FERMILAB-PUB-23-0827-SQMS-V

On the approximability of random-hypergraph MAX-3-XORSAT problems with quantum algorithms

Eliot Kapit[1,3],* Brandon A. Barton[2,3], Sean Feeney[3], George Grattan[3], Pratik Patnaik[1,3], Jacob Sagal[3], Lincoln D. Carr[1,2,3], and Vadim Oganesyan[4,5,6]

[1] *Department of Physics, Colorado School of Mines, 1523 Illinois St, Golden CO 80401*
[2] *Department of Applied Mathematics and Statistics, Colorado School of Mines, 1500 Illinois St, Golden CO 80401*
[3] *Quantum Engineering Program, Colorado School of Mines, 1523 Illinois St, Golden CO 80401*
[4] *Department of Physics and Astronomy, College of Staten Island, CUNY, Staten Island, NY 10314, USA*
[5] *Physics program and Initiative for the Theoretical Sciences, The Graduate Center, CUNY, New York, NY 10016, USA and*
[6] *Center for Computational Quantum Physics, Flatiron Institute, 162 5th Avenue, New York, NY 10010, USA*

A huge range of important problems in computer science–including task optimization, formal logic, encryption, and machine learning–can be solved by finding the sequence of binary variables that optimizes a cost function defined by a series of few-variable constraint relationships. These problems define the complexity class NP, and are in the worst case, and often the typical case, exponentially hard in the number of variables for all known methods. This hardness applies both to exact and approximate optimization, e.g. finding configurations with a value within a defined fraction of the global optimum. Fundamentally, the lack of any guided local minimum escape method ensures the hardness of both exact and approximate optimization classically, but the intuitive mechanism for approximation hardness in quantum algorithms based on Hamiltonian time evolution is not well understood. In this work, using the prototypically hard MAX-3-XORSAT problem class, we explore this question. We conclude that the mechanisms for quantum exact and approximation hardness are fundamentally distinct. We qualitatively identify why traditional methods such as high depth quantum adiabatic optimization algorithms are not reliably good approximation algorithms. We propose a new spectral folding optimization method that does not suffer from these issues and study it analytically and numerically. We consider random rank-3 hypergraphs including extremal planted solution instances, where the ground state satisfies an anomalously high fraction of constraints compared to truly random problems. We show that, if we define the energy to be $E = N_{\rm unsat} - N_{\rm sat}$, then spectrally folded quantum optimization will return states with energy $E \leq AE_{\rm GS}$ (where $E_{\rm GS}$ is the ground state energy) in polynomial time, where conservatively, $A \simeq 0.6$. We thoroughly benchmark variations of spectrally folded quantum optimization for random classically approximation-hard (planted solution) instances in simulation, and find performance consistent with this prediction. We do not claim that this approximation guarantee holds for all possible hypergraphs, though our algorithm's mechanism can likely generalize widely. These results suggest that quantum computers are more powerful for approximate optimization than had been previously assumed.

Contents

*Electronic address: ekapit@mines.edu

This manuscript has been authored by Fermi Research Alliance, LLC under Contract No. DE-AC02-07CH11359 with the U.S. Department of Energy, Office of Science, Office of High Energy Physics.

I. INTRODUCTION

Combinatorial optimization of constraint satisfaction problems (CSPs) is an enormously important–and often, enormously difficult–area of modern computer science [1]. Specifically, a huge array of problems in optimization, cybersecurity, machine learning, and more amount to finding low-energy configurations of large collections of M few-body constraints over N binary variables, see Fig. 1. Generically, the energy landscape of these cost functions is extremely rough (see Fig. 3) with exponentially many local minima, making it very difficult to find the true ground state or even a sufficiently low energy configuration.

From the point of view of statistical physics these cost functions are often equivalent to the Hamiltonian of a *disordered spin glass*, and the core hardness mechanism comes from the inability of the system to efficiently escape local minima by flipping small numbers of spins at each step [2–13]. In the hardest problems–and often, even, in the typical case–the time to find the solution grows exponentially for all known classical and quantum methods. Remarkably, these problems are not only hard to solve, but also hard to approximate, where approximation is defined as finding any configuration within a defined fraction of the global optimum [14]. And while a host of clever algorithms have been proposed over the years to attack these problems, supplemented by rapid growth in computing power since the invention of integrated circuits, the exponential worst-case difficulty scaling remains, and is believed (in the as yet unproven statement that P $\neq$ NP) by most computer scientists to be a fundamental and insurmountable fact of our reality, at least for classical computers.

To make further progress, a host of heuristic quantum algorithms have been developed, such as analog quantum annealing [15–21], and its closed system and digital cousins, adiabatic quantum computing (AQC) [22, 23] and quantum approximate optimization algorithms (QAOAs) [24]. In all of these methods the spin glass problem Hamiltonian (diagonal in the computational z basis) is combined with a transverse field term (typically, a uniform field along x), which allows the system to escape from local minima by multiqubit tunneling, a collective process where large clusters of qubits all change configuration simultaneously to tunnel from one minimum to another [25]. In some artificial problem instances this process has been shown to produce exponential speedups compared to classical simulated annealing, with polynomial speedups observed in experiment for short-ranged graphs [21, 26–28]. However, generally applicable beyond-quadratic speedups for NP optimization problems have not been realized, and given the present noisy state of quantum hardware and the projected severe overhead of error correction [29], this degree of speedup is insufficient for practical advantage over classical machines.

In this work, we approach the question of approximation hardness with the goal of identifying both core intuitive mechanisms ensuring it, and opportunities for exponential quantum advantage. In particular, for MAX-3-XORSAT (a particularly difficult CSP class, defined in detail in section II D), we argue the following:

- The hardness mechanism(s) for directly finding the ground state of a given problem Hamiltonian H_P with heuristic quantum methods can be readily identified, and likely cannot be circumvented in the worst cases.

- Unlike classical algorithms based on local updates, the mechanisms which ensure that it is hard to find the ground state do not generalize to ensure approximation hardness for quantum algorithms.

- However, there are good reasons to believe that traditional quantum approaches such as AQC or QAOA are not effective approximators (e.g. reliably returning low energy states in polynomial time) in the worst case. We present relatively large scale numerical simulations that support this expectation.

- Understanding why this is the case suggests a novel *spectral folding* quantum strategy, that transforms H_P and then solves the transformed Hamiltonian through more traditional means. For some variations the performance of spectrally folded quantum optimization can be predicted analytically and promises an efficient approximation guarantee for an extremely large fraction of instances, well into the classically hard regime. For random hypergraphs in this regime, spectrally folded quantum optimization provides an exponential speedup for returning low energy states.

We back up all of these claims with extensive theoretical analysis and numerical tests of all core predictions, to the

largest feasible system sizes for simulating quantum algorithms, and to the largest sizes needed to ensure asymptotic scaling has been reached for classical algorithms. In doing so, we define instance construction rules that ensure classical approximation hardness, at least for algorithms based on local updates, and numerically establish that high depth QAOA does not exhibit meaningful quantum advantage in finding either exact or approximate solutions to these hard instances. We analytically and numerically show that spectrally folded quantum optimization can efficiently approximate these problems, in practice in a linearly growing number of cost function evaluations.

This paper is structured as follows. In section II, we first provide an overview of classical and quantum approximation hardness and the MAX-3-XORSAT problem, and define the construction rules for the problems studied in this work. In section III, we define spectrally folded quantum optimization and propose two variations of our algorithm. Then, in section IV, we develop an extensive-order resummed quantum perturbation theory capable of predicting the performance of these algorithms, and establish an approximation guarantee for random hypergraphs. To verify all of these claims, in section V, we present extensive numerical tests and simulations for a variety of algorithms and problem parametrizations. A summary of the key simulation results is presented in table I. We finally offer concluding remarks, and include additional technical details in the appendices.

II. HARDNESS MECHANISMS, PROBLEM DEFINITIONS AND PREVIOUS APPROACHES

A. Inability of previous methods to find or even approach the ground state at large N

To motivate our novel methods, it is important to first review the qualitative reasons classical and established quantum approaches are unable to efficiently solve or approximate these problems. The classical failure mechanism is straightforward: hard problems display a high density of poor quality local minima, e.g. high energy as compared to the true ground state. Once a local minimum is reached, as no general mechanism for guided local minimum escape exists it is impossible to know in general how close one is to a ground state, either in energy or Hamming distance (number of bit flips separating two states), at least unless the found minimum happens to satisfy a large fraction of constraints.

We emphasize that we are concerned here with an approximation guarantee, not just a method that works well in practice. For completely random problems, one can often predict the average ground state energy using statistical physics arguments, such as the famous Parisi solution to the Sherrington-Kirkpatrick model [30]. If one is able to find configurations close to this energy in a given instance, that is not sufficient to rule out the existence of some other, much deeper minimum far away in configuration space, even if randomly drawing problems with such deep minima is exponentially unlikely. In other words, there is no efficient classical algorithm to know if a given instance is extremal in this way unless P=NP [14].

Since there are in many cases exponentially many more poor quality minima than low-lying ones, the basin of attraction of the true ground state is generally an exponentially small fraction of total configuration space and thus very hard to find, a phenomenon that has been referred to as an "entropic barrier" to problem solving [12]. In particular, the authors of [12] showed that for the 3-XORSAT problem they considered, once a local minimum was found it was more efficient to simply restart the algorithm from a random state instead of attempting to climb out of the found minimum through penalized local operations, as in simulated annealing or parallel tempering [31]. We expect this may be a generic feature in some of the hardest CSP classes. And interestingly, these arguments apply equally well to approximation hardness, not just finding the optimal solution. While for a given class and system size approximation is nominally an easier problem, given that there are many more valid approximate solutions, both tasks scale exponentially in the worst case, for fundamentally the same reason. We now turn to quantum algorithms, for which the situation is considerably more complex.

B. Mechanisms for quantum solution hardness: exponentially small gaps and transverse field chaos

We consider a broad class of heuristic quantum algorithms, which derive from quantum annealing, AQC and QAOA. These algorithms can be supplemented with additional gate model techniques, such as amplitude amplification [32], which improve performance but do not circumvent exponential runtimes. In these algorithms the system is initialized in the ground state of a trivial Hamiltonian, which is then slowly interpolated into H_P; the system is then measured. Other variations based on energy matching [33–36] initialize a known or planted low energy state of H_P and then use collective quantum tunneling to try to find other low energy states.

These algorithms are fundamentally distinct from classical approaches in two ways. The first is the presumed mechanism for quantum advantage: multiqubit quantum tunneling (or collective quantum phase transitions in general), where local minima can be efficiently escaped through many-body quantum effects that have no classical analog. Second and more subtly, where classical local update algorithms all start from a random high energy state and attempt to cool to low energy states, quantum algorithms start below the energy of the problem ground state. They then attempt to tunnel into the global minimum and into other low energy states as the energy of the initial state crosses the target state as the total Hamilto-

nian changes.

Unfortunately, when compared to other applications such as quantum simulation or factoring large numbers with Shor's algorithm, the realistic performance advantage of these algorithms is generally much more modest. In the worst case–and for many problem classes, the typical case–the macroscopic quantum tunneling rate into the ground state decreases exponentially in N. This is a fairly generic expectation, as in many cases, including the MAX-3-XORSAT problem discussed below, the tunneling rate at the crossing point can be computed using Nth order perturbation theory and the convergence of such a method implies exponential decay. For some specific problems this can be circumvented by introducing additional terms into the evolution, such as the well-known result of ramping transverse field terms down one-by-one in mean-field p-spin ferromagnets [37], though generalizations of this and other methods [20, 38–42] to realistic disordered problems are not expected to show similar advantages at large scales.

Further, for a given class, even if one can somehow ensure that the paramagnet-to-spin-glass transition decays polynomially (as it does in the Sherrington-Kirkpatrick problem [43, 44] and, likely, for MAXCUT [45, 46]), that is not sufficient to ensure that the solution can be found in polynomial time. This is because of a phenomenon known as transverse field chaos (TFC) [47, 48], where energetic corrections from the transverse field can change the energy hierarchy of classical minima in the quantum spin glass phase, potentially pushing local minima below the energy of the true ground state of H_P (when all transverse terms are turned off). Consequently, optimization methods will steer the system toward these false ground states first, with additional phase transitions that occur as the transverse field is further weakened. As these transitions occur at weak field values, from the analysis in section IV they are generically exponentially slow, and indeed, engineering this effect intentionally is an elegant way to craft hard benchmark problems for quantum algorithms [49]. This effect can be avoided by restricting the algorithm to unstructured driver Hamiltonians [50] or very weak transverse fields, but in either case performance is very poor. The combination of exponentially small gaps and TFC make it extremely unlikely that any quantum algorithm of this type can reliably and directly find the ground state of NP-complete problems.

C. Approximation hardness and conventional AQC/QAOA

Approximation hardness, however, is another story. For many NP-hard problems guaranteeing an approximation better than random guessing by a constant fraction is also NP-hard [14, 51, 52]. As TFC involves crossings between states that were close in energy to begin with, it cannot by itself lead to quantum approximation hardness at this level. So for example, if the random state energy of a given class is chosen to be zero and the ground state $-N$, a sufficiently general convention, then any algorithm which could return states in polynomial time with energy $\leq -cN$ for constant $c > 0$ for all instances would promise a potentially exponential speedup. Thus, a hypothetical algorithm which always returned states with energy $-N/3$ or below in polynomial time assuming spectral continuity would still promise an exponential speedup for the hardest problem classes even if TFC reduced that guarantee to $-N/4$. In other words, while TFC can prove ruinous for finding an exact solution when the problem exhibits a clustering phase [13, 53–58] and there are exponentially many states very close to the ground state in energy but well separated from it in Hamming distance, it is not going to push zero energy states into competition with $O(1)$ fractions of the ground state energy.[1]

Exponentially small gaps are a more serious problem, but those too cannot so easily be assumed to ensure approximation hardness. This is because the empirical "difficulty exponents" of phase transitions in low-order CSPs are often quite small; for random hypergraph MAX-3-XORSAT instances, the minimum gap at the paramagnet-spin glass transition scales as approximately $\Omega_0(N) \sim 2^{-cN}$ with $c \simeq 0.14$ (see section IV B). And while the fraction of states $p_E(N)$ below the approximation threshold is typically exponentially small, the total number of such states is exponentially large. If we could assume that the mixing rate between the paramagnetic initial state and the dressed excited states of H_P is equivalent to that of the ground state, then the naive product of $\Omega_0^2(N) \times 2^N \times p_E(N)$ is often exponentially large as well, which would ensure fast approximation. But this naive analysis is inaccurate, because as we argue momentarily, mixing matrix elements with excited states decay more quickly, and the decay exponents grow with excitation energy. The exponential number of target states in approximation problems allow an algorithm suffering from both TFC and exponentially decaying matrix elements to still guarantee an efficient approximation, provided, at least, that those decay exponents are not too small. In some sense, our novel spectrally folded quantum algorithm achieves fast approximation of classically hard instances by ensuring that the exponential decay rate of tunneling into the problem's excited states matches or exceeds that of the ground state in more traditional methods.

Given all this, should we expect that AQC or QAOA will efficiently approximate NP-hard problems? Formally, it has not been proven that these algorithms are not efficient approximators when the circuit depth is allowed to grow as a low-order polynomial in system size,

[1] It's important to note that for some problems where the classical approximation hardness threshold is relatively close to 1, such as MAXCUT, TFC might well become a serious obstacle to achieving quantum advantage for approximation.

though there are good reasons to be doubtful, supported by a number of recent works [59–62]. Let us consider, qualitatively, how AQC/QAOA solves such a problem, assuming the limit of quasi-continuous time; we present a quantiative analysis below in Sec. IV B. The system is initialized in the paramagnetic ground state of a uniform transverse field Hamiltonian H_D, and the system evolves in time interpolating between H_D and H_P by lowering the coefficient of one and raising the coefficient of the other, raising the energy of the paramagnetic state until it crosses the problem ground state from below. We again emphasize the fundamental distinction of crossing from below, rather than cooling from above, in quantum and classical algorithms. The minimum gap at the transition is expected to decay exponentially in N and is approximately given by the overlap of the dressed problem ground state $|G_D\rangle$ with the paramagnet state $|S\rangle$; we perform this calculation in Sec. IV B. If the dressings are weak enough that $|G_D\rangle = |G\rangle$ (the Grover limit [63]), then $\langle S|G_D\rangle = 2^{-N/2}$; however for random MAX-3-XORSAT problems the perturbative corrections spread $|G_D\rangle$ over a more significant (if still exponentially small) fraction of Hilbert space and reduce the decay exponent to around a quarter of that in the Grover case.

Assuming that we evolve time too quickly (e.g. not exponentially long) and miss the primary phase transition, we can ask how efficiently $|S\rangle$ will mix with the problem's excited states, as these rates ultimately determine the algorithm's efficacy as an approximator. Calculating them directly is very difficult, but we can qualitatively predict that they should be much smaller for two reasons. First, these crossings occur as the transverse field strength κ is reduced toward zero, and since the dressings that reduce the decay exponent all scale with extensive powers of κ, they will be much reduced by any reduction in κ itself. Second, when considering a dressed excited state $|E_D\rangle$, the perturbative dressings that come from mixing with states with lower energy now have opposite sign and destructively interfere with other corrections in the overlap with $|S\rangle$, in contrast to $|G_D\rangle$ where all higher order terms are positive definite. Both of these effects can, and do, considerably worsen the mixing rates with excited states and make AQC/QAOA a poor approximator in the worst cases, and often in practice.

D. The MAX-3-XORSAT problem and approximation-hard instance construction

Given the severe overhead of fault tolerance [29], quantum hardware is expected to exhibit enormous prefactor disadvantages as compared to parallel silicon, particularly when the comparison is made to hardware with equivalent financial value (e.g. millions of USD). Quantum algorithms thus have the most promise when the problem is hard or outright impossible for classical machines. NP-hard constraint satisfaction problems are no exception, so when benchmarking a proposed quantum algorithm it is important to ensure that the problem classes we consider are sufficiently hard for classical machines, and in the present NISQ era, exhibit their exponential difficulty at small enough N that numerical simulations of quantum algorithms can demonstrate meaningful improvements. MAX-3-XORSAT problems are ideal for benchmarking quantum algorithms because their exponential difficulty scaling is obvious at small N for both classical and prior quantum approaches, in contrast to other problems where the asymptotic exponential scaling often does not set in until system sizes that are prohibitively large for simulation.

