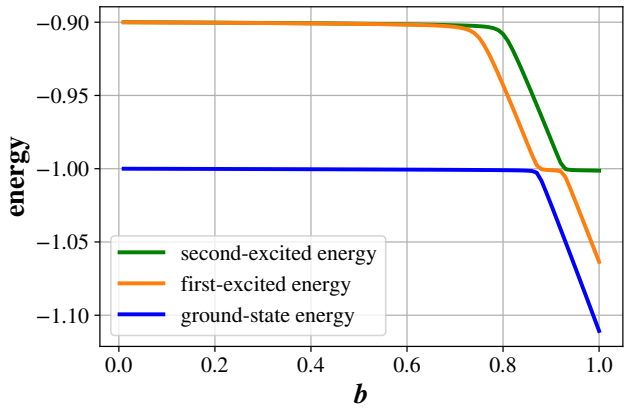

Figure 2: Plot of the relevant overlap values for the same $n = 20$, $\eta = 0.5$ 3-spin instance from Fig. 1. Overlaps are determined by exact diagonalization of H_b. The overlap $|\langle + | \psi_b \rangle|$ (blue) determines the runtime of step 2 of Algorithm 1, and the overlap $|\langle z^* | \psi_b \rangle|$ (red) determines the runtime of step 3.

We present an example instance to confirm that the theoretical picture of the algorithm holds in practice. Recall that in Fig. 1, we presented the lowest three energy levels of H_b as a function of b for an $n = 20$ instance chosen from the 3-spin ensemble, with $\eta = 0.5$. In Fig. 2, we show the overlap profile for the same instance, that is $|\langle + | \psi_b \rangle|$ and $|\langle z^* | \psi_b \rangle|$ as a function of b.

Our theoretical picture is largely confirmed. The ground state energy stays remarkably unshifted from -1 and the spectral gap remains open until b is quite large, around $b = 0.85$ for this instance. Meanwhile, the overlap $|\langle z^* | \psi_b \rangle|$ which determines the runtime of the algorithm grows exponentially with b, even as the overlap $|\langle + | \psi_b \rangle|$ remains close to 1 (the wavefunction remains localized in the 2-norm even as it becomes increasingly de-localized in the 1-norm, see Sec. 4.1). This confirms that for this instance the first step of the algorithm will succeed with little need for amplification even when we take b as large as 0.8, and that the second step has success probability significantly better than Grover's algorithm.

This instance also illustrates how there might be considerable room for improvement in the theoretical analysis. Firstly, in Fig. 2, the quantity $|\langle z^* | \psi_b \rangle|$ appears to grow *super*-exponentially with b, even as $|\langle + | \psi_b \rangle|$ remains close to 1, whereas the theoretical lower bound from Lemma 2 predicts only exponential growth. Secondly, the numerics indicate that the algorithm will work well for this $\eta = 0.5$ instance all the way up to $b = 0.8$, whereas the theoretical analysis only guarantees success up to a much smaller value of b. Looking back at Eq. (35) and the text beneath, for $k = 3$ and $\eta = 0.5$, the large-excited-energy condition (Cond. 1) is shown to hold only for $b \leq 1.02 \times 10^{-4}$ (and only for *most* random instances in the ensemble). From the theory perspective, the important conclusion is that this value is independent of n, but these numerics suggest that there is a considerable gap between the theoretical bounds and the empirical reality.

5.2 Estimation of Actual Size of Super-Grover Speedup

Our theoretical analysis shows a super-Grover speedup but with a very small lower bound on the size of the speedup. We illustrate numerically that the actual super-Grover speedup might be much more substantial. By performing exact diagonalization on 30 random $\eta = 0.5$, $b = 0.7$ instances from the 3-spin ensemble at each value $n = 17, 18, \ldots, 23$, we determine the growth of the advantage with n. We chose $b = 0.7$ to give some breathing room between the maximum value at which the large-excited-energy condition (Cond. 1) failed (roughly $b = 0.8$) for the instance depicted in Fig. 1. Indeed, we found that for all of the instances we diagonalized, $b = 0.7$ fell comfortably below the avoided level crossing between the ground-state and first-excited energies, and that the large-excited-energy condition was satisfied. Thus, by Theorem 1, the algorithm would succeed with runtime dependent on $|\langle z^* | \psi_{0.7} \rangle|^{-1}$. In Fig. 3, we plot the quantity $|\langle z^* | \psi_{0.7} \rangle|^{-1}$ for each instance, as well as the median among all 30 instances at each value of n. Fitting the medians to a line on a log scale, we find that the best fit is $|\langle z^* | \psi_{0.7} \rangle|^{-1} \approx 0.28 \times 2^{0.427n}$; the 95% confidence interval on the 0.427 value is $[0.415, 0.439]$. While the quality of the numerical fit is encouraging, we caution that robust conclusions are difficult to assert based on an exponential fit to just 7 data points that span the small range from roughly 40 to roughly 300, less than even one order of magnitude.