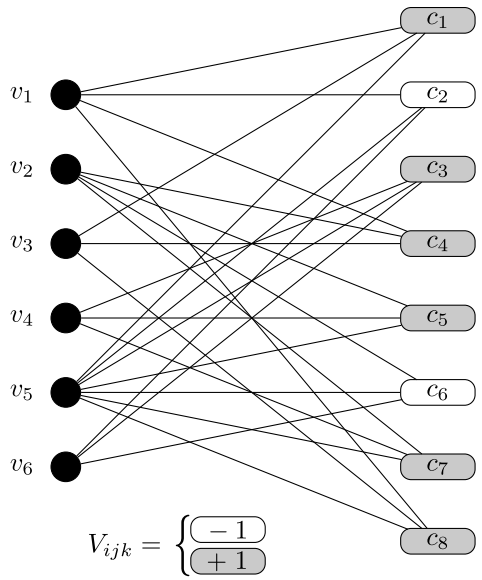

FIG. 1: Graphical representation of a random 3-uniform hypergraph $G_{\mathcal{H}} = (\mathcal{V}, \mathcal{E})$ used in the MAX-3-XORSAT problem. The set of vertices $\mathcal{V}$ labelled $\{v_1, \ldots, v_6\}$ are connected to a hyperedge (i.e., constraint) from the set $\mathcal{E}$ on the right labeled $\{c_1, \ldots, c_8\}$ if the vertex is in constraint. Each constraint necessarily contains a random set of three unique vertices. The legend specifies the value of the variable $V_{ijk} = \pm 1$ where the grey and white filled c_j boxes denote a $(+1)$, and (-1) valued constraint respectively. This example has a constraint density of $N_C = \frac{4}{3}N$ for 6 vertices and 8 constraints, although we use other values of N_C throughout this work.

To prototype our quantum algorithms, we thus consider MAX-3-XORSAT [14], a well-studied problem that consists in the Pauli basis of a hypergraph of N_C three-body constraint terms, as sketched in figure 1:

$$H_P = -\sum_{ijk}^{N_C} V_{ijk} Z_i Z_j Z_k, \;\; V_{ijk} = \pm 1. \tag{1}$$

A constraint is said to be satisfied if, for a given bitstring, $V_{ijk} Z_i Z_j Z_k = 1$, and unsatisfied otherwise. This is called a hypergraph because V_{ijk} has three indices rather than the usual two found in graph theory. Thanks to the linearity of the problem, one can use Gaussian elimination

to check if a solution exists that satisfies all the constraints in $O\left(N^3\right)$ time, but if the problem is not fully satisfiable, finding the lowest energy state(s) is NP hard. A random state satisfies half the constraints on average and is thus energy zero. Further, it was shown by Håstad [52] that if the true ground state satisfies a fraction $(1-\epsilon)$ of the constraints, then finding any configuration that satisfies more than $(1/2+\epsilon)$ of them is also NP-hard. The hardest instances are thus those with small but finite ϵ, e.g. almost satisfiable problems, as both finding the true ground state, and even finding an approximate solution, is exponentially difficult. Note that if the problem graph is sparse (e.g. N_C/N is on the order of 1) finding approximate solutions can still be easy, since one can randomly select a fraction of the constraints $< cN$ (for some $O(1)$ constant c), and solve that new, much easier problem; solutions to this sub-problem will satisfy half the remaining constraints, on average.

To ensure that we are studying problems that are both hard to solve and hard to approximate for all known methods, we consider a family of instances we call planted partial solution problems (PPSPs). To construct a PPSP, we choose a small unsatisfied fraction ϵ and pick a random hypergraph of $N_C \gg N$ unique triplets; we use $\epsilon = 0.1$ in all simulations here. We then pick a random bitstring G and randomly select $(1-\epsilon)N_C$ of the constraints to be satisfied in G, by picking the sign of V_{ijk} appropriately, with the rest unsatisfied. If ϵ is small and $N_C/N \gg 1$, G will be the problem ground state with very high probability, as the SAT/UNSAT transition for this problem is at $N_C/N \sim 0.92$ [64] and at densities much higher than the SAT/UNSAT threshold ground states for random graphs satisfy $N_C/2 + O\left(\sqrt{N_C}\right)$ constraints. This property also ensures that G is a unique ground state with high probability at large N. When we refer to random hypergraph problems throughout this work, we refer to this construction rule: a random, potentially fairly dense, hypergraph where one can optionally randomly chose an anomalously large fraction of constraints to be satisfied by matching the signs to a randomly chosen ground state bitstring.

Our PPSP construction is necessary because truly approximation hard problems–where the practical polynomial time approximation difficulty approaches the random guessing limit of the complexity class separation–are rare in the space of all possible instances. Sufficiently sparse problems are approximation-easy, and for denser random problems one can always find strings that satisfy $N_C/2 + O\left(\sqrt{N_C}\right)$ constraints in polynomial time [65], with a smaller prefactor in front of the $\sqrt{N_C}$ than the prefactor in the average satisfied in the ground state. We formulated our PPSP construction to ensure our algorithm was being benchmarked on instances with a plausible claim to true classical approximation hardness. We note that commonly studied 3-regular problems [12, 66] do not display strong approximation hardness, as they are sparse, and can be solved efficiently if satisfiable. And intriguingly, we show that, by some metrics, the performance of our novel algorithm progressively improves with increasing N_C/N in this regime, at constant ϵ.

E. Absence of angle fine tuning in this work

Before presenting our algorithm and main results, we want to make one last point about what we mean by QAOA when talking about quantum algorithms. Specifically, following the original proposal [24], most studies of QAOA allow the individual angles governing the magnitudes of the driver and problem Hamiltonians, H_D and H_P, to vary independently at each timestep. For finding ground states with high depth circuits, this technique is analogous to the schedule fine tuning concept that dates back to the adiabatic formulation of Grover's algorithm [63], and yields a quadratic speedup. And for many problems, these algorithms exhibit *concentration*, where the set of angles that is optimal for one randomly generated instance is close to optimal for another with high probability [44, 61]. More general formulations such as ADAPT-QAOA [67] can potentially offer more significant speedups, albeit with the challenge of a much larger search space for optimizing control parameters.

We avoid this approach in our work, and instead focus only on simple heuristics where the schedule is determined from smooth functions based on intuitive guessing and a small amount of trial and error. We do this for two reasons. First, as argued in [68], quadratic speedups from schedule fine tuning are extremely fragile and unlikely to be viable at large N except in very narrow circumstances. Second, as the truly approximation hard instances of MAX-3-XORSAT (and we suspect, many other CSPs) are extremal in the space of random problems, it is less obvious that concentration arguments would apply to the cases we consider. So even if, for example, these instances exhibit concentration and some set of angles is near-optimal for a specific ϵ and N_C/N scaling, we do not think it can be easily assumed that those angles would generalize to other extremal parametrizations.

The fundamental challenge of approximation hardness is to find, or rule out the existence of, one or more very deep minima in an exponentially large search space. Consequently, optimizing angles based on average energies returned is not expected to improve algorithm performance. If the algorithm finds a state in the basin of attraction of the deep minimum, we can likely halt as a successful approximation has been found, but if it does not, sets of angles that return lower energies for random, uncorrelated shallower minima are not expected to improve the probability of reaching states near the deep minimum. And again, even if a given subset of extremal problems does exhibit concentration for angle optimization, those angles may not generalize to any other extremal part of parameter space.

We therefore restrict all of our simulations in this work to linearly increasing total runtime with N, and smooth, simple functions to control the relative magnitudes of

H_D and H_P. Unsurprisingly, the best scaling and prefactor choices differ somewhat for QAOA and variations of spectrally folded optimization, and the results here are not optimal for any specific algorithm variation or PPSP subclass, but rather represent a decent choice, found by intuition and trial and error, for a broad set of parameters.

III. SPECTRALLY FOLDED QUANTUM OPTIMIZATION ALGORITHMS

The core idea of spectral folding (and related, more general spectral deformations) is to modify how the Hamiltonian is applied to the quantum state through the introduction of a filter function. Specifically, the algorithms we consider solve problems through simulating the time evolution of a quantum state, as $|\psi\rangle \to e^{-iH(t)dt}|\psi\rangle$, with the exponentiated Hamiltonian discretized as a series of layers e.g. $e^{iaH_D}e^{ibH_P}$. The driver Hamiltonian, and any other additional Hamiltonian terms, are not changed by spectral folding, so we will ignore them for now and focus on the problem Hamiltonian itself. Specifically, we write $|\psi\rangle$ in the computational basis as $|\psi\rangle = \sum_{m=0}^{2^N-1} c_m |m\rangle$, where m is the decimal integer representation of a given bitstring. Then, for (arbitrary) control angle γ:

$$e^{i\gamma H_P}|\psi\rangle = \sum_{m=0}^{2^N-1} e^{i\gamma E(m)} c_m |m\rangle, \tag{2}$$

$$E(m) = \langle m| H_P |m\rangle = -\sum_{ijk}^{N_C} \langle m| V_{ijk} Z_i Z_j Z_k |m\rangle. \tag{3}$$

In other words, the phase of each component state advances proportionally to its energy under the problem Hamiltonian, and that energy is computed at each step by applying a sequence of gates to implement each constraint.

In spectral folding, the phase of each component instead advances proportional to an arbitrary function f of the diagonal H_P,

$$|\psi\rangle \to e^{i\gamma f(H_P)}|\psi\rangle = \sum_{m=0}^{2^N-1} e^{i\gamma f(E(m))} c_m |m\rangle. \tag{4}$$

This can be accomplished by introducing a register of auxiliary qubits, applying a gate sequence that maps the sum of the constraint terms to a fraction of that register to store E, using a second fraction of that register to compute $f(E)$, applying a sequence of controlled-phase gates to advance the phase by $f(E)$, and then uncomputing the previous steps to return the register to its initial state. The entire process is sketched in figure 2. Provided f is a relatively simple function, this adds a multiplicative overhead which is polylogarithmic in N, since $E(m)$ is bounded by a polynomial in N and each arithmetic operation takes $O(\log N)$ steps.[2] We define *spectral warping* as any f which applies a nonlinear rescaling of E, such as $f(E) = cE^2$, and *spectral folding* as a choice of f that mirrors E about a specific value, e.g. $f(E) = |E - E_t|$. The core idea is sketched in figure 3, and these two methods can of course be composed. Incorporating this operation enormously expands the space of quantum optimization algorithms we can define; in this work we focus on two choices, *linear* and *quadratic* spectral folding. Specifically, if we choose our problem normalization via including a multiplicative constant[3] so that the ground state energy is $E_{\rm GS} = -N$, and let $E_t = AN$ for a constant A, then we define linear and quadratic spectral folding as

$$f_{\rm lin}(E) = \frac{|E + AN|}{A}, \quad f_{\rm quad}(E) = \frac{(E + AN)^2}{A^2 N} \tag{5}$$

Here, $A < 1$ defines the approximation target. And critically, it is defined using the conventions that random states have energy zero, so returning a state with energy $AE_{\rm GS}$ approximates, by a factor of A, the degree to which the true ground state itself improves on random states. These normalization choices ensure that the energy difference between the new ground states of the folded problem, and random states, is N as in the original renormalized problem. Making this choice simplifies the analysis significantly.

A good choice of A is important for spectrally folded quantum optimization to succeed; if A is chosen to be too close to 1, then we risk failing to well-approximate H_P due to the interference effects mentioned in section II C. A choice of A which is too small will return a suboptimal approximation ratio, and if A is too close to zero, cause instabilities from having a poorly defined problem to solve. Fortunately, for random hypergraph MAX-3-XORSAT instances–and here random refers to the graph itself and not, critically, on how many constraints are satisfied in the ground state–we can predict the threshold A for which we expect a polynomial depth circuit to return states with $E \simeq AE_{\rm GS}$ from first principles. The ideal value of A depends both on the problem class and on the variation of spectral folding employed; for MAX-3-XORSAT, $A \geq 0.6$ is achievable as derived below in section IV D. This is a significant leap over the best known classical approximation algorithm for this

[2] We note that for any choice of f more complex than multiplying E by a constant (something that does not require auxiliary qubits to begin with), any spatial locality the graph might have is lost in this step, since $E(m)$ is a global quantity which we are deforming with f.

[3] Formally, this choice assumes that we know the fraction of the N_C constraints which are satisfied in the ground state, something that we cannot know in advance of running our quantum optimization algorithm! However, we can simply repeatedly run the algorithm with different normalization choices to guess its value, a prefactor overhead of at most $O(N_C)$.

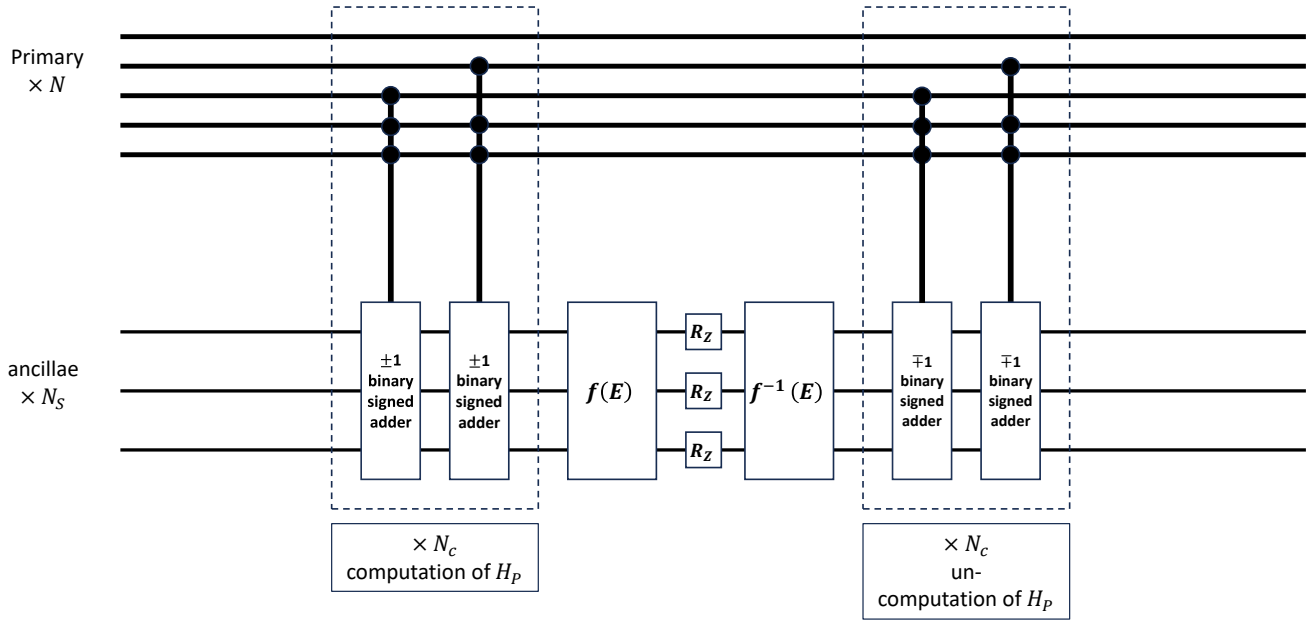

FIG. 2: A schematic for implementing spectrally deformed time evolution $U = \exp(if(H_P)dt)$. A set of gates maps the value of each of the N_C constraint to adding or subtracting 1 to a register of ancilla qubits, using binary signed adder circuits controlled by the value of $V_{ijk}Z_iZ_jZ_k$. A sequence of additional gates computes $f(E)$ from E (likely using more ancilla qubits), and then a set of local Z rotations is applied to the register storing $f(E)$ to advance the phase of each component of $|\psi\rangle$ proportionally. The computation of $f(E)$ and entangling with "constraint-controlled" gates are then uncomputed, returning the ancillas to their initial state and disentangling them from the N primary qubits over which the problem is defined. The net result of this entire process is to enact the operation in Eq. (4), evolving time under an arbitrary function of the diagonal Hamiltonian (transverse field layers and other operations on the primary qubits are not shown). For relatively simple functions, the net overhead of this entire process (compared to enacting $\exp(iH_Pdt)$ directly) is polylogarithmic in N.

problem [69], which offers a weaker guarantee with much more restricted viability, in the worst case, equivalent to $A \to 0$ in our notation. It is also a significant leap over recent quantum approaches to this problem [70, 71]. We note that minimizing $(H-E)^2$ is not itself a novel idea and has been used in classical and quantum algorithms for finding states close to specific energies in chemical and many-body systems [72–78]. To our knowledge, however, the use of spectral folding for approximate optimization of CSPs is novel, both in concept and in the analysis we present below to choose A and understand at a deeper level why it presents significant advantages over optimizing the problem H_P directly.

From hereon, we let H_{fold} be the spectrally folded problem Hamiltonian. Having defined it, there are a number of ways we can attempt to find its ground states. The simplest choice, and one that performs well, is AQC/QAOA-inspired state preparation, where we interpolate in Trotterized evolution between the transverse field driver and the folded problem over a time t_f:

$$H(t) = f(t)\,H_D + g(t)\,H_{\text{fold}}, H_D = -\sum_j X_j, \quad (6)$$

$$f(0) = g(t_f) = 1, \;\; f(t_f) = g(0) = 0. \quad (7)$$

This prescription, with the quadratic folding choice in Eq. (5), is the most straightforward to benchmark using standard quantum simulation packages as each call requires $O\left(N_C^2\right)$ multiqubit $Z_iZ_jZ_kZ_l...$ rotations. Classically simulating the linear folding prescription requires auxiliary qubits to implement the absolute value operation or, much more practically, saving the phase oracle as a pre-computed diagonal operator.