A scaling of $O^*(2^{0.43n})$ would represent a material improvement over Grover, if not a practically transformative one. A larger speedup, closer to quartic compared with exhaustive enumeration (corresponding to a quantum algorithm with $2^{0.25n}$ runtime), would be needed to make the prospect of actual quantum advantage more likely [11]. However, as we have not made much effort at optimizing the parameters b and η, we are optimistic that future examination will reveal additional speedup over Grover than what we report here.

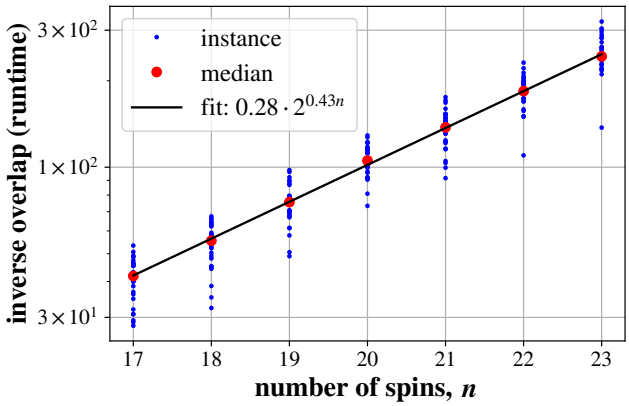

Figure 3: Plot of $|\langle z^*|\psi_b\rangle|^{-1}$ for several randomly chosen instances from the 3-spin model at $b = 0.7$ and $\eta = 0.5$ at values of n ranging from $n = 17$ to $n = 23$. For each value of n, the median value is also plotted. A fit of the medians to an exponential yields the fit $|\langle z^*|\psi_b\rangle|^{-1} = 0.276 \cdot 2^{0.427n}$.

6 CONCLUSION

We present a quantum algorithm that has a rigorous proof of a speedup compared to Grover's algorithm in certain cases. On the one hand, as it stands, the algorithm is not likely to lead to a practical advantage. For one, the size of the speedup in our proof is very small, although we believe that further optimization of parameters and extensions of the theory would lead to improvements. Additionally, while numerical experiments suggest that the actual speedup is non-negligible, it still appears to fall short of a cubic or quartic speedup (in comparison to exhaustive enumeration), and thus, after accounting for other overheads, it is unlikely to lead to practical benefit without further improvements [11]. Finally, in most cases where our algorithm has rigorous guarantees, the speedup over the *best* classical algorithm is less than quadratic, owing to the existence of clever classical algorithms that exploit structure to beat exhaustive enumeration.

On the other hand, the mechanism by which our algorithm generates a speedup has no obvious classical analogue. This contrasts with many other speedups over exhaustive enumeration, which are essentially a Grover speedup combined with a classical technique. The inherent quantumness of our speedup positions our algorithm to at least have a fighting chance at achieving a super-quadratic speedup, even if it cannot yet do so when compared to mature classical algorithms that have been improved over the course of decades. Toward that end, an important open question is whether our algorithm and the speedup mechanism behind it can be combined with any of the classical techniques for exploiting problem structure, such as backtracking [45] and branch-and-bound [46].

ACKNOWLEDGMENTS

We are grateful to Mario Berta, András Gilyén, Michael Kastoryano, Sam McArdle, Ashley Montanaro, and Grant Salton for helpful conversations. We thank Robbie King for discovering an inaccuracy in the statement of Lemma 2 that appeared in an earlier version.

REFERENCES

[1] Dimitris Achlioptas, Assaf Naor, and Yuval Peres. 2007. On the Maximum Satisfiability of Random Formulas. *J. ACM* 54 (2007), 10–es. https://doi.org/10.1145/1219092.1219098 arXiv:math/0305151

[2] Dimitris Achlioptas and Yuval Peres. 2003. The Threshold for Random k-SAT is $2^k(\ln 2 - O(k))$. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing* (San Diego, CA, USA). Association for Computing Machinery, New York, NY, USA, 223–231. https://doi.org/10.1145/780542.780577 arXiv:cs/0305009

[3] Dorit Aharonov, Itai Arad, Zeph Landau, and Umesh Vazirani. 2011. The 1D Area Law and the Complexity of Quantum States: A Combinatorial Approach. In *2011 IEEE 52nd Annual Symposium on Foundations of Computer Science*. 324–333. https://doi.org/10.1109/FOCS.2011.91

[4] Dorit Aharonov and Amnon Ta-Shma. 2003. Adiabatic Quantum State Generation and Statistical Zero Knowledge. In *Proceedings of the Thirty-Fifth Annual ACM Symposium on Theory of Computing* (San Diego, CA, USA). Association for Computing Machinery, New York, NY, USA, 20–29. https://doi.org/10.1145/780542.780546 arXiv:quant-ph/0301023

[5] Tameem Albash and Daniel A. Lidar. 2018. Adiabatic quantum computation. *Rev. Mod. Phys.* 90 (2018), 015002. Issue 1. https://doi.org/10.1103/RevModPhys.90.015002 arXiv:1611.04471

[6] Boris Altshuler, Hari Krovi, and Jérémie Roland. 2010. Anderson localization makes adiabatic quantum optimization fail. *Proceedings of the National Academy of Sciences* 107 (2010), 12446–12450. https://doi.org/10.1073/pnas.1002116107 arXiv:0912.0746

[7] Andris Ambainis, Kaspars Balodis, Jānis Iraids, Martins Kokainis, Krišjānis Prūsis, and Jevgēnijs Vihrovs. 2019. Quantum speedups for exponential-time dynamic programming algorithms. In *Proceedings of the Thirtieth Annual ACM-SIAM Symposium on Discrete Algorithms*. SIAM, 1783–1793. https://doi.org/10.1137/1.9781611975482.107 arXiv:1807.05209

[8] Andris Ambainis and Martins Kokainis. 2017. Quantum Algorithm for Tree Size Estimation, with Applications to Backtracking and 2-Player Games. In *Proceedings of the 49th Annual ACM SIGACT Symposium on Theory of Computing* (Montreal, Canada). Association for Computing Machinery, New York, NY, USA, 989–1002. https://doi.org/10.1145/3055399.3055444 arXiv:1704.06774

[9] Itai Arad, Zeph Landau, and Umesh Vazirani. 2012. Improved one-dimensional area law for frustration-free systems. *Phys. Rev. B* 85 (2012), 195145. Issue 19. https://doi.org/10.1103/PhysRevB.85.195145 arXiv:1111.2970

[10] Itai Arad, Zeph Landau, Umesh Vazirani, and Thomas Vidick. 2017. Rigorous RG algorithms and area laws for low energy eigenstates in 1D. *Communications in Mathematical Physics* 356 (2017), 65–105. https://doi.org/10.1007/s00220-017-2973-z arXiv:1602.08828

[11] Ryan Babbush, Jarrod R. McClean, Michael Newman, Craig Gidney, Sergio Boixo, and Hartmut Neven. 2021. Focus beyond Quadratic Speedups for Error-Corrected Quantum Advantage. *PRX Quantum* 2 (2021), 010103. Issue 1. https://doi.org/10.1103/PRXQuantum.2.010103 arXiv:2011.04149

[12] Aleksandrs Belovs. 2013. Quantum walks and electric networks. (2013). https://doi.org/10.48550/arXiv.1302.3143 arXiv:1302.3143

[13] Sergio Boixo, Emanuel Knill, and Rolando D Somma. 2010. Fast quantum algorithms for traversing paths of eigenstates. (2010). https://doi.org/10.48550/arXiv.1005.3034 arXiv:1005.3034

[14] Sergio Boixo, Gerardo Ortiz, and Rolando Somma. 2015. Fast quantum methods for optimization. *The European Physical Journal Special Topics* 224 (2015), 35–49. https://doi.org/10.1140/epjst/e2015-02341-5 arXiv:1409.2477

[15] Gilles Brassard, Peter Høyer, Michele Mosca, and Alain Tapp. 2002. Quantum Amplitude Amplification and Estimation. In *Quantum Computation and Quantum Information: A Millennium Volume*. Contemporary Mathematics Series, Vol. 305. AMS, 53–74. https://doi.org/10.1090/conm/305/05215 arXiv:quant-ph/0005055