We can also consider trial minimum annealing (TMA), originally proposed in [79]. We explore the TMA formulation in depth because we can predict its scaling analytically. In this scheme, a simple classical algorithm is used to find an initial local minimum of H_P; the quality of the minimum does not particularly matter and for approximation-hard instances we assume it is far above the true ground state energy, in the worst case asymptotically approaching random guessing. Let this classical minimum state be $|L\rangle$. We will use the linear folding prescription in Eq. 5 for H_P itself, and add to it a new, diagonal *lowering Hamiltonian* H_L which has $|L\rangle$ as its ground state, and assign to it a time-dependent coefficient $C(t)$. Our total cost function Hamiltonian is

$$H_{\text{cost}}(t) = \frac{|H_P + AN|}{A} + C(t)\,H_L. \quad (8)$$

Recall that H_P is normalized so that its ground state energy is $-N$. To go further, we need to specify a form for H_L. For this analysis will choose a new random hypergraph of N_C triples which, critically, has no correlation

to the hypergraph of H_P; we choose the same N_C as the problem for convenience here but any $O(N)$ quantity should be fine. We choose the signs of all constraints so that $|L\rangle$ satisfies all of them. H_L is not included in the folding procedure so applied separately in time evolution. We then choose $C(t=0)$ such that the initial energy of $|L\rangle$ (defined by $H_{\text{fold}} + C(t) H_L$) is well below $-N$ but remains $O(N)$. Our algorithm simulates appropriately discretized time evolution in the following sequence:

- Initialize $|L\rangle$, with H_{cost} always on, and evolve time smoothly ramping up the transverse field from 0 to κ in time t_r. We assume t_r increases linearly with N and choose $\kappa \leq \kappa_c$; $\kappa_c \simeq 1.3$ for MAX-3-XORSAT but can vary for other problems, and we expect weak variation from one instance to the next. We want to choose κ at or just below this value, so we remain in the dressed problem phase (DPP, defined in section IV) at all times. Leaving and then re-entering the DPP does not mean the algorithm will fail but makes predictions harder. This smoothly evolves the state to $|L_D\rangle$, the dressed version of $|L\rangle$.

- Evolve time for a total time T, likely also $O(N)$, where $C(t)$ is smoothly ramped down to zero, ensuring that $|L\rangle$ crosses the hyperspherical shell of ground states of $H_{\text{fold}} = |H_P + AN|/A$. Note that this crossing occurs when the ground state energy of $C(t) H_L$ is $O(-N)$, and if we assume the initial minimum was uncorrelated with the true ground state $|G\rangle$, the mean Hamming distance between $|L\rangle$ and any of the ground states of the folded Hamiltonian is $N/2$ flips. Consequently, H_L adds an $O\left(\sqrt{N}\right)$ energy uncertainty to these states that has no meaningful impact on the approximation ratio.

- Finally, ramp the transverse down to zero smoothly over t_r and measure the system in the z basis. For an appropriate $O(1)$ choice of A and a random problem hypergraph, this algorithm will return states with energies close to AE_{GS} with constant probability. We can optionally repeat the algorithm many times, starting from different choices of $|L\rangle$, to ensure a fairer sampling of states in that energy range.

The total gate count of this algorithm is as follows. We have a factor of $O(N + N_C \,\text{polylog}(N))$ per timestep for the layers of transverse field, H_{fold} and H_L terms, which we simplify to $N_C \,\text{polylog}(\text{N})$. We obtain, in the worst case, a factor of $O(N_C)$ for the number of guesses one needs to make to correctly set the normalization for a chosen A. We assume, on empirical grounds, that the total quantum evolution time is $O(N)$.This choice works well in practice in our simulations, and more intuitively, the very simplest classical optimization routine, steepest descent, requires $O(N)$ Hamiltonian calls to halt. We do not think it reasonable, ultimately, that a quantum algorithm should perform well with fewer steps per shot. Finally, in the worst case we expect a timestep $dt \sim 1/N$, for graphs where a small number of variables connect to significant fractions of the N_C constraints, but dt constant or increasing logarithmically is empirically and intuitively fine in the typical case. Taken together, and we emphasize assuming that the algorithm is capable of returning states with $E < AE_{\text{GS}}$ in constant probability, we estimate a total runtime between $O\left(N_C^2 N^2 \,\text{polylog}(N)\right)$ in the worst case and $O(N_C N \,\text{polylog}(N))$ in more typical cases.

All that said, justifying the assumption of constant success probability is the central task of the paper. We now provide a theoretical analysis of the performance of this formulation of spectral folding on random hypergraph PPSPs, and a more qualitative analysis of the expected performance of other variations.

IV. ANALYTICAL PERFORMANCE PREDICTIONS

A. Preliminaries

In this section, we will predict from first principles the average macroscopic quantum tunneling rate–and thus, achievable approximation ratio–for the TMA variation of spectrally folded quantum optimization. This version may not be the optimal choice, and there are intuitive reasons to believe that other variations could offer better performance, but being able to make direct analytical predictions enormously strengthens our argument and bolsters the scaling expectations one can infer from our numerical results. To do so, we have developed a somewhat novel resummed extensive order perturbation theory based on previous *forward approximation* results [25, 33, 80–84].

For the random hypergraph problems we study here, the two key factors in determining the macroscopic quantum tunneling rate are the transverse field strength κ_c where a phase transition occurs between the paramagnet and the quantum spin glass, which we call the dressed problem phase (DPP) in this work, and the energy cost $E(x)$ for x random flips away from the ground state. We first derive the energy cost, and remarkably, for random flip sequences it turns out to be graph independent. Specifically, for MAX-3-XORSAT, our problem is defined as a hypergraph of N_C p-body constraints (e.g. $V_{ijk} Z_i Z_j Z_k$) over N variables, where $p = 3$ here, and each constraint returns ± 1 and flips to the opposite value when any one of the spins flips. Let us say the system is in some classical configuration s; the energy is then given by $E(s) = N_C (n_{\text{unsat}} - n_{\text{sat}})$, where a sat constraint returns 1 in this notation, and n implies a density.

Now we flip one spin at random. Each spin participates in, on average, pN_C/N constraints, and consequently, the

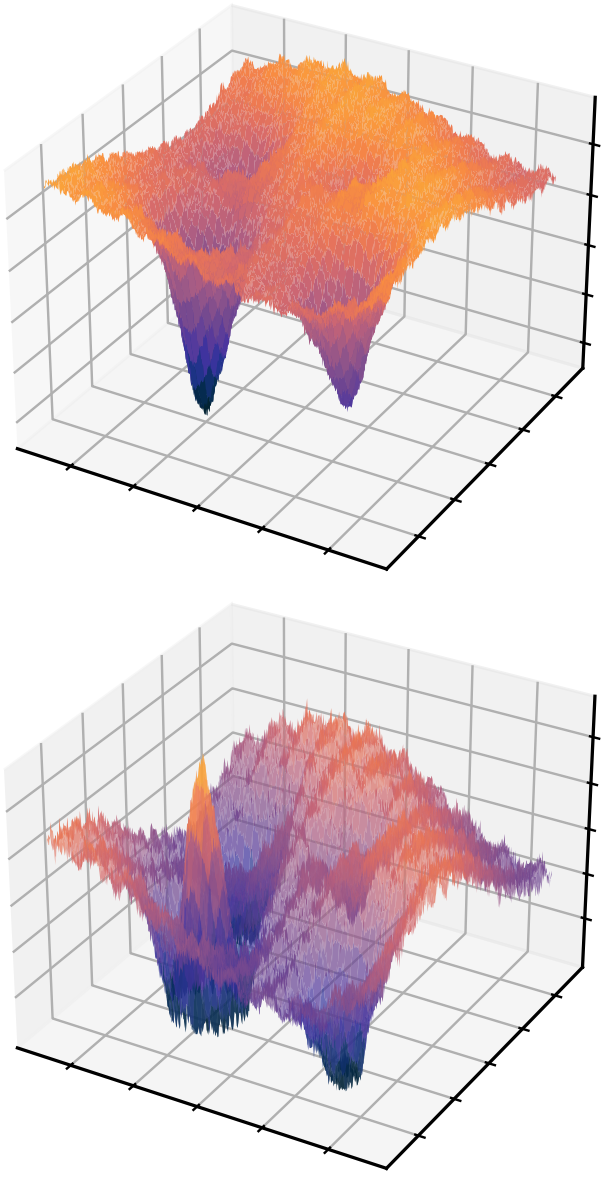

FIG. 3: *Illustration of the spectral folding procedure.* (Top) Sketch of the rough energy landscape of an approximation-hard CSP, with a single deep minimum whose basin of attraction is an exponentially small fraction of the configuration space. Directly optimizing this cost function through quantum approaches often misses the deep minimum entirely, for reasons explained in the text. (Bottom) Spectrally folded energy landscape in Eq. (5), where the problem Hamiltonian energies are mirrored around an approximation target $E = AE_{\rm GS}$. This can be implemented in a gate model algorithm with modest overhead, as shown in the text. Doing so promotes the states near the fold to an exponentially large ground state band while eliminating an interference effect that reduces tunneling into them from trivial initial states; for a wide range of problem instances (and likely, low-order problem classes), this works out to an approximation guarantee. Detailed performance predictions, and numerical benchmarking, are shown in the text.

average energy change for a single spin flip is

$$\Delta E_{\rm avg} = +2p\frac{N_C}{N}\left(N_{\rm sat} - N_{\rm unsat}\right) = -2pE \tag{9}$$

Now imagine we have flipped y spins from our initial configuration. If we flip one more spin at random, once again $\Delta E_{\rm avg}(y) = -2pE(y)\Delta y$. However, we have already flipped y spins, so when we flip one more at random, with probability $1 - y/N$ we have flipped a spin back and are computing the energy change associated with reducing y by 1. Consequently

$$\left(1 - \frac{2y}{N}\right)\frac{\Delta E_{\rm avg}(y)}{\Delta y} = -2pE(y) \tag{10}$$

If we interpret this as a differential equation, it has a straightforward solution: starting from the ground state G, for x unique random flips away the average energy is

$$E_{\rm avg}(x) = E_{\rm GS}\left(1 - \frac{2x}{N}\right)^p. \tag{11}$$

Note that this statement is graph independent, and is only an average; individual trajectories will of course display substantial variations. It is a rederivation of a familiar result for dense hypergraphs with Gaussian distributed constraint energies [33, 85], but is applicable in much broader contexts and is easy to confirm numerically. Throughout this work, we rescale all problems by a multiplicative constant so that the ground state energy is $-N$.

B. Paramagnet to spin glass transition scaling for MAX-3-XORSAT

Before proceeding to our main calculation, we note that, from this result, we can predict the typical case difficulty scaling of finding the ground state with AQC or QAOA, for random hypergraphs as defined in section II D. To predict the phase transition rate, we will use high order perturbation theory to compute the perturbative dressings to the problem ground state; call this state $|G_D\rangle$. Using fourth order perturbation theory, for uniform transverse field strength κ, the energy of the dressed ground state is, again on average[4]

$$E_{\rm GS} \simeq -N\left(1 + \frac{\kappa^2}{2p} + \frac{\kappa^4}{8p^3} + ...\right). \tag{12}$$

[4] In this step the graph structure of H_P can potentially be important. This is because the energy denominator $2p$, the average cost per flip just a few flips away from G, is more sensitive to the details of the graph than the energy cost many flips away, which can in turn shift κ_c. We thus do not claim this result is valid for all graphs. However, we find that for our random PPSPs the transition is consistently near the κ_c we predict here and the scaling of the uniform field result closely matches our prediction.

This crosses the energy of the paramagnetic state, which is $-N\kappa$ and minimally perturbed by the problem Hamiltonian, at $\kappa_c \simeq 1.29$ for our family of $p = 3$ PPSPs. The splitting $2\Omega_0$ at the phase transition is expected to be proportional to the overlap of the uniform superposition state $|S\rangle$, the ground state of H_D, with the dressed state $|G_D\rangle$:

$$\Omega_0 \propto \langle S|G_D\rangle \,. \tag{13}$$

To compute Ω_0, we must thus compute the perturbative corrections to $|G_D\rangle$, to high orders. Again using Eq. (11), if we let $\left|G^{(i,j,k\ldots)}\right\rangle \equiv X_i X_j X_k \ldots |G\rangle$, and $\tilde{E}_{\text{avg}}(k) \equiv E_{\text{avg}}(k) - E_G$, at the transition point we have

$$\begin{aligned} |G_D\rangle \simeq |G\rangle + \tfrac{\kappa_c}{\tilde{E}_{\text{avg}}(1)} \textstyle\sum_j \left|G^{(j)}\right\rangle \\ +2!\tfrac{\kappa_c^2}{\tilde{E}_{\text{avg}}(1)\tilde{E}_{\text{avg}}(2)} \textstyle\sum_{i\neq j} \left|G^{(i,j)}\right\rangle \\ +3!\kappa_c^3 \textstyle\prod_{m=1}^{3} \tfrac{1}{\tilde{E}_{\text{avg}}(m)} \textstyle\sum_{i\neq j\neq k} \left|G^{(i,j,k)}\right\rangle + \ldots \end{aligned} \tag{14}$$

The factorials come from the combinatorics of ordering the m spin flips to reach each term. Now, since all states are present in $|S\rangle$ with equal amplitude $2^{-N/2}$ and all terms in $|G_D\rangle$ are positive definite, we can immediately conclude

$$\Omega_0(N) \simeq 2^{-N/2}\left(1 + \sum_{m=1}^{N} \kappa_c^m \binom{N}{m} m! \prod_{n=1}^{m} \frac{1}{\tilde{E}_{\text{avg}}(n)}\right) \tag{15}$$

For $\kappa_c = 1.29$, this function is well fit by $\Omega_0 = a\sqrt{N}2^{-bN}$, where $b \simeq 0.14$, in decent agreement with the result for a mean-field $p = 3$-spin ferromagnet derived in [86]. We note also the hardness equivalence between dense and random sparse graphs drawn in [59], though we emphasize that the results we derive here are not limited to completely random graphs. If we are in the diabatic regime for $t_f \ll 1/\Omega_0(N)$, the probability of finding the ground state after a sweep is $P_{\text{GS}}(t_f) \simeq \Omega_0^2 t_f/W$, where $W \sim O(N)$ is the energy range swept over in the evolution; for $t_f \propto N$ we thus have $P_{\text{GS}}(t_f, N) \simeq N2^{-2bN}$. Our numerical simulations for PPSPs, shown below, are in good agreement with this result. Note that this assumes a single avoided crossing and does not consider transverse field chaos, e.g. crossings late in evolution between states close in energy; such transitions can in principle be substantially more difficult [87], but are not relevant to approximation hardness in this problem as argued earlier.

C. Scaling of inter-valley tunneling in a $p = 3$ quantum spin glass

This calculation is less easy to generalize to a spectrally folded Hamiltonian, however, where we have exponentially many competing ground states, clustered in a thin hyperspherical shell around the true minimum, at least in approximation-hard problems. So instead we will consider tunneling between two p-spin wells in an N spin system, spaced $N/2$ flips apart, and show that the average per-state tunneling rate in the TMA formulation of spectrally folded optimization should have near-identical scaling. Our total N-spin Hamiltonian, consequently, is

$$\begin{aligned} H = -\frac{1}{N^{p-1}}\left[\left(\sum_j Z_j a_j\right)^p + \left(\sum_j Z_j b_j\right)^p\right] \\ -\kappa\sum_j X_j, \end{aligned} \tag{16}$$

where the a_j and b_j are all equal to ± 1 and specify the minimum position for each of the two terms. We assume for the remainder of this writeup that a and b correspond to bitstrings $M = N/2$ flips apart; we let these states be $|0\rangle$ and $|1\rangle$. We will also define the bare classical energy, with no corrections from transverse fields. We start from one of the two minima, and consider a state which is $m+n$ random flips away from it. We let m of these flips be ones which move toward the other minimum (e.g. reduce the Hamming distance to it), and the n flips be flips that move away from it. Then the bare energy $E^{(0)}_{m,n}$ is given by:

$$E^{(0)}_{m,n} = -N\left[\left(1 - 2\frac{m+n}{N}\right)^p + \left(2\frac{m-n}{N}\right)^p\right]. \tag{17}$$

We assume that the transverse field strength κ is below κ_c, the p-dependent critical point where a transition to the paramagnetic state occurs. The ground states are thus the symmetric and antisymmetric combinations of the two dressed classical minima, with splitting $2\Omega_0$, where Ω_0 decays exponentially in N and our goal in this section is to predict its decay rate.

Computing Ω_0 proceeds through the following steps:

- We compute the renormalized cost per flip away from either minimum, incorporating transverse field corrections, which we will then use in the energy denominators of our Mth order perturbation theory. This step is analogous to commonly used resummation schemes in diagrammatic quantum field theory, where self-energy corrections are incorporated into the propagators used to compute higher order processes.

- We divide the system between primary spins, which flip between the classical minima, and secondary spins, which do not. We then compute the dressed states $|0_D\rangle$ and $|1_D\rangle$ that comprise all the primary flip sequences up to order $M/2$ away from each minimum. It is at this order that the two states have nonzero overlap.

- These dressed states are then normalized; incorporating this normalization, their overlap gives the primary spin contribution to the tunneling rate, $\Omega_0^{(p)}$.

- We then compute the secondary spin contributions to tunneling, which take two forms: an increase of the tunneling rate from the constructive contribution of many additional tunneling sequences in which secondary spins participate, and a decrease from normalization corrections and the spread of the classical minima away from the core classical configurations from which the tunneling calculation begins.

- Incorporating both sets of secondary spin contributions gives us a closed form expression for Ω_0 which can then be evaluated numerically and compared to exact diagonalization.

We first want to compute the energy shifts, in second order perturbation theory, to these states. These corrections arise from a single spin being flipped and flipped back, and are opposite in sign to the cost of the local flip. Let $u_{m,n}$ be the difference between energies of a state m, n and a ground state, incorporating these corrections. Then

$$u_{m,n} = E^{(0)}_{m,n} + N\left(1+\frac{\kappa^2}{2p}\right) - \left(\frac{N}{2}-2m\right)\frac{\kappa^2}{\partial_m E^{(0)}_{m,n}} - \left(\frac{N}{2}-2n\right)\frac{\kappa^2}{\partial_n E^{(0)}_{m,n}} + O\left(\kappa^4\right). \quad (18)$$

Note that $\partial_m E_{m,n}|_{m,n=0} = \partial_n E_{m,n}|_{m,n=0} = 2p$, so $u_{0,0} = 0$. One can observe that if $p = 2$, $\partial_m E_{m,0} = 4\left(1-\frac{4m}{N}\right)$, and the transverse field corrections to state energies are m-independent, so that the energy barrier between the two competing ground states is not renormalized by the transverse field. But for $p = 3$ and higher these corrections are nontrivial and act to reduce the effective energy barrier between the states, increasing the tunneling matrix element. This process is effectively a resummation of higher order corrections and is necessary to obtain quantitatively accurate results.

We start by computing the dressed states, summing over primary spin corrections only. They take the form

$$|0_D\rangle \equiv |0\rangle + \sum_j \frac{\kappa}{u_{1,0}} X_j|0\rangle + \sum_{j,k(j\neq k)} \frac{2\kappa^2}{u_{1,0}u_{2,0}} X_j X_k|0\rangle + \sum_{j,k,l(j\neq k\neq l)} \frac{3!\kappa^3}{u_{1,0}u_{2,0}u_{3,0}} X_j X_k X_l|0\rangle + ... \quad (19)$$

The expression for $|1_D\rangle$ is identical. We stop our expansion at order $M/2$, which is the lowest nontrivial order needed to connect the states. For simplicity we assume M is even though the argument is easy to generalize to odd M as well. Note that this state is not normalized, and in fact the norm of the state written above is exponentially large, so we will need to incorporate normalization corrections into the definition of the states. Thanks to the dressing of the states, we obtain a primary-spin energy splitting

$$\frac{1}{2}\left(\langle 0_D| + \langle 1_D|\right) H_P \left(|0_D\rangle + |1_D\rangle\right) = 2\Omega_0^{(p)},$$
$$\frac{1}{2}\left(\langle 0_D| - \langle 1_D|\right) H_P \left(|0_D\rangle - |1_D\rangle\right) = 0. \quad (20)$$

Evaluating these expressions, the degeneracy splitting from only considering primary spins is:

$$\Omega_0^{(p)} = \kappa^M \binom{M}{M/2} \frac{1}{u_{M/2,0}} \left(\frac{M}{2}! \prod_{k=1}^{M/2-1} \frac{1}{u_{k,0}}\right)^2 \times \mathcal{N}^{(p)},$$
$$\mathcal{N}^{(p)} = \left(1 + \sum_{k=1}^{M/2} \binom{M}{k} \left(\kappa^k k! \prod_{j=1}^{k} \frac{1}{u_{j,0}}\right)^2\right)^{-1}. \quad (21)$$

Here, $\mathcal{N}_p$ is the normalization correction. This covers the primary spin portion of the macroscopic quantum tunneling rate.

We now turn to the secondary spins. To introduce secondary spin corrections, consider a single secondary spin j out of the $(N-M) \sim N/2$ total, whose bit value is the same in both classical minima. Since the same transverse field is acting on it as all other spins, when we consider the sum of all processes that connect the two minima, we can now divide them between those where $M/2$ primary spins flip from each minima to meet in the middle, with secondary spin j unchanged, and a new set of processes where spin j flips starting from each minimum and the two wavefunctions overlap at the set of states where $M/2$ primary spins have flipped along with j. The first set of processes is what was considered in Eq. 21; the second is new, and we want to calculate its matrix element. It is most useful to express these matrix elements as a ratio of the new term to the original, primary-spin-only process, since both decay exponentially in M. Let the primary spin perturbative matrix element to reach $M/2$ flips be $\xi_{M/2}$, so that

$$\xi_{M/2} = \kappa^{M/2}\left(\frac{M}{2}! \prod_{k=1}^{M/2} \frac{1}{u_{k,0}}\right). \quad (22)$$

To define the analogous process where j flips, we need to sum over all the points during the perturbative sequence when that can happen. We thus have:

$$\xi^s_{M/2} = \kappa^{M/2+1} \frac{M}{2}! \sum_{n=1}^{M/2} \prod_{k=1}^{n} \frac{1}{u_{k,0}} \prod_{k=n}^{M/2} \frac{1}{u_{k,1}}. \quad (23)$$

And noting that we have to make this insertion in the matrix elements from both minima, the total tunneling term is increased by

$$\Omega_0^{(p)} \to \left(1 + \left(\frac{\xi^s_{M/2}}{\xi_{M/2}}\right)^2\right)\Omega_0^{(p)}, \quad (24)$$

$$\left(1 + \left(\frac{\xi^s_{M/2}}{\xi_{M/2}}\right)^2\right) \equiv \gamma_T. \quad (25)$$

We now need to consider the rest of the secondary spins. Formally of course, the additional energy cost of each secondary spin flip changes as more secondary spins flip, but since these corrections are fairly weak (though they are appreciable and necessary for an accurate prediction of the scaling exponent) the total tunneling rate is going to be dominated by the set of processes where a comparatively small fraction of secondary spins have flipped, and we can thus approximate them as independent contributions. In this limit, since there are $N-M$ secondary spins,

$$\Omega_0^{(p)} \to \Omega_0^{(p)} \gamma_T^{N-M}. \tag{26}$$

Alongside this, the secondary spin corrections also spread the competing ground state wavefunctions out over Hilbert space, which exponentially reduces the weight of the core classical configurations from which the tunneling calculation begins. To be consistent with the independence approximation made above, we simply compute all the corrections to the ground state from each secondary spin independently and multiply them. Noting that we must apply this calculation to both competing ground states, this reduces the tunneling rate by

$$\Omega_0^{(p)} \to \Omega_0^{(p)} \left(\frac{\gamma_T}{\gamma_R}\right)^{N-M}, \quad \gamma_R \equiv 1 + \left(\frac{\kappa}{u_{0,1}}\right)^2. \tag{27}$$

Note that, if we set the cost per flip of a given secondary spin to some constant U, independent of the configuration of the other spins, that would imply it is disconnected from the primary spins as there are no couplings to shift the energy. In this limit a direct evaluation of the two functions shows that $\gamma_T = \gamma_R$ (for any choice of U) and this now disconnected spin plays no role in tunneling at all. This factorization of disconnected spins is reassuring, and lends support to the correctness of this approach. Taking into account all these effects, our total tunneling rate is

$$\Omega_0 = \kappa^M \binom{M}{M/2} \frac{1}{u_{M/2,0}} \left(\frac{M}{2}! \prod_{k=1}^{M/2-1} \frac{1}{u_{k,0}}\right)^2 \times \mathcal{N}^{(p)} \times \left(\frac{\gamma_T}{\gamma_R}\right)^{N-M}. \tag{28}$$

Taking all of these effects into account yields a highly accurate prediction of the minimum gap scaling for a wide range of values for p and κ, with only an $O(1)$ discrepancy in the prefactor and few percent discrepancies in the scaling exponent (empirically, Eq. 28 tends to slightly overestimate the decay compared to the exponent extracted from numerical diagonalization). We refer to the Appendix A for more details.

D. Achievable approximation ratio with spectrally folded trial minimum annealing

With this result in hand, we will now predict the macroscopic quantum tunneling rate–and thus, achievable approximation ratio–for the spectrally folded Hamiltonian. From this, we can calculate our target value of A, assuming that there is a single deep minimum far below the energy of any local minima. Relaxing this assumption will improve the performance of the algorithm by virtue of there being many more target states, as confirmed in our simulations for lower constraint densities. We consider the protocol in section III, with an initial state $|L\rangle$. We assume that the process of ramping the transverse field up and down is itself at least roughly adiabatic, i.e., we can assume approximate spectral continuity with respect to the folding and lowering Hamiltonians, noting that the lowering Hamiltonian will itself create $O\left(\sqrt{N}\right)$ shifts to the energies of states near the fold. It follows from our assumption that the ramping process itself does not meaningfully heat the system. We then consider the set $\mathcal{T}$ of all states within $O(1)$ shifts of $-AE_{\rm GS}$ in H_P, the states closest to the fold, and compute, as a function of all our various algorithm parameters, the total probability of tunneling into any one of them.

Since the tunneling rate into any individual state is exponentially small, and the time over which we slowly turn off the lowering Hamiltonian is $T \propto O(N)$, we can assume that tunneling will be diabatic with respect to any individual state. A Fermi's Golden rule analysis as in [68] suggests that

$$P_{\rm tot} \propto \frac{T}{W} \sum_{j \in \mathcal{T}} \Omega_{0,Lj}^2, \tag{29}$$

where $W \sim O(N)$ is the energy range swept over by reducing $C(t)$ to 0, and $\Omega_{0,Lj}$ is the tunnel splitting at degeneracy between $|L_D\rangle$ and the target state $|j_D\rangle$, which we assume are an average of $\sim N/2$ flips apart. To go further, we need to compute the average value of $\Omega_{0,Lj}^2$, noting that while of course there will be substantial state-to-state variations, given that there are exponentially many states in $\mathcal{T}$ the average value is going to dominate Eq. (29). As in the previous calculation, the most important quantity here is the average cost per flip away from the typical ground state in the folded Hamiltonian, which remarkably turns out to be A-independent.

To see this, we start from the average cost per flip away from $|G\rangle$, given by Eq. (11), and note that we can invert that equation to find the mean number of flips $x_A N$ for which $\langle E(x_A N)\rangle = -AN$. To be specific,

$$x_A = \frac{1 - A^{1/3}}{2}. \tag{30}$$

We can therefore assume that the typical state in $\mathcal{T}$ is $x_A N$ flips away from $|G\rangle$. If we consider the sequences of primary spin flips connecting $|j_D\rangle$ and $|L_D\rangle$, the typical

flip sequence starts $x_A N$ flips away from $|G\rangle$, and notice that with probability $1 - x_A$ an additional random flip towards $|L\rangle$ will also move closer to $|G\rangle$. Taking all these effects and the division by A into account, so that the bare unperturbed energy of both $|j\rangle$ and $|L\rangle$ when they cross are both $\sim -N$, a bit of algebra shows that for y flips away from a ground state of the folded Hamiltonian, not only is the total average cost $\Delta E\,(y)$ A-independent, it is precisely equal to the cost given by Eq. (11).

This is again only an average, but noting that since it appears the denominators of equations like (28), variations about it are more likely to increase the tunneling rate than decrease it. And likewise, since H_L is a random 3-XORSAT problem itself the mean cost per flip away from $|L\rangle$ is going to be given by Eq. 11 as well, so Eqns. 16 through 28 can faithfully predict the average tunneling rate between $|L\rangle$ and a randomly chosen ground state of the folded Hamiltonian.[5] For more discussion of this approximation, see the appendix.

Of course, this rate decays exponentially; assuming the two states are $M \sim N/2$ flips away for $\kappa = 1.29$, $\Omega_0 \propto 2^{-bN}$ where $b \simeq 0.2$. But this is balanced by the fact that there are on the order of $\binom{N}{x_A N}$ target states. We can further note that out of these states, while the mean distance to $|L\rangle$ is $M \sim N/2$, ones which are k flips closer have tunneling rates which are larger by a factor of 2^{2bk} on average, and though those states are proportionally rare their increased weight is enough to meaningfully impact our choice of A. Since our total runtime is linear, simple diabatic scaling predicts that the probability of tunneling into the typical state $M - k$ flips away is proportional to Ω_0^2, e.g. $2^{-4b(M-k)}$.

We now take this result and plug it into Eq. (29), so that we can determine the choice of A where the returned P_{tot} provides an approximation guarantee. If we use Stirling's approximation to write the binomial coefficients as exponentials, and ignore slowly varying polynomial factors, the total number of states in $\mathcal{T}$ scales as:

$$N_{\mathcal{T}} \propto \exp\left(-\left[x_A \ln x_A + (1 - x_A)\ln\left(1 - x_A\right)\right] N\right). \quad (31)$$

Likewise, if the average probability of tunneling into a target state k primary spin flips closer to $|L\rangle$ is increased by a factor of at least 2^{4bk}, the weighted per-state average of the diabatic tunneling rate into states in $\mathcal{T}$ can be approximated as

$$\frac{\log\left\langle \Omega_0^2 \right\rangle}{-N} \approx 2b\ln 2 - x_A\left(\ln\left(1 + 2^{4b}\right) - \ln 2 - 2b\ln 2\right) \quad (32)$$

Note that this average comes from considering only $x_A N$ flips away from $|G\rangle$ but varying Hamming distance from $|L\rangle$ and thus neglects the influence of comparatively rarer states larger distances from $|G\rangle$. Taking all these terms into account, the probability of returning a state with $E \simeq AE_{\text{GS}}$ as measured relative to the original H_P becomes constant, or at least stops decaying exponentially, when

$$\begin{aligned} &-2b\ln 2 - \left[x_A \ln x_A + (1 - x_A)\ln\left(1 - x_A\right)\right] + \\ &x_A\left(-\ln 2 - 2b\ln 2 + \ln\left(1 + 2^{4b}\right)\right) = 0. \end{aligned} \quad (33)$$

The achievable approximation ratio is thus determined by the per-state decay exponent b, computed in section IV C as a function of N and κ by fitting Eq. (28) to $\Omega_0\,(N) \propto \sqrt{N}2^{-bN}$, and then choosing x_A using Eq. (30) to solve Eq. (33). This analysis only counts states within $O\,(1)$ shifts of AE_{GS} (recall that $E_{\text{GS}} = -N$ in our normalization) and ignores low-order polynomial prefactors; for $b = 0.2$, which again depends on $\kappa = 1.29$ in this calculation, this is solved when $x_A \simeq 0.08$, or $A \simeq 0.59$.

This means that if the true ground state satisfies a constraint fraction F beyond random guessing—e.g. a total fraction $1/2+F$, so F is at most $1/2$ here—our algorithm will return states which satisfy a fraction $1/2 + AF$ with high probability. If we choose A to be too large compared to the target value established by expressions like Eq. (33), we risk failing to well-approximate the problem; conversely, choosing A below it will reduce the returned approximation ratio to A and thus perform suboptimally. And, we emphasize again, this prediction assumes a random, potentially dense hypergraph but is fundamentally independent of the fraction satisfied in E_{GS} itself and so applies to the planted partial solution instances we use for numerical benchmarking below.

E. Further Comments and Caveats

We expect that this analysis underestimates the choice of A that will return states with $E \leq AE_{\text{GS}}$ with constant probability. This is because our counting here only counts states very close to the fold, when in reality the probability of tunneling into states a small extensive fraction larger than AE_{GS} is still going to be appreciable due to the continued exponential growth of the number of targets, even if the per-state tunneling rate does tend to decrease with increasing E due to the interference effect mentioned earlier, in which perturbative corrections that mix with states of lower energy have opposite sign. In addition to this consideration, because the target ground states and low lying excitations of the folded Hamiltonian in $\mathcal{T}$ very roughly form a hyperspherical shell, any individual target state will have other states in $\mathcal{T}$ that are relatively close to it in Hamming distance. Consequently we expect these states to have a band dispersion, centered around the mean energy given by the corrections in Eq. (18). Since we are already assuming off-resonant

[5] We expect that using the average cost per flip in Eq. 17 will if anything underestimate the per-state tunneling rate in real disordered problems. This is because all of these energy costs appear in denominators, which leads to likely small asymmetries in how much the deviations from the average in any individual flip sequence contribute to the total matrix element, giving lower energy sequences proportionally higher weight. We do not really expect this effect to be significant but rather highlight it as another point where our prediction is conservative by design.

tunneling, i.e. a per-state tunneling rate $\propto \Omega_0^2$, and so summing over squared matrix elements, if we consider states near the band center where the density is highest this will alter the average tunneling rate by at most a prefactor. However, there are good reasons to suspect that tunneling into extremal states near the bottom of the band can be substantially enhanced, enough to increase the optimal value of A. This calculation is difficult to do quantitatively so we do not attempt it here; we instead see the approximation of considering only states near the band center of $\mathcal{T}$ as another choice that likely underestimates the achievable approximation ratio. We discuss this point in more detail in Appendix B.

We also want to emphasize that this variation is not necessarily the optimal spectrally folded optimization algorithm, but instead merely the one where we were able to analytically compute the threshold A. For example, one can perform trial minimum annealing with a simple local Z bias lowering Hamiltonian (e.g. $H_L = \sum_j h_j Z_j$), or standard AQC interpolation using the quadratic folding procedure in Eq. (5) as the cost function. The linear lowering Hamiltonian is expected to have equal or better tunneling rates to a 3-XORSAT-based minimum as the overall cost-per-flip curve is shallower, though the local energy shifts to the ground states of H_{fold} from H_L are expected to be larger. The total gate count at each time step is lower. Empirical performance in testing up through $N = 25$ showed fairly similar performance to 3XOR-based H_L for all other parameters equal, but with more significant non-monotonic behaviors that made fitting difficult; see Sec. V for details. That the two schemes could asymptotically converge to the same achievable approximation ratios seems plausible to us but we cannot simulate large enough systems to be sure.

For quadratic folding AQC as in Eq. (6), if we choose $A = 1$ the gap is efficiently computable using the methods in [86] and decays as $\Omega_0 \sim 2^{-0.16N}$. Given that, like linear folding, the cost per flip curve is A-independent, if we assume that the tunneling rate per state for $A < 1$ is basically equal to this, then the total decay exponent vanishes if $x_A \simeq 0.06$ and $A \simeq 0.68$ using the arguments of the previous few paragraphs. We do not think that can be simply assumed as easily as with tunneling between semiclassical minima and a linearly folded problem Hamiltonian, in the DPP, and more theoretical work is needed here to analytically determine the optimal choice of A. Interestingly however, our simulation data in Sec. VI supports this conclusion, with a worst case polynomial time approximation ratio of 0.7 found in our simulations. These simulations show that this method performs similarly too, or slightly worse than, the 3XORSAT-TMA algorithm, which is better able to outperform the approximation guarantee of ~ 0.6 derived here.[6] We also expect that the average tunneling rate–and thus, achievable approximation ratio–can likely be further increased by using other, potentially many-frequency, AC methods such as RFQA [25, 49, 68, 88, 89]. For simplicity, we do not incorporate these methods in this work, but they could be a novel way to further improve the performance of this algorithm and are worth exploring in future research.

In summary, through a relatively novel resummed extensive order perturbation theory, we have shown that random hypergraph MAX-3-XORSAT instances, including extremal ones with planted partial solutions, are efficiently approximable to a fairly large constant fraction through the spectrally folded quantum optimization algorithms. We do not expect this to be the case for QAOA, an expectation supported by the numerical evidence we present below. We similarly do not expect such guarantees to be possible for directly finding global optima, for the reasons set forth in the introduction. Evidently, one cannot so easily summit a mountain in hyperspace, but one can reach the rim of a crater. We now present a series of numerical simulations to further support these claims.