[16] Sergey Bravyi. 2015. Monte Carlo Simulation of Stoquastic Hamiltonians. *Quantum Information & Computation* 15 (2015), 1122–1140. https://doi.org/10.26421/QIC15.13-14-3 arXiv:1402.2295

[17] Sergey Bravyi, David P Divincenzo, Roberto I Oliveira, and Barbara M Terhal. 2008. The complexity of stoquastic local Hamiltonian problems. *Quantum Information & Computation* 8 (2008), 0361–0385. https://doi.org/10.26421/QIC8.5-1 arXiv:quant-ph/0606140

[18] Earl Campbell, Ankur Khurana, and Ashley Montanaro. 2019. Applying quantum algorithms to constraint satisfaction problems. *Quantum* 3 (2019), 167. https://doi.org/10.22331/q-2019-07-18-167 arXiv:1810.05582

[19] Nicolas J. Cerf, Lov K. Grover, and Colin P. Williams. 2000. Nested quantum search and structured problems. *Phys. Rev. A* 61 (2000), 032303. Issue 3. https://doi.org/10.1103/PhysRevA.61.032303

[20] Don Coppersmith, David Gamarnik, Mohammad Hajiaghayi, and Gregory B. Sorkin. 2003. Random MAX SAT, Random MAX CUT, and Their Phase Transitions. In *Proceedings of the Fourteenth Annual ACM-SIAM Symposium on Discrete Algorithms* (Baltimore, Maryland). Society for Industrial and Applied Mathematics, USA, 364–373. arXiv:math/0306047

[21] Elizabeth Crosson and Aram W. Harrow. 2016. Simulated Quantum Annealing Can Be Exponentially Faster Than Classical Simulated Annealing. In *2016 IEEE 57th Annual Symposium on Foundations of Computer Science (FOCS)*. 714–723. https://doi.org/10.1109/FOCS.2016.81 arXiv:1601.03030

[22] Elizabeth Crosson and Samuel Slezak. 2020. Classical Simulation of High Temperature Quantum Ising Models. (2020). https://doi.org/10.48550/arXiv.2002.02232 arXiv:2002.02232

[23] Alexander M. Dalzell, Nicola Pancotti, Earl T. Campbell, and Fernando G. S. L. Brandão. 2022. Mind the gap: Achieving a super-Grover quantum speedup by jumping to the end. (2022). https://doi.org/10.48550/arXiv.2212.01513 arXiv:2212.01513

[24] B. Derrida. 1980. Random-Energy Model: Limit of a Family of Disordered Models. *Phys. Rev. Lett.* 45 (1980), 79–82. Issue 2. https://doi.org/10.1103/PhysRevLett.45.79

[25] Ronen Eldan, Frederic Koehler, and Ofer Zeitouni. 2022. A spectral condition for spectral gap: fast mixing in high-temperature Ising models. *Probability Theory and Related Fields* 182 (2022), 1035–1051. https://doi.org/10.1007/s00440-021-01085-x arXiv:2007.08200

[26] Edward Farhi, Jeffrey Goldstone, David Gosset, Sam Gutmann, Harvey B Meyer, and Peter Shor. 2009. Quantum adiabatic algorithms, small gaps, and different paths. *Quantum Information & Computation* (2009). https://doi.org/10.26421/QIC11.3-4-1 arXiv:0909.4766

[27] Edward Farhi, Jeffrey Goldstone, Sam Gutmann, and Michael Sipser. 2000. Quantum computation by adiabatic evolution. (2000). https://doi.org/10.48550/arXiv.quant-ph/0001106 arXiv:quant-ph/0001106

[28] Reza Gheissari and Aukosh Jagannath. 2019. On the spectral gap of spherical spin glass dynamics. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques* 55 (2019), 756 – 776. https://doi.org/10.1214/18-AIHP897 arXiv:1608.06609

[29] Lov K Grover. 1996. A fast quantum mechanical algorithm for database search. In *Proceedings of the twenty-eighth annual ACM symposium on Theory of computing*. 212–219. https://doi.org/10.1145/237814.237866 arXiv:quant-ph/9605043

[30] Thomas Dueholm Hansen, Haim Kaplan, Or Zamir, and Uri Zwick. 2019. Faster k-SAT Algorithms Using Biased-PPSZ. In *Proceedings of the 51st Annual ACM SIGACT Symposium on Theory of Computing* (Phoenix, AZ, USA). Association for Computing Machinery, New York, NY, USA, 578–589. https://doi.org/10.1145/3313276.3316359