V. NUMERICAL TESTS OF APPROXIMATION HARDNESS FOR TRADITIONAL METHODS

A. Setup and summary of results

To confirm our predictions–or at least, verify that any serious issues with our calculations and interpretation of the problem are subtle and not apparent at system sizes within reach of present or near-future classical simulations–we performed a series of numerical simulations of various classical and quantum algorithms applied to our PPSPs. For all quantum simulation tasks we used the Qulacs package [90]. For smaller systems and algorithm prototyping we ran our simulations on local workstations; this includes all the spectral folding TMA simulations. For all QAOA and spectral folding AQC simulations presented, we used the Fujitsu Quantum Simulator, a classical HPC system. This allowed us to probe larger system sizes while still averaging over enough instances to have reliable statistics for the problem class. Unless otherwise stated, each datapoint represents the average over 960 or 1000 (for the QAOA and spectral folding AQC results) random problem instances constructed with the prescriptions outlined in section II D. In all cases in this work we used an unsatisfied fraction $\epsilon = 0.1$ for our partial planted solutions; we expect similar phenomenolog-

[6] For smaller systems we also tested linear folding AQC and quadratic folding TMA. In very preliminary studies we found that quadratic AQC modestly outperformed linear AQC, and linear TMA more significantly outperformed quadratic TMA. So we chose not to pursue those methods for larger simulations and do not present those results here, but they may be viable or even superior for other problem classes.

ical behavior for other constant ϵ values, though we expect the precise thresholds we measure will vary with ϵ for classical optimization algorithms and QAOA (but not folded optimization, below its predicted performance floor).

The results of our simulations are summarized in Table I, which lists the best per-shot polynomial time approximation ratio achievable through each algorithm studied, both classical and quantum. To estimate these values, we measured the average per-shot probability $P_q(N) = P(E \leq qE_{\mathrm{GS}})$ of returning states with energies at or below qE_{GS}, where $q \leq 1$ is the approximation ratio and E_{GS} is negative in our conventions. These probabilities were computed by choosing bins of size $0.05\,E_{\mathrm{GS}}$.

To determined the polynomial time hardness threshold, we applied a very simple rule where we fitted $P_q(N)$ to a simple exponential function, assuming any observed decay faster than $2^{-0.005N}$ corresponded asymptotically to exponential decay, and any positive exponents, e.g. exponential growth, represent small-N growth toward some constant saturation value. This threshold of $2^{-0.005N}$ was chosen to reduce the influence of uncertainty from fitting in a small handful of cases. For the greedy algorithm with $N_C/N = 1.5\sqrt{N}$ we fitted decay to an exponential in $\sqrt{N}$ as discussed below. The polynomial time approximation hardness threshold for a given algorithm, unsatisfied fraction ϵ and N_C/N scaling choice is defined to be q_a, the largest values of q for which we do not observe exponential decay. For $N_C/N = 3\sqrt{N}/2$ this threshold is decaying with system size, and the asterisk next to the result for QAOA is to highlight the fact that we only ran these simulations out to $N = 28$ so are likely not capturing the asymptotic threshold. As discussed elsewhere, the number of cost function calls is $O(N)$ for all approaches studied. Cases for which a range is quoted are where we felt there was some ambiguity to the fitting, and all values are the result of extrapolating fits to numerical simulations and are naturally somewhat approximate.

B. Performance of greedy classical algorithms

To explore the classical difficulty of our PPSPs, we applied a greedy local search algorithm adapted from [12]. This algorithm is straightforward; we start with a random bitstring. Then, beginning at each step, we calculate $k =$ "number of unsatisfied constraints minus the number of satisfied constraints" associated with each bit. We then calculate the fraction of bits, f_k, belonging to each k value. Note that we only care when $k > 0$ because these are the cases where flipping a particular bit will lower the energy. Using some weight function, $w(k)$, we select a k value with normalized probabilities $\propto w(k) * f_k$ and flip a bit with that k value. If there are multiple bits belonging to the k value chosen, we choose a bit in this set with uniform probability to flip. This is repeated until the configuration finds itself in a minimum, and the algorithm halts.

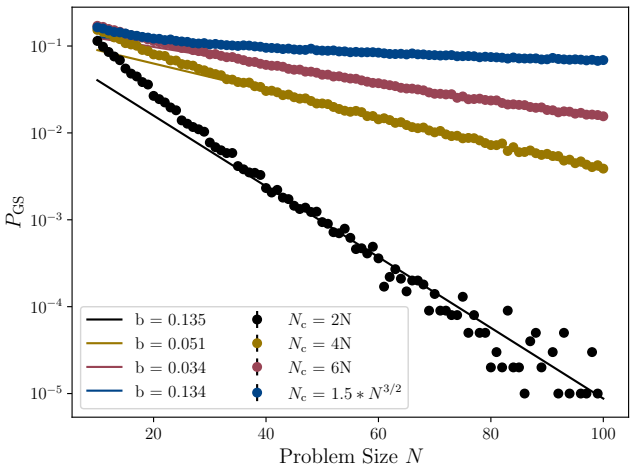

FIG. 4: Per-shot probability of finding the true ground state with the quasi-greedy classical algorithm defined in section V B as adapted from [12], for the four constraint densities studied in this work and planted solution unsatisfied fraction $\epsilon = 0.1$. The probability decays superpolynomially with N; fits are to $a2^{-bN}$ for $N_C = \{2, 4, 6\}\,N$ and $a2^{-b\sqrt{N}}$ for $N_C = 1.5N^{3/2}$; the scaling fairly well matches the empirically observed Eq. (34). Each datapoint is the average of 10^5 shots. As the constraint density increases, the basin of attraction widens, and the problem becomes easier–though still exponentially scaling–for local update classical routines. Each shot consists of $O(N)$ local updates so the time to solution scales essentially as the inverse of this probability.

We found the algorithm performed best with a weight profile quadratic with k, however this can be experimented with for different results. Notably, [12] found that when applying a highly optimized version of this search to 3-regular 3-XORSAT problems it performed well even when compared to more sophisticated algorithms such as simulated annealing and parallel tempering. The intuitive reason for this can be inferred from the typical energy landscapes of these problems, which are rough and contain exponentially many high energy local minima. Once one is found, it is more efficient to simply restart the algorithm from a new random configuration instead of attempting to "climb out" using penalized operations in simulated annealing or parallel tempering. As the locations of these minima are uncorrelated with the true ground state, finding one provides no useful information in a ground state search.

This expected inefficiency of simulated annealing/parallel tempering for this problem can easily be inferred from the results plotted in Figs. 4 and 5. Namely, in all cases the algorithm will find a single relatively deep minimum with high probability at each shot, leading to a super-polynomial cost to escape from it in algorithms simulating a thermal bath. Interestingly, as N_C/N increases for fixed unsatisfied ground state fraction ϵ, we find that the decay exponent of the per-shot probabil-

N_C/N	Classical	QAOA	AQC/0.75	AQC/0.85	TMA-3/0.75	TMA-3/0.85	TMA-L/0.75
2	0.75-0.8	0.75	0.75	0.75	0.75	0.75	0.7
4	0.55	0.55	0.7	0.7	0.75	0.8	0.75
6	0.45	0.45	0.75	0.7	0.8	0.8	0.75
$3\sqrt{N}/2$	decaying	0.25/decaying*	0.75	0.75	0.8	0.8	0.75-0.8

TABLE I: Approximation hardness thresholds for the classical greedy search, high-depth QAOA and spectral folding variations. This table lists q_a, the largest value of the approximation ratio q before exponential decay is reported, drawn from the numerical experiments in figures 5 through 8 and 14. Spectral folding results are labeled as protocol/A (where A is the approximation target); AQC is quadratic spectral folding in the AQC formulation, TMA-3 is trial minimum annealing with a linear folded Hamiltonian and 3-XORSAT lowering Hamiltonian, and TMA-L is the same with local Z biases for the lowering Hamiltonian. These results support the predictions in section IV that random hypergraph problems are efficiently approximable through spectrally folded quantum optimization.

ity of finding the ground state, $P_{\mathrm{GS}}(N)$, monotonically decreases, suggesting that the basin of attraction of the true ground state is widening as the problem becomes more extremal. In fact, for $N_C \geq 2N$, we empirically observe that the per-shot probability of finding the planted ground state has the approximate scaling

$$\log\left(P_{\mathrm{GS}}(N)\right) \simeq -c_g \frac{N^2}{N_C}. \quad (\text{PPSPs}, N_C \geq 2N) \quad (34)$$

However, the approximation ratio q_a–defined as the minimum energy for which the probability of finding states at or below it stays constant as N increase–steadily worsens. We attribute this to there being a high density of local minimia with energies $\geq q_a E_{\mathrm{GS}}$ (recall E_{GS} is negative in our conventions), but below that threshold the number of minima quickly decreases and the probability of finding one decreases exponentially. This results in the scaling collapse seen in Fig. 5–for sufficiently low energy there are no minima aside from the ground state, so the approximation probability scales nearly identically to $P_{\mathrm{GS}}(N)$. This *high energy clustering phase* is a feature of our PPSP construction, and is responsible for its classical approximation hardness. We again contrast this to problems near the statistical SAT/UNSAT threshold such as three-regular instances, where the *clustering energy*, the lowest energy where there are still exponentially many local minima and they are thus easy to find, is close to E_{GS} and they are not approximation-hard in practice as a result. We conjecture that high energy clustering behavior is a generic feature of low-degree constraint problem classes that are approximation-hard for local update algorithms.

As shown in Fig. 5, these construction rules yield in a set of instances which are hard to approximate in practice. If we let N_C/N grow slowly with N, e.g. as $\ln(N)$ or $\sqrt{N}$, then as $N \to \infty$, the probability of finding any states with energies any $O(1)$ fraction better than random guessing, decays superpolynomially in N. And since the unsatisfied fraction ϵ in the ground state is small but nonzero, Gaussian elimination cannot be used to efficiently find the solution, forcing classical computers to rely on local update algorithms stymied by entropic barriers. It is of course possible that some clever algorithm could be written to exploit our PPSP structure to efficiently solve or approximate these instances classically; we merely claim hardness for generic methods based on local updates. Our PPSP construction rules can easily be generalized to other CSPs, and we suggest that they could prove to be a useful tool for exploring practical approximation hardness in other contexts.

C. Performance of high-depth QAOA for this problem

To make firm points of comparison, alongside the simulations of spectrally folded optimization itself we extensively benchmarked high-depth QAOA and quasi-greedy [12, 66] classical algorithms on these instances. For our high-depth QAOA simulations, we formulated our algorithm to mimic trotterized time evolution over a total time t_f, with:

$$|\psi(t+dt)\rangle = e^{-2\pi i f(t) dt H_D} e^{-2\pi i g(t) dt H_P} |\psi(t)\rangle, \quad (35)$$
$$f(t) = \sqrt{1 - t/t_f},\ g(t) = \sqrt{t/t_f}.$$

In all the presented data we used $t_f = N/32$ and $dt \simeq 0.05$. Individual shots use a random evolution time between $2t_f/3$ and $4t_f/3$; this runtime averaging produces substantially more reliable scaling, particularly when probabilities are small. These parameters were chosen by trial and error for smaller systems; we observed that the probabilities of finding the ground state and other low-energy states increased sublinearly with t_f beyond this point. The relative improvements of the probabilities of returning low-lying states were similarly sublinear. No sophisticated numerical or iterative optimization methods were used, for the reasons discussed in section II E.

The results of our QAOA simulations, of 1000 random PPSPs for each choice of N running from 8 to 30, are shown in Fig. 6. The probability of finding the ground state, shown in Fig. 11, decays exponentially with an exponent very close to that in Eq. (15), which we find remarkable given the simplifying assumptions in that derivation, and that it does not use the more sophisticated techniques used to compute tunneling rates between semiclassical minima. Further, this exponent dis-

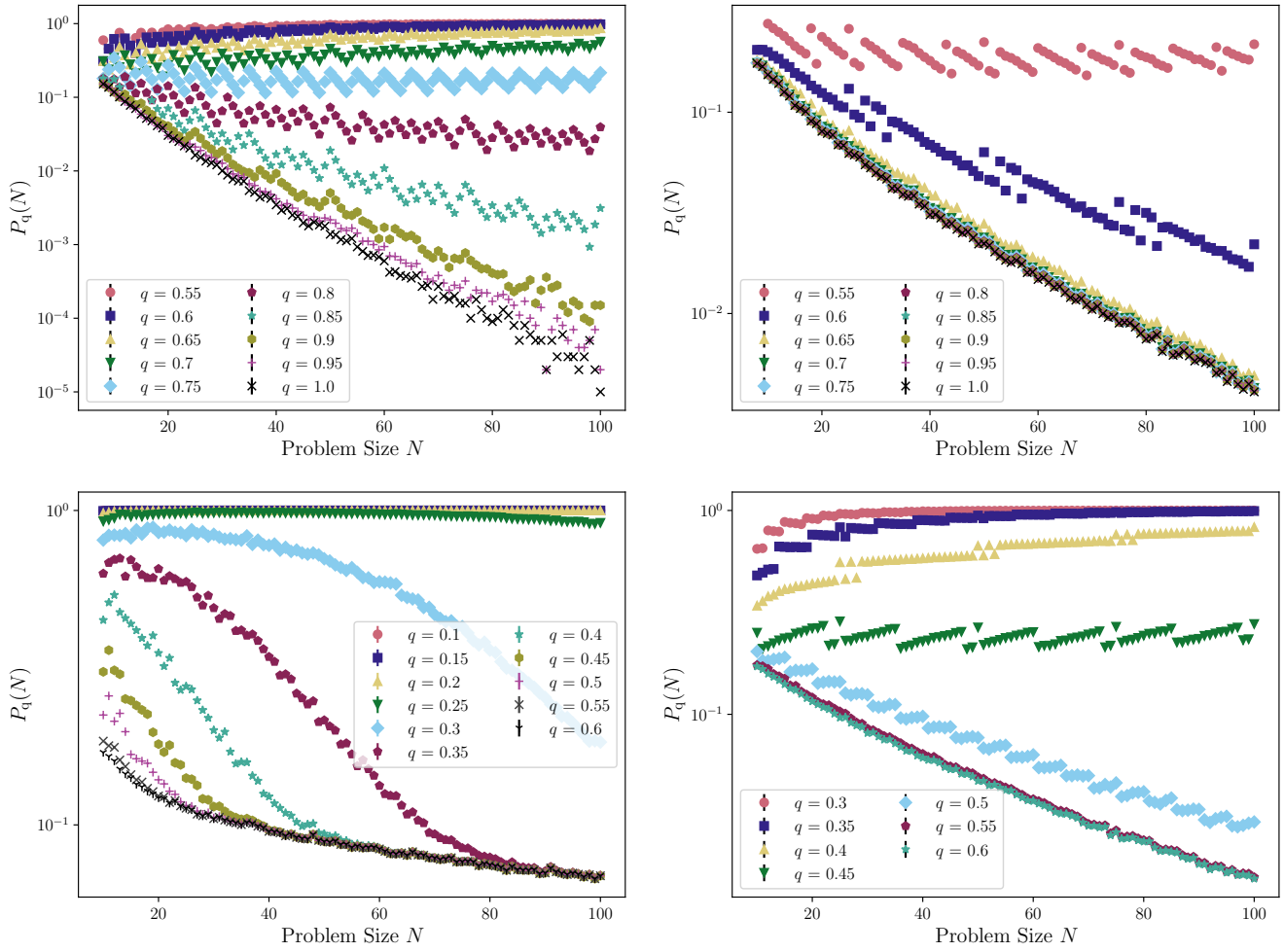

FIG. 5: Classical approximation hardness of PPSPs using the quasi-greedy classical algorithm, for constraint densities (clockwise from top left) $N_C = \left\{2N, 4N, 6N, 1.5N^{3/2}\right\}$. In each figure we plot the probability that a given shot returns an energy below qE_{GS} for various choices of q. For all the fixed N_C/N problem classes we find an empirical approximation threshold $q = A_g$ below which finding states becomes superpolynomially hard, and that this value decreases as N_C/N increases. For the case $N_C = 1.5N^{3/2}$ (bottom left), this value steadily drops, as discussed in the text.

plays only small variations with constraint density and is nearly identical in all four cases. Smaller system studies for other constraint densities all yielded very similar results for P_{GS}, as predicted by Eq. (15).

Turning to approximation hardness, being relatively sparse, the $N_C = 2N$ problems are fairly well-approximated by QAOA, with the algorithm returning strings within $q = 0.75$ with constant or saturating probability; we attribute this to the presence of many competing minima with energies not far from E_{GS}. In contrast, for $N_C = 4N$ the algorithm's performance for approximation degrades, with clear exponential decay for approximation ratios better than $q = 0.55$. For higher constraint densities approximation becomes even more difficult, decaying exponentially below $q = 0.45$ for $N_C = 6N$ and 0.25 for $N_C = 3N^{3/2}/2$. We expect decay at sufficiently large N for any constant fraction in that case, but cannot simulate larger system sizes. Crucially, the thresholds q_a we measure are nearly identical to those found by the classical greedy algorithm (figure 5), and no signatures of an exponential quantum advantage in these instances can be seen.

Interestingly, as N_C/N increases, the probabilities of finding states comparatively close to $|G\rangle$ in Hamming distance improve (see Fig. 11), but the probabilities of finding states close in energy worsen. We attribute this behavior to the high energy clustering phase conjectured in Sec. V B. Empirically for our PPSPs there is a high density of local minima with energies $E \geq q_a E_{\mathrm{GS}}$, and if q_a is relatively close to 1 it becomes harder for high-depth QAOA to find local excitations near the planted ground state, as the probability amplitudes will be spread over increasingly many competing minima and their own basins of attraction. Conversely, as q_a decreases with

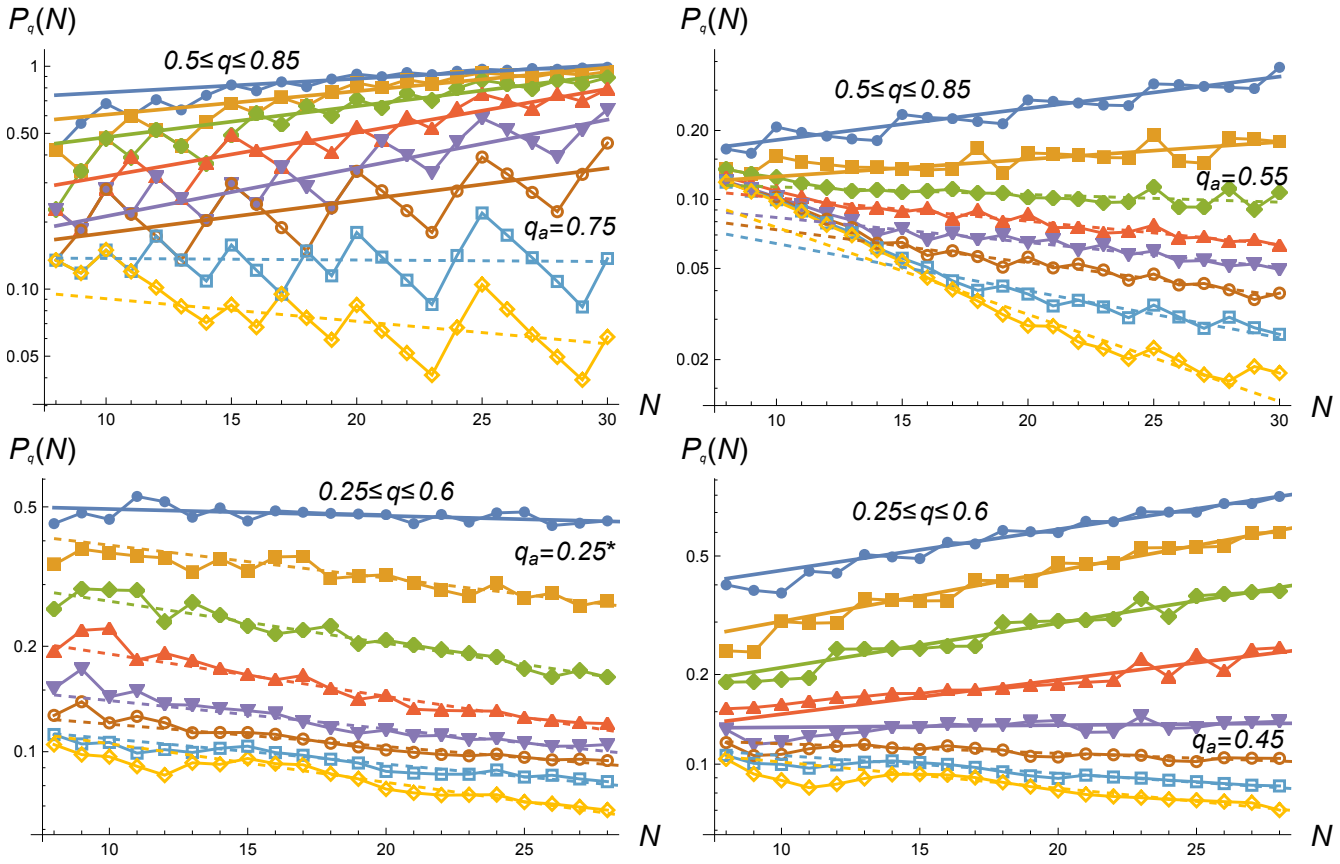

FIG. 6: Performance of QAOA as an approximator for $N_C = \left\{2, 4, 6, 3\sqrt{N}/2\right\} N$ (clockwise from top left), with problem and algorithm parameters described in-text. Plotted are the probabilities $P_q(N) = P(E \leq qE_{\mathrm{GS}})$ for q running from 0.5 to 0.85 in steps of 0.05 (top panels) and 0.25 to 0.6 (bottom panels). The results for the larger constraint densities thus plot a weaker approximation range. Thick straight lines correspond to simple exponential fits where $P_q(N)$ is not decaying, dashed straight lines correspond to exponential decay, and thin lines between points are included for visual clarity. As summarized in table I, these results do not represent a meaningful improvement over the classical greedy result (Fig. 5), though in some cases where $P_q(N)$ decays exponentially for both approaches, the exponent for QAOA may be better. These results were obtained using the Fujitsu Quantum Simulator, a classical HPC system.

increasing N_C/N, the probability of finding local excitations relatively near $|G\rangle$ increases though still decays exponentially, as the low energy minima far from $|G\rangle$ are at proportionally higher energies and thus do not compete directly with few-flip states. That the thresholds q_a for QAOA match those of the greedy classical algorithm further supports the interpretation of an energy threshold above which local minima become common.

Comparing the classical and established quantum methods, we find that, for these approximation hard instances, QAOA performs very poorly for finding the ground state but less poorly for approximation below the classical hardness threshold q_a (see Table I), with decay exponents that are much closer to the classical result. In some higher constraint density cases our fits produced favorable exponents for approximation with QAOA but our range of N here is smaller than we would prefer to claim any relative quantum advantage absent theoretical justification. Nonetheless, both methods show clear super-polynomial decay per-shot for approximate optimization below the energy range where local minima are dense. With these results in hand, we now turn to folded quantum optimization, which maintains an approximation guarantee regardless of the problem's constraint density.

VI. NUMERICAL TESTS OF SPECTRAL FOLDING VARIATIONS

Having numerically confirmed the expectation that our PPSPs are superpolynomially hard for classical and prior quantum approaches, for both exact and approximate optimization, we now present the results of our folded quantum optimization simulations. We simulated both the trial minimum annealing and interpolation (e.g. AQC) variations. In all cases we chose runtimes increased linearly in N, albeit with larger prefactors than in the QAOA simulations (which used $t_f = N/32$). A longer runtime further helps reduce the potential influence of

diabatic local heating as the Hamiltonian parameters are varied. Run and ramp times that are too short can lead to artificially poor scaling, arising from the formation of local excitations as the transverse field is turned on or off too quickly. This is fundamentally a different, and much more prosaic, issue, than decaying collective tunneling rates, but can be difficult to distinguish when our only measures are energy and Hamming distance from $|G\rangle$.

A. Spectrally folded adiabatic interpolation performance

We first present simulations of the AQC variation of spectral folding, in figures 7 and 12. For the AQC variation, we followed the procedure in Eq. (35), with a quadratic H_{fold} (Eq. 5) in place of H_P and $f(t) = (1-t/t_f)^{1/4}$ instead of $\sqrt{1-t/t_f}$, with an average runtime $t_f = N/24$. This schedule modification was found to improve scaling at higher approximation ratios. In all cases individual shots are runtime averaged between $2t_f/3$ and $4t_f/3$ as in our QAOA simulations.

All these parameter choices are the result of intuition, trial and error, and a desire for simulations to complete in reasonable amounts of time; none are particularly optimal. Nonetheless, as shown in figure 7, in all but one case we were able to meet the approximation target $A = 0.75$, but not exceed it; for $N_C = 4N$. Our spectral folding methods are also much better at reliably returning states close to G in Hamming distance, consistent with the approximation guarantee (see Fig. 12). We likewise tested increasing A to 0.85 in simulations up to $N = 24$ with this variation, and found improved prefactors but no improvement in scaling (see Fig. 14). This suggests that we have found the performance ceiling for this approach. Interestingly, the worst case $q_a = 0.7$ observed for this method is very close to the approximation ratio of ~ 0.68 predicted in section IV E, using a more simplified analysis than was employed for the TMA variation.

B. Trial minimum annealing performance

We also tested the TMA formulation of spectral folding–the formulation for which we can make the most reliable analytical predictions–as plotted in Fig. 8. For these variations we used runtimes $t_f = N/12$ and $dt = 0.025$; note that this choice of t_f is a factor of 8/3 larger than in our QAOA simulations but with the same scaling. In smaller system studies similar qualitative performance was observed for shorter t_f (such as $N/24$ or $N/32$). For this variation we used a 3-XORSAT H_L–the formulation for which we could predict performance in Sec. IV–with the minimum energy set to $-2N$ via $C(t)$, which was linearly ramped down to zero by t_f, and simple sinusoidal ramp profiles with $t_r = N/24$; the transverse field strength κ during the main evolution was 1.3. The careful reader may note that choosing $C(t)$ to set the minimum energy to $-2N$ instead of $-N$ is naively suboptimal, as all other things being equal sweeping over a larger energy range increases W in Eq. (29), and should reduce the returned probabilities $P_q(N)$ by an appropriate prefactor. However in our simulations this choice consistently improved both the prefactors and scaling, e.g. the value of q_a, as compared to choosing a minimum energy $-N$ for H_L. We suspect this has to do with the band structure considerations described in Appendix B but due to the complexity of the problem, are unable to make a quantitative prediction.

The performance of the two approaches is qualitatively similar with subtle differences as we vary the returned approximation ratio q. At lower approximation ratios the AQC formulation returns higher probabilities, at lower total gate count since there are no ramping steps and no additional gates associated with adding H_L. However, at higher approximation ratios the TMA formulation appears to be better able to approach the approximation target $A = 0.85$. As discussed in the algorithm definition, folded optimization will definitionally fail to consistently return energies significantly below AE_{GS}, and we expect it to break down as A gets too close to 1 given that QAOA and similar methods fail to reliably approximate these problems. Choosing the best value for A is thus a subtle issue that depends on the problem class; for extensions of this method to hard CSPs it will necessarily change from one problem class to the next.

Likewise, as mentioned in section IV E, one can replace the random 3-XORSAT lowering Hamiltonian in TMA for a simpler set of linear Z biases, reducing the gate count per timestep and, potentially, increasing the per-state tunneling rate by implementing a shallower cost-per-flip curve. In comparison to the 3-XORSAT variation discussed in the previous paragraph, to achieve good performance we needed to double the ramp time. This protocol seemed to be more sensitive to performance degradation from heating during ramps. The algorithm also benefitted from adjusting $C(t)$ so that the minimum energy of H_L was $-3N$, as compared to $-2N$ for the 3-XORSAT H_L. Relative performance for $A = 0.75$ is comparable to the other variations, as illustrated in Fig. 9, though the individual $P_q(N)$ show more significant non-monotonicity that makes reliable curve fitting challenging. This issue is even more pronounced for $A = 0.85$ (data not shown), to the point that we did not quote q_a values for that variation in the Table I.

C. Discussion of our spectrally folded optimization results

The reader may note that all of these simulations targeted energies in the range $A = 0.75$ to $A = 0.85$, well above the theoretical prediction of ~ 0.6. We also tested $A = 0.65$ at smaller scales (data not shown) for both formulations but found no region of the parameter space where it showed meaningful improvements over choosing

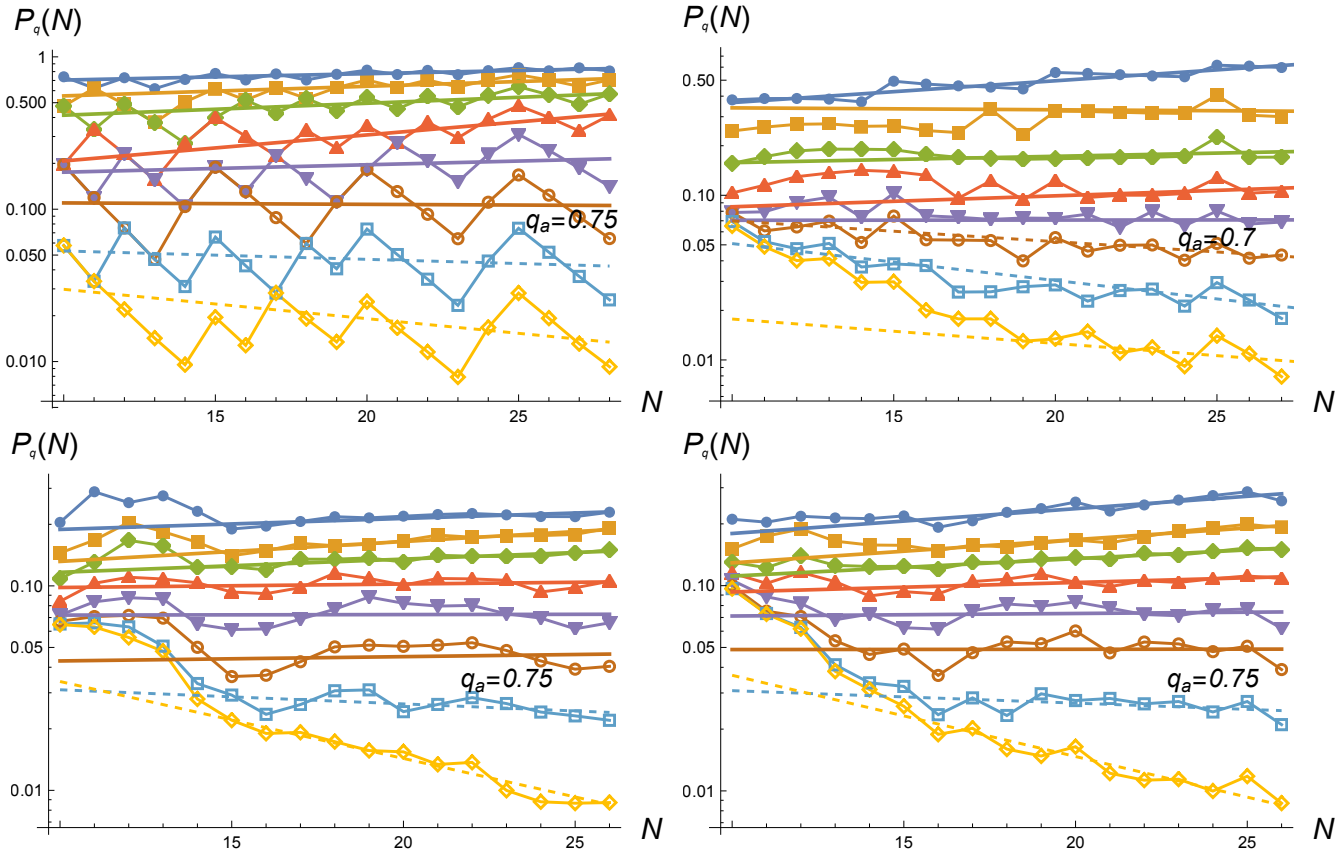

FIG. 7: Performance of the quadratic AQC formulation of spectral folding, for $N_C = \left\{2, 4, 6, 3\sqrt{N}/2\right\} N$ (clockwise from top left), with N running from 10 to 27 (top row) or 26 (bottom row), $A = 0.75$, $dt = 0.0325$, $t_f = N/24$ and other parameters as stated in text. In each plot the 8 curves plot $P_q(N) = P(E \leq qE_{\rm GS})$ for q running from 0.5 to 0.85 (top to bottom) in steps of 0.05. Thick straight lines correspond to simple exponential fits where $P_q(N)$ is not decaying, dashed straight lines correspond to exponential decay, and thin lines between points are included for visual clarity. In all four cases spectrally folded optimization is able to meet its approximation target of $A = 0.75$, returning states at or below this energy with constant probablity in a linearly growing number of cost function calls. This is in stark contrast to our classical greedy algorithm (FIG. 5) and QAOA (FIG. 6) results, where the achievable polynomial time approximation ratio steadily worsens with increasing N_C/N, and supports the theoretical analysis of section IV. These results were obtained using the Fujitsu Quantum Simulator, a classical HPC system.

$A = 0.75$ or higher. When increasing A to 0.85 we were not able to meet this target with constant probability, though for some cases we did see improvements in the returned achievable approximation ratio q_a when compared to $A = 0.75$. We did not run tests with $A > 0.85$ due to system size constraints, though comparing performance between $A = 0.75$ and $A = 0.85$ in table I suggests we have found the performance ceiling for these methods. Given the comparatively small system sizes available to classical simulation our ability to draw meaningful scaling distinctions with very small changes in A is limited. Considering the results of all our numerical simulations, the minimum achievable approximation ratio for our PPSP problem classes is at least 0.7 (often 0.75) for the quadratic AQC variation, and at least 0.75 (often 0.8) for the 3-XORSAT TMA variation.

As summarized in table I, the contrast between the clear super-polynomial decay of higher approximation ratios with classical methods and QAOA (Figs. 4-6), and the constant probabilities returned by folded quantum optimization (Figs. 7-9) is stark. In particular, for the established methods, the polynomial time approximation threshold is set by the problem structure, specifically the relative energy of the high energy clustering phase, and as our PPSPs become more extremal this threshold steadily worsens as a fraction of the energy of the planted ground state. We found no evidence for an exponential separation in approximation power between local classical searches and QAOA/AQC. At lower constraint densities, the clustering energy is not far above the ground state and spectrally folded quantum optimization provides no benefits for approximation, though it can still show scaling advantages for returning states close in Hamming distance to the ground state (see appendices). At higher constraint densities however, spectral folding is able to return states close to G, in both energy and Hamming distance, consistently and without degradation as N_C/N increases. The performance of all spectral folding vari-

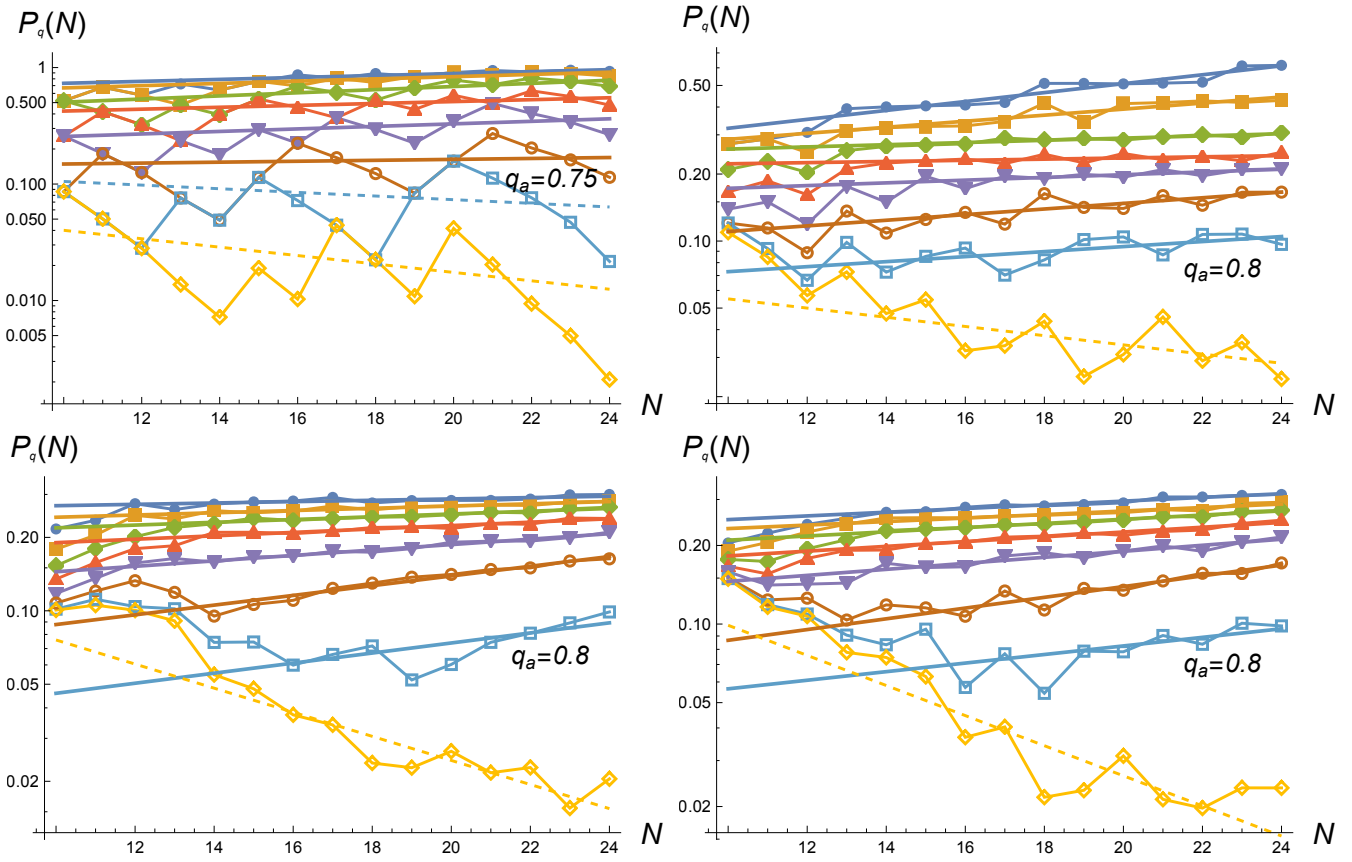

FIG. 8: Performance of the trial minimum annealing formulation of spectral folding with a 3-XORSAT lowering Hamiltonian H_L, with $A = 0.85$, plotted for constraint densities (clockwise from top left) $N_C = \left\{2N, 4N, 6N, 1.5 \times N^{3/2}\right\}$ and approximation ratios between $q = 0.5$ and 0.85 with N running from 10 to 24, for the parameters detailed in the text. All data is derived from averaging over 960 random instances and choices of t_f. Compared to the AQC formulation shown in figure 7, the achievable approximation ratio q_a is often slightly higher, though the total gate count in this formulation is larger by a constant prefactor. In all cases q_a well exceeds the value of 0.6 conservatively predicted for this formulation.

ations tested was broadly similar, with the 3-XORSAT TMA variation returning the highest approximation ratio but at the highest prefactor cost in gate count.