[31] Matthew B Hastings. 2013. Obstructions to classically simulating the quantum adiabatic algorithm. *Quantum Information & Computation* 13 (2013), 1038–1076. https://doi.org/10.26421/QIC13.11-12-8 arXiv:1302.5733

[32] M. B. Hastings. 2018. A short path quantum algorithm for exact optimization. *Quantum* 2 (2018), 78. https://doi.org/10.22331/q-2018-07-26-78 arXiv:1802.10124

[33] Matthew B Hastings. 2018. Weaker Assumptions for the Short Path Optimization Algorithm. (2018). https://doi.org/10.48550/arXiv.1807.03758 arXiv:1807.03758

[34] M. B. Hastings. 2019. The short path algorithm applied to a toy model. *Quantum* 3 (2019), 145. https://doi.org/10.22331/q-2019-05-20-145 arXiv:1901.03884

[35] Itay Hen and A. P. Young. 2011. Exponential complexity of the quantum adiabatic algorithm for certain satisfiability problems. *Phys. Rev. E* 84 (2011), 061152. Issue 6. https://doi.org/10.1103/PhysRevE.84.061152 arXiv:1109.6872

[36] Timon Hertli. 2014. 3-SAT Faster and Simpler—Unique-SAT Bounds for PPSZ Hold in General. *SIAM J. Comput.* 43 (2014), 718–729. https://doi.org/10.1137/120868177 arXiv:1103.2165

[37] Michael Jarret, Stephen P. Jordan, and Brad Lackey. 2016. Adiabatic optimization versus diffusion Monte Carlo methods. *Phys. Rev. A* 94 (2016), 042318. Issue 4. https://doi.org/10.1103/PhysRevA.94.042318 arXiv:1607.03389

[38] Michael Jarret and Kianna Wan. 2018. Improved quantum backtracking algorithms using effective resistance estimates. *Phys. Rev. A* 97 (2018), 022337. Issue 2. https://doi.org/10.1103/PhysRevA.97.022337 arXiv:1711.05295

[39] Sergey Knysh and Vadim Smelyanskiy. 2010. On the relevance of avoided crossings away from quantum critical point to the complexity of quantum adiabatic algorithm. (2010). https://doi.org/10.48550/arXiv.1005.3011 arXiv:1005.3011

[40] Jessica Lemieux, Bettina Heim, David Poulin, Krysta Svore, and Matthias Troyer. 2020. Efficient Quantum Walk Circuits for Metropolis-Hastings Algorithm. *Quantum* 4 (2020), 287. https://doi.org/10.22331/q-2020-06-29-287 arXiv:1910.01659

[41] Frederic Magniez, Ashwin Nayak, Jeremie Roland, and Miklos Santha. 2007. Search via Quantum Walk. In *Proceedings of the Thirty-Ninth Annual ACM Symposium on Theory of Computing* (San Diego, California, USA) *(STOC '07)*. Association for Computing Machinery, New York, NY, USA, 575–584. https://doi.org/10.1145/1250790.1250874 arXiv:quant-ph/0608026

[42] Simon Martiel and Maxime Remaud. 2020. Practical Implementation of a Quantum Backtracking Algorithm. In *SOFSEM 2020: Theory and Practice of Computer Science*. Springer International Publishing, Cham, 597–606. https://doi.org/10.1007/978-3-030-38919-2_49 arXiv:1908.11291

[43] B. Monien and E. Speckenmeyer. 1985. Solving satisfiability in less than 2^n steps. *Discrete Applied Mathematics* 10 (1985), 287–295. https://doi.org/10.1016/0166-218X(85)90050-2

[44] Ashley Montanaro. 2015. Quantum speedup of Monte Carlo methods. *Proceedings of the Royal Society A: Mathematical, Physical and Engineering Sciences* 471 (2015), 20150301. https://doi.org/10.1098/rspa.2015.0301 arXiv:1504.06987

[45] Ashley Montanaro. 2018. Quantum-Walk Speedup of Backtracking Algorithms. *Theory of Computing* 14 (2018), 1–24. https://doi.org/10.4086/toc.2018.v014a015 arXiv:1509.02374

[46] Ashley Montanaro. 2020. Quantum speedup of branch-and-bound algorithms. *Phys. Rev. Research* 2 (2020), 013056. Issue 1. https://doi.org/10.1103/PhysRevResearch.2.013056 arXiv:1906.10375

[47] Ramamohan Paturi, Pavel Pudlák, Michael E. Saks, and Francis Zane. 2005. An Improved Exponential-Time Algorithm for k-SAT. *J. ACM* 52 (2005), 337–364. https://doi.org/10.1145/1066100.1066101

[48] Robert Rodošek. 1996. A new approach on solving 3-satisfiability. In *International Conference on Artificial Intelligence and Symbolic Mathematical Computing*. 197–212. https://doi.org/10.1007/3-540-61732-9_59

[49] Alex Samorodnitsky. 2008. A modified logarithmic Sobolev inequality for the Hamming cube and some applications. (2008). https://doi.org/10.48550/arXiv.0807.1679 arXiv:0807.1679

[50] Yuval R. Sanders, Dominic W. Berry, Pedro C.S. Costa, Louis W. Tessler, Nathan Wiebe, Craig Gidney, Hartmut Neven, and Ryan Babbush. 2020. Compilation of Fault-Tolerant Quantum Heuristics for Combinatorial Optimization. *PRX Quantum* 1 (2020), 020312. Issue 2. https://doi.org/10.1103/PRXQuantum.1.020312 arXiv:2007.07391

[51] Dominik Scheder and John P. Steinberger. 2017. PPSZ for General k-SAT - Making Hertli's Analysis Simpler and 3-SAT Faster. In *32nd Computational Complexity Conference (CCC 2017)*, Vol. 79. Schloss Dagstuhl–Leibniz-Zentrum fuer Informatik, Dagstuhl, Germany, 9:1–9:15. https://doi.org/10.4230/LIPIcs.CCC.2017.9

[52] Uwe Schöning. 2002. A probabilistic algorithm for k-SAT based on limited local search and restart. *Algorithmica* 32 (2002), 615–623. https://doi.org/10.1007/s00453-001-0094-7

[53] David Sherrington and Scott Kirkpatrick. 1975. Solvable Model of a Spin-Glass. *Phys. Rev. Lett.* 35 (1975), 1792–1796. Issue 26. https://doi.org/10.1103/PhysRevLett.35.1792

[54] R. D. Somma, S. Boixo, H. Barnum, and E. Knill. 2008. Quantum Simulations of Classical Annealing Processes. *Phys. Rev. Lett.* 101 (2008), 130504. Issue 13. https://doi.org/10.1103/PhysRevLett.101.130504 arXiv:0804.1571

[55] Daniel W Stroock and Boguslaw Zegarlinski. 1992. The equivalence of the logarithmic Sobolev inequality and the Dobrushin-Shlosman mixing condition. *Communications in Mathematical Physics* 144 (1992), 303–323. https://doi.org/10.1007/BF02101094

[56] M. Szegedy. 2004. Quantum speed-up of Markov chain based algorithms. In *45th Annual IEEE Symposium on Foundations of Computer Science*. 32–41. https://doi.org/10.1109/FOCS.2004.53

[57] Dave Wecker, Matthew B. Hastings, and Matthias Troyer. 2016. Training a quantum optimizer. *Phys. Rev. A* 94 (2016), 022309. Issue 2. https://doi.org/10.1103/PhysRevA.94.022309 arXiv:1605.05370

[58] Pawel Wocjan and Anura Abeyesinghe. 2008. Speedup via quantum sampling. *Phys. Rev. A* 78 (2008), 042336. Issue 4. https://doi.org/10.1103/PhysRevA.78.042336 arXiv:0804.4259

[59] Theodore J. Yoder, Guang Hao Low, and Isaac L. Chuang. 2014. Fixed-Point Quantum Search with an Optimal Number of Queries. *Phys. Rev. Lett.* 113 (2014), 210501. Issue 21. https://doi.org/10.1103/PhysRevLett.113.210501 arXiv:1409.3305

[60] A. P. Young, S. Knysh, and V. N. Smelyanskiy. 2010. First-Order Phase Transition in the Quantum Adiabatic Algorithm. *Phys. Rev. Lett.* 104 (2010), 020502. Issue 2. https://doi.org/10.1103/PhysRevLett.104.020502 arXiv:0910.1378

[61] Leo Zhou, Sheng-Tao Wang, Soonwon Choi, Hannes Pichler, and Mikhail D. Lukin. 2020. Quantum Approximate Optimization Algorithm: Performance, Mechanism, and Implementation on Near-Term Devices. *Phys. Rev. X* 10 (2020), 021067. Issue 2. https://doi.org/10.1103/PhysRevX.10.021067 arXiv:1812.01041

Received 2022-11-07; accepted 2023-02-06