This illustrates the fundamentally different structure of collective tunneling in this approach. Where high depth QAOA, AQC and other direct methods attempt to find the ground state, asymptotically fail (given exponential scaling), and return approximate states passed after this missed opportunity, spectrally folded optimization deforms the cost function to search for states in an exponentially large hyperspherical shell, avoiding the interference and weakening field issues that are qualitatively responsible for QAOA's lack of obvious quantum advantage. In the interest of fair comparisons the total runtime per-shot of all routines is $O(N)$ Hamiltonian calls. With this simple linear scaling we are able to ensure that any exponential scaling of the time to solution not visible from our figures or fits must involve very small exponents, though of course we cannot rule out such behavior on numerics alone. We thus do not see our numerical simulations as a precision scaling benchmark of folded quantum optimization, but rather a test of the basic veracity of our conservative analytical predictions in section IV. The results of our simulations, for a wide range of parameters, all suggest that our theory is sound.

D. Extensions of these results

Having thoroughly explored MAX-3-XORSAT in this work, it would be interesting to test spectral folding methods on other hard CSPs or problems that can be straightforwardly formulated as such. We expect that the derivation of the achievable approximation ratio through the overlap of dressed states could generalize with some modifications to many other problem classes. We focused on MAX-3-XORSAT due to the simplicity of its structure, classical approximation hardness, and the fact that its exponential difficulty scaling is obvious at small N for more standard classical and quantum approaches.

It also strikes us as noteworthy that while spectral folding can be implemented in traditional classical heuristics,

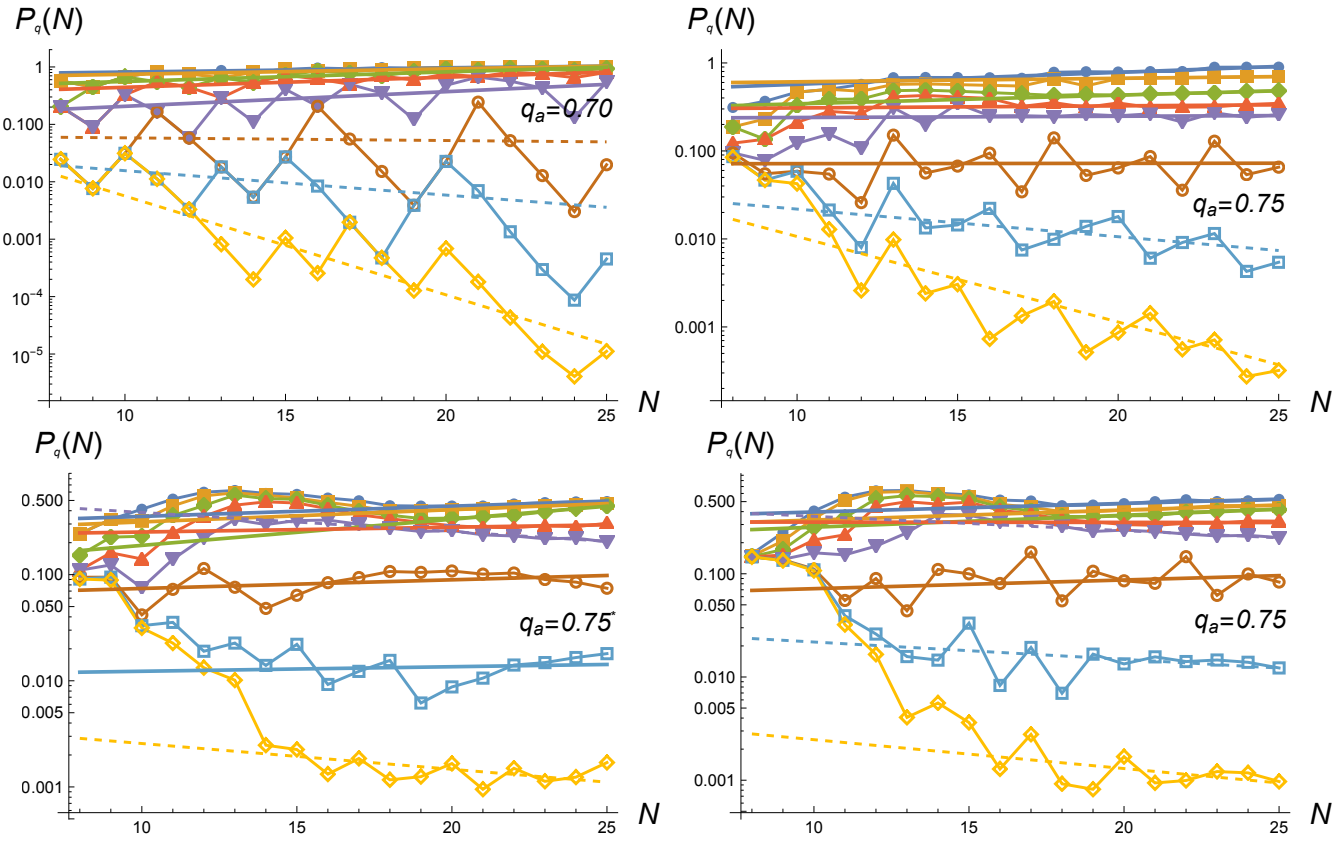

FIG. 9: Performance of the trial minimum annealing formulation of spectral folding with a local Z bias H_L, with $A = 0.75$, for N running from 8 to 25. q_a in each case is comparable to other variations, though greater non-monotonicity in the individual $P_q\,(N)$ curves makes fitting more difficult.

there is no benefit to doing so. As we discussed at length in Sec. II, classical algorithms generally start from high energy states and attempt to cool towards the ground state through local updates. In the case of linear spectral folding, the folding procedure makes no difference whatsoever to the returned energy until the fold energy has been reached, and in the worst case reaching that energy is an exponentially difficult task for classical computers unless P=NP, due to entropic barriers as discussed in Sec. II and confirmed in our simulations. Changes to the high energy spectrum from quadratic folding are similarly not expected to make approximation easier. So while spectral folding is not in and of itself a quantum operation, we expect that it is only valuable in quantum algorithms and we consider it an irreducibly quantum method as a result. Further, because of its nonlocal nature, even in the locally gapped dressed problem phase we expect volume-scaling entanglement in low-lying states of the folded quantum spin glass, and significant classical simulation difficulty.

One interesting potential exception to this argument is quantum Monte Carlo. Being a stoquastic problem, a folded quantum spin glass can in principle be efficiently simulated using QMC [91–95], which for a uniform field has in some cases proven to be an effective quantum-inspired classical solver [26, 96]. Incorporating spectral folding into these algorithms is possible, though the loss of locality makes evaluating each update much more expensive. As QMC is, fundamentally, an energy-based classical local update rule combined with many replicas, we do not expect it to overcome the entropic barrier in MAX-3-XORSAT the way that true quantum evolution can.[7] All that said, there may be some narrow cases where incorporating spectral folding in a QMC calculation could prove useful as a classical solver; this would be an interesting avenue for future research. If it turns out that many or all of these instances are efficiently approximable classically through QMC with a folded spectrum, that would be a very significant discovery in its own right.

[7] As remarked earlier, we found the scaling of linear spectral folding in the AQC framework, which has no influence on the high energy spectrum at all, to be close to that of quadratic folding for the same schedule. As finding the deep minimum is an entropic barrier challenge for classical algorithms, we do not expect simulated quantum annealing through QMC to be able to replicate the true quantum result.

VII. CONCLUSIONS AND OUTLOOK

Using the NP-hard MAX-3-XORSAT problem class, we explored the question of classical and quantum approximation hardness from a practical, mechanism-focused point of view. Guided by theoretical intuition, we proposed a class of instances called planted partial solution problems (PPSPs) which we showed are empirically hard for both exact and approximate optimization for classical searches and established quantum methods such as the quantum approximate optimization algorithm (QAOA) and adiabatic quantum computing (AQC). Through extrapolation and qualitative analysis of a rigorous calculation of the typical-case minimum gap at the first order paramagnet to spin glass transition bottlenecking MAX-3-XORSAT, we were able to identify two effects that significantly impede high-depth QAOA's ability to approximate the hardest instances, namely weakening transverse field and destructive interference. We then proposed a novel algorithmic update, called *spectral folding*, that does not suffer from these two issues. Spectral folding is conceptually simple; the analytical analysis of its performance is perhaps less so. And rather intriguingly, it works by applying classical modifications to the classical cost function being optimized, which only provide benefits for quantum optimization.

Using a resummed extensive-order perturbation theory, we were able to predict a constant fraction approximation guarantee for our difficult random hypergraph PPSPs, and consequently, an exponential quantum speedup in the classically hard regime. To further support our claims, we performed a series of numerical simulations of high-depth QAOA, classical optimization routines, and spectral folding methods, out to the largest sizes that we could feasibly reach while still being able to gather good statistics. These numerical results support our claims and we did not discover any meaningful discrepancies or red flags in them; every major prediction here has been numerically checked to the extent we found reasonably possible with current supercomputing resources.

The implications of an effective fast approximation guarantee through spectral folding–or any quantum algorithm, for that matter–are profound, and even restricting ourselves to MAX-3-XORSAT it is important to ask what types of hypergraph might cause it to fail. To be clear, "failure" in this case means that, for a class of hypergraphs, some property invalidates the analytical predictions we made and restricts polynomial-time approximation to values of A near zero. If a subsequent, more sophisticated analysis shows, for example, that due to some subtle effect missed in our resummed extensive order perturbation theory, the variations we propose are asymptotically limited to $A = 0.4$ for random hypergraphs instead of our prediction of ≥ 0.6, we would not consider that a general failure of the algorithm. Given the exponential density of target states, for folded quantum optimization to fail the per-state tunneling rate needs to be much worse than what our calculations and simulations return; even doubling the per-state tunneling decay exponent in Eq. (33) yields $A \simeq 0.18$, a much worse approximation ratio than we predict and observe in simulations, but still a constant fraction better than random guessing.

We are not so hubristic as to claim it is impossible that there is some hidden effect that substantially worsens scaling at large N for random hypergraphs, which is not captured by our theory and is invisible in our numerics. But we see no evidence for it, have no idea what such an issue could be, and formulated our theory such that the simplifying approximations we made were more likely to underestimate the achievable approximation ratio A than overestimate it. Excellent empirical performance in simulation supports this interpretation. We see it as much more likely that one can structure PPSP hypergraphs in some non-random way as to violate the core assumptions of our calculations and become inapproximable. The conditions on such hypergraphs are, however, fairly strict; besides needing to more than double the exponential decay rate of per-state tunneling as compared to random hypergraphs, whatever property is responsible for the slowdown must be resilient both to the addition of a random, uncorrelated problem with similar constraint density and ground state energy, and to spectral deformations such as quadratic folding. Further, such graphs must be relatively dense, as sparse problems are easy to approximate by solving random sub-problems, and their construction rules must not inadvertently render them amenable to classical optimization.

Identifying hypergraph properties that cause this entire method to break down could lead to valuable new discoveries about macroscopic quantum tunneling physics and problem hardness, and further algorithm innovations in the steps needed to mitigate their effects. As we stated in the introduction, the underlying reasons for classical approximation hardness are generic and intuitive if often only applicable to extremal problem instances, but their quantum equivalents are not, and much more opaque. The correctness of our predictions here would imply that quantum approximation hardness may not be generic at all, and instead specific to as-yet undiscovered sets of problem properties. Let's go exploring.

VIII. ACKNOWLEDGMENTS

We would like to thank Matthew Jones, Chris Laumann, Gianni Mossi, Eleanor Rieffel, Antonello Scardicchio and Davide Venturelli for valuable discussions of the issues in this work. We would like to thank Caleb Rotello for detailed feedback on the manuscript. We would also like to thank Takuto Komatsuki and Joey Liu for support with HPC calculations. This work was supported by the DARPA Reversible Quantum Machine Learning and Simulation program under contract HR00112190068, as well as by National Science Founda-

tion grants PHY-1653820, PHY-2210566, DGE-2125899, and by the U.S. Department of Energy, Office of Science, National Quantum Information Science Research Centers, Superconducting Quantum Materials and Systems Center (SQMS) under contract number DE-AC02-07CH11359. The SQMS Center supported EK's advisory role in this project, as well as time improving and fine tuning the algorithms and writing the paper. Many of the numerical simulations in this work were performed with a generous grant of HPC access from the Fujitsu Corporation. Part of this research was performed while the one of the authors (BAB) was visiting the Institute for Pure and Applied Mathematics (IPAM), which is supported by the National Science Foundation (Grant No. DMS-1925919). The Flatiron Institute is a division of the Simons Foundation.

Appendix A: Tunneling between two p-spin wells

To benchmark our prediction in Eq. 28, we performed a series of numerical simulations, shown in FIG. 10. (p, κ) values used were $\{(2, 1.25), (3, 1.25), (4, 1.2), (5, 1.1), (7, 1.06), (9, 1.03)\}$, and the resulting exponents found from numerical fitting of $\Omega_0(N) = a\sqrt{N}2^{-bN}$, as compared to Eq. 28, are (listed as (p, b_{num}, b_{th})) $\{(2, 0.175, 0.180), (3, 0.235, 0.239), (4, 0.275, 0.287), (5, 0.482, 0.512), (7, 0.591, 0.623), (9, 0.676, 0.722)\}$, demonstrating excellent overall agreement, even with the prefactor, as shown in FIG 10.

While fully sufficient for our purposes here (where $p = 3$), we want to highlight two issues with the formulation in Eq. 18 that suggest a more refined treatment should be developed to tackle $p > 3$ and/or asymmetric minima. First, $\partial_m E_{m,n}$ is guaranteed to vanish at some point, since the energy curve at the peak of the barrier is smooth, and since it enters the calculation in the denominator, it predicts a diverging scale for energy corrections. In symmetric cases, this divergence is canceled by the factor of $N/2 - 2n$ in the numerator and the result is finite, but if the energy curve is not symmetrical between the two minima those two factors won't generally coincide (as the barrier peak won't occur at $M/2$ flips for asymmetric minima) and the divergence will not be canceled.

Second, even we consider the symmetric case and the divergence *is* canceled, this formulation of perturbation theory can still lead to unphysical results. For $p \geq 4$, while the energy corrections remain finite throughout, for κ relatively close to κ_c, the renormalized energy barrier peaks at some $m = m_c < M/2$, and decreases from there. This is unphysical, and we can mitigate it with the ad hoc prescription that $u_{m,0}$ simply stops increasing beyond $m = m_c$ and maintains that value until $m = M/2$. The sum in Eq. 23 is similarly modified so that the insertion of an additional spin flip cannot lower the energy. Using this prescription produces quantitatively good results for p running from 4 to 9, as shown in FIG. 10.

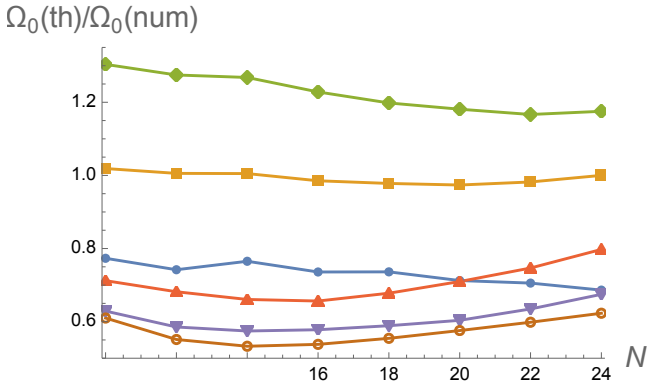

FIG. 10: Comparison of Eq. 28 to the tunnel splitting found from exact diagonalization of the Hamiltonian in Eq. 16, for $p = \{2, 3, 4, 5, 7, 9\}$ (blue, gold, green, purple, red, brown) and κ stated in the text. There are no free parameters in this; exponents found by extrapolating Eq. 28 out to $N = 80$ agree with the numerical results to within few percent. To make sensible predictions for $p \geq 4$ an additional modification was needed to regularize the transverse field corrections to state energies– see text for details. Only $p = 3$ is directly relevant to this work; simulation of other p values is to demonstrate the accuracy and flexibility of our theory.

One can also work through this analysis for the splitting $p = 2$ mean field all-to-all ferromagnet, where the transition out of the ferromagnet phase occurs when $\kappa = 2$ for this normalization. In that case all N spins must flip and there are no secondary spin corrections; further, due to a symmetry cancellation the self energy corrections (e.g. Eq. 18) are energy-independent and thus do not renormalize the tunneling barrier. Repeating the same steps for this somewhat simpler calculation yields

$$\Omega_0(N\kappa) \simeq \sqrt{\frac{N}{2\pi}} \frac{\kappa^N}{w(N,\kappa)^2} \left(\frac{e}{4}\right)^N. \tag{A1}$$

We of course still need to evaluate the normalization factor $w(N, \kappa)$. Explicitly, it is given by

$$w(N,\kappa) = \sqrt{1 + \sum_{k=1}^{N/2} \binom{N}{k} \kappa^{2k} \left(k! \prod_{j=1}^{k} \epsilon_k^{-1}\right)^2} \tag{A2}$$

This function is extremely well fit by a simple exponential in N, with a coefficient that depends on κ. Empirically, it can be well approximated as,

$$w(N,\kappa)^2 \propto \left(1 + a\kappa^b\right)^N, \tag{A3}$$

where fitting a range of κ values from 0.4 to 1.6 gives $a = 0.066$ and $b = 2.25$, in fairly good agreement with the result $\left(1 + \kappa^2/16\right)^M$ one can derive from simple second order perturbation theory.

Of note is that Eq. A2 is both of similar quantitative accuracy as the results plotted in FIG. 10, and it

predicts that the decay exponent smoothly approaches zero as $\kappa \to 2$, at which point the minimum gap decays inverse polynomially. Similar behavior–a crossover to polynomial scaling when theory predicts the decay exponent must vanish–was observed both analytically and numerically in the 1d transverse field Ising chain in [25]. We see this as lending further indirect support to our polynomial time approximation hardness prediction in Eq. 33–though the situation is fairly different all these calculations identify a crossover to polynomial scaling by the point at which a predicted decay exponent vanishes in careful many-body theory.

Appendix B: Band structure considerations for linear spectral folding

In the simulations in section VI, we noticed that the quadratic AQC formulation of spectrally folded quantum optimization performed similarly to the rough prediction in section IV, returning a worst case approximation ratio of $q_a = 0.7$ compared to the approximate theoretical estimate 0.68. In contrast, the linear trial minimum annealing versions more notably outperformed the analyitical prediction (returning $q_a = 0.75$ as compared to 0.6), and in numerical experiments the best choice for the lowering parameter $C(t)$ in Eq. 8 was 2-3 times that one would naively expect. We suspect that this has to do with the band structure of the dressed eigenstates in $\mathcal{T}$ (the band of classical states at or very near the fold energy) when the transverse field κ is nonzero, and in this subsection argue why this might be the case. We present these arguments as suggestions and not a rigorous proof, and think more work on this issue could be valuable for shedding light on the detailed structure of these optimization algorithms.

Let $\{|T_j\rangle\}$ be the set of all $N_{\mathcal{T}}$ bitstring states in $\mathcal{T}$; states in $\mathcal{T}$ all have $E \simeq AE_{\rm GS}$ before any corrections from the lowering Hamiltonian. Further let $|\psi_k\rangle$ be a dressed eigenstate whose spectral weight is concentrated in $\mathcal{T}$. Thus

$$|\psi_k\rangle \simeq \sum_{j\in\mathcal{T}}^{N_{\mathcal{T}}} c_{jk}\,|T_j\rangle\,, \quad \left\langle \left|c_{jk}^2\right| \right\rangle \simeq \frac{1}{N_{\mathcal{T}}}. \tag{B1}$$

Finally, let Ω_{Lj} be the M-spin tunneling matrix element into a bitstring state $|T_j\rangle$ from $|L_D\rangle$, where $\langle \Omega_{Lj}\rangle_j = \Omega_0$ as calculated in Eq. 28.

We choose our gauge and basis so that all transverse field terms are negative and H is real, which is straightforward since H is stoquastic. In this gauge we can assume all the Ω_{Lj} matrix elements are negative, though their magnitudes can of course vary substantially over $\mathcal{T}$ (but we expect are fairly well-correlated locally, when considering states in $\mathcal{T}$ only a few flips apart). Because the states in $\mathcal{T}$ are all near-zero energy, the local transverse field-induced "hopping" matrix elements connecting them are either direct (if a given pair of states in $\mathcal{T}$ are one flip apart), or the result of short ranged, few-flip tunneling through local excited states. In either case, the resulting matrix element is negative, and the band of $|\psi_k\rangle$ states are (approximately) the eigenstates of a hopping-like model on a sparse, disordered graph of $N_{\mathcal{T}}$ sites, where most hopping matrix elements are negative and there is local potential disorder. With all these quantities defined, we now want to estimate the tunneling matrix element $\tilde{\Omega}_{Lk}$ from $|L_D\rangle$ into $|\psi_k\rangle$.

We first consider states $|\psi_k\rangle$ near the band center, e.g. where the energy shifts from the transverse field are solely due to the second order local processes captured in section IV and the "hopping energy" is nearly zero. For such states, we can assume the signs of the amplitudes c_{jk} are randomized. Then by the law of large numbers:

$$\begin{aligned}\left\langle \tilde{\Omega}_{Lk}^2 \right\rangle_k &\simeq \left\langle \left(\sum_{j\in\mathcal{T}}^{N_{\mathcal{T}}} c_{jk}\Omega_{Lj} \right)^2 \right\rangle_k \\ &\propto N_{\mathcal{T}} \left\langle c_{jk}^2 \Omega_{Lj}^2 \right\rangle_j \propto \left\langle \Omega_{Lj}^2 \right\rangle .\end{aligned} \tag{B2}$$

Thus, the average squared matrix element to tunnel into a state in the band center has the same scaling as one obtains from a more naive calculation that ignores the band structure entirely. Since the majority of states in the band are near the center in this respect, this calculation shows that ignoring the local band structure in $\mathcal{T}$ is a decent approximation for obtaining the average collective tunneling rate that enters into equations like 33.

But what happens when we consider states that are more extremal, e.g. near the top or bottom of the band? Near the top of the band, we can assume significant local alternation in the signs/phases of the c_{jk} coefficients, and thus the assumption of randomization is still fairly good and Eq. B2 likely accurately captures the scaling. For states near the bottom of the band, however, the situation changes. First, these states will be lower in energy by an $O(N)$ factor, which means $C(t)$ must be further reduced for $|L_D\rangle$ to cross them, and the individual tunneling matrix elements Ω_{Lj} into such states may scale differently as a result. We label these matrix elements Ω'_{Lj} and do not attempt to predict what, if any, changes to scaling may arise.

Second, to minimize the energy the signs/phases of the c_{jk} of nearby (in Hamming distance) configurations will be synchronized given that the matrix elements that couple them are real and (mostly) negative (again, in this gauge choice). And this synchronization can dramatically enhance tunneling rates. Let us guess, for example, that in the band ground state $|\psi_0\rangle$, *most* c_{j0} are positive. Then:

$$\left|\tilde{\Omega}_{L0}^2\right| \simeq \left\langle \left(\sum_{j\in\mathcal{T}}^{N_{\mathcal{T}}} c_{j0}\Omega'_{Lj} \right)^2 \right\rangle \propto N_{\mathcal{T}} \left\langle \left(\Omega'_{Lj}\right)^2 \right\rangle . \tag{B3}$$

Since $N_{\mathcal{T}}$ is exponentially large in N (Eq. 31), the tunneling matrix element into the band ground state can

be exponentially larger than the average matrix element for tunneling into the band center, though if Ω'_{Lj} decays more quickly with N that may reduce or erase this effect. We expect that collective tunneling into low-lying states in the band could be similarly enhanced even if there is more variation in the signs of the c_{jk} terms, as the matrix elements Ω_{Lj} are locally correlated in magnitude. And though there are exponentially fewer states near the band bottom compared to the band center, if the tunneling matrix elements are exponentially larger the weighted average over all states in the band can be substantially increased–we saw this effect already in averaging over the relative distance to $|L\rangle$ of different states in $\mathcal{T}$ (Eq. 32). It is thus quite plausible that band structure effects could further increase collective tunneling rates and be responsible for the relative overperformance of TMA spectral folding in our simulations, as compared to the analytical prediction.

All that said, we did not include this analysis in the main text because we do not presently have the mathematical tools to quantitatively predict the band structure and nature of the dressed eigenstates, nor can we predict the potential scaling changes in the collective tunneling matrix elements Ω'_{Lj} to states near the bottom of the band. So we present these results to justify our claim that ignoring band structure is likely to underestimate the achievable approximation ratio, and to suggest interesting directions for further research.

[1] B. H. Korte, J. Vygen, B. Korte, and J. Vygen, *Combinatorial optimization*, vol. 1 (Springer, 2011).

[2] M. R. Garey, D. S. Johnson, and L. Stockmeyer, in *Proceedings of the sixth annual ACM symposium on Theory of computing* (1974), pp. 47–63.

[3] J. D. Ullman, Journal of Computer and System sciences **10**, 384 (1975).

[4] G. S. Grest, C. Soukoulis, and K. Levin, Physical Review Letters **56**, 1148 (1986).

[5] P. Crescenzi, V. Kann, and M. Halldórsson, *A compendium of np optimization problems* (1995).

[6] D. S. Hochba, ACM Sigact News **28**, 40 (1997).

[7] R. Monasson, Journal of Physics A: Mathematical and General **31**, 513 (1998).

[8] G. J. Woeginger, in *Combinatorial optimization: eureka, you shrink!* (Springer, 2003), pp. 185–207.

[9] V. Bapst, L. Foini, F. Krzakala, G. Semerjian, and F. Zamponi, Physics Reports **523**, 127 (2013).

[10] C. J. Hillar and L.-H. Lim, Journal of the ACM (JACM) **60**, 1 (2013).

[11] D. Venturelli, S. Mandrà, S. Knysh, B. Gorman, R. Biswas, and V. Smelyanskiy, Physical Review X **5**, 031040 (2015).

[12] M. Bellitti, F. Ricci-Tersenghi, and A. Scardicchio, *Entropic barriers as a reason for hardness in both classical and quantum algorithms* (2021), 2102.00182.

[13] C. Jones, K. Marwaha, J. S. Sandhu, and J. Shi, arXiv preprint arXiv:2210.03006 (2022).

[14] S. Arora and B. Barak, *Computational complexity: a modern approach* (Cambridge University Press, 2009).

[15] A. Finnila, M. Gomez, C. Sebenik, C. Stenson, and J. Doll, Chemical physics letters **219**, 343 (1994).

[16] T. Kadowaki and H. Nishimori, Physical Review E **58**, 5355 (1998).

[17] A. Das and B. K. Chakrabarti, Reviews of Modern Physics **80**, 1061 (2008).

[18] M. W. Johnson, M. H. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, et al., Nature **473**, 194 (2011).

[19] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, Nature Physics **10**, 218 (2014).

[20] P. Hauke, H. G. Katzgraber, W. Lechner, H. Nishimori, and W. D. Oliver, arXiv preprint arXiv:1903.06559 (2019).

[21] A. D. King, J. Raymond, T. Lanting, R. Harris, A. Zucca, F. Altomare, A. J. Berkley, K. Boothby, S. Ejtemaee, C. Enderud, et al., Nature pp. 1–6 (2023).

[22] E. Farhi, J. Goldstone, S. Gutmann, and M. Sipser, arXiv:quant-ph/0001106 (2000).

[23] T. Albash and D. A. Lidar, arXiv:1611.04471 (2017).

[24] E. Farhi, J. Goldstone, and S. Gutmann, arXiv preprint arXiv:1411.4028 (2014).

[25] G. Grattan, B. A. Barton, S. Feeney, G. Mossi, P. Patnaik, J. C. Sagal, L. D. Carr, V. Oganesyan, and E. Kapit, *Exponential acceleration of macroscopic quantum tunneling in a floquet ising model* (2023), 2311.17814.

[26] T. Albash and D. A. Lidar, Physical Review X **8**, 031016 (2018).

[27] A. D. King, J. Carrasquilla, J. Raymond, I. Ozfidan, E. Andriyash, A. Berkley, M. Reis, T. Lanting, R. Harris, F. Altomare, et al., Nature **560**, 456 (2018).

[28] S. Ebadi, A. Keesling, M. Cain, T. T. Wang, H. Levine, D. Bluvstein, G. Semeghini, A. Omran, J.-G. Liu, R. Samajdar, et al., Science **376**, 1209 (2022).

[29] R. Babbush, J. R. McClean, M. Newman, C. Gidney, S. Boixo, and H. Neven, PRX Quantum **2**, 010103 (2021).

[30] G. Parisi, Physical Review Letters **43**, 1754 (1979).

[31] D. J. Earl and M. W. Deem, Physical Chemistry Chemical Physics **7**, 3910 (2005).

[32] R. Shaydulin, C. Li, S. Chakrabarti, M. DeCross, D. Herman, N. Kumar, J. Larson, D. Lykov, P. Minssen, Y. Sun, et al., arXiv preprint arXiv:2308.02342 (2023).

[33] C. Baldwin and C. Laumann, Physical Review B **97**, 224201 (2018).

[34] V. N. Smelyanskiy, K. Kechedzhi, S. Boixo, S. V. Isakov, H. Neven, and B. Altshuler, arXiv preprint arXiv:1802.09542 (2018).

[35] K. Kechedzhi, V. Smelyanskiy, J. R. McClean, V. S. Denchev, M. Mohseni, S. Isakov, S. Boixo, B. Altshuler, and H. Neven, arXiv preprint arXiv:1807.04792 (2018).

[36] V. N. Smelyanskiy, K. Kechedzhi, S. Boixo, H. Neven, and B. Altshuler, arXiv preprint arXiv:1907.01609 (2019).

[37] Y. Susa, Y. Yamashiro, M. Yamamoto, I. Hen, D. A. Lidar, and H. Nishimori, Phys. Rev. A **98**, 042326 (2018), URL `https://link.aps.org/doi/10.`

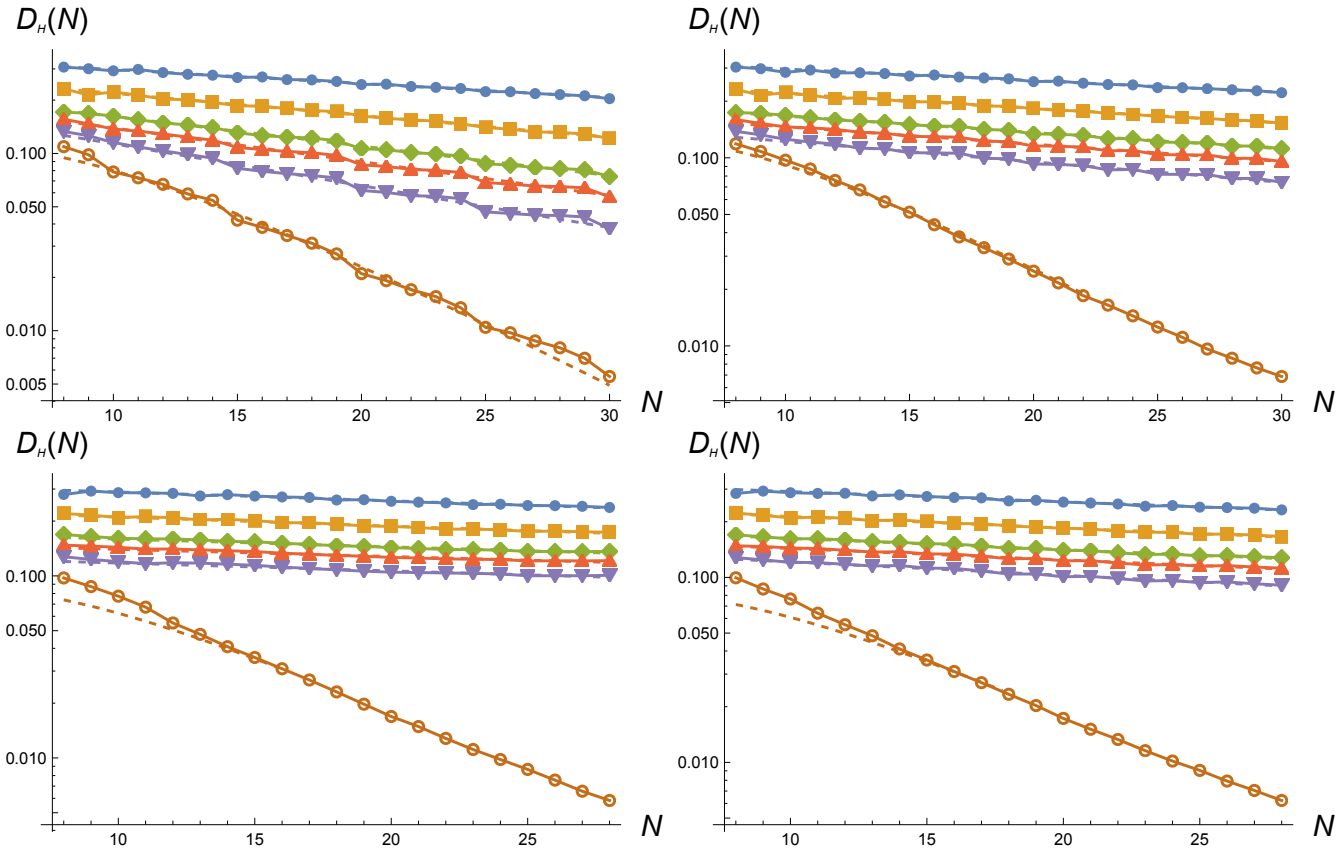

FIG. 11: Probabilities of returning states within Hamming distance $D_H = \{2/5, 1/3, 1/4, 1/5, 1/8, 0\}\, N$ flips from the ground state G for QAOA (top to bottom curves), for $N_C = \left\{2, 4, 6, 3\sqrt{N}/2\right\} N$ (clockwise from top left) and application parameters in text. As seen in the figures, the per-shot probability of finding the ground state is essentially independent of constraint density and tracks the prediction $N2^{-0.28N}$ in Eq. 15. On the other hand, the probabilities of returning states at various extensive fractional Hamming distances, e.g. $N/4$ or fewer flips, decay much more slowly, and the task of finding states comparatively close to G becomes easier as N_C/N increases. That said, these probabilities all decay exponentially with N in all cases, consistent with the worsening approximation ratios observed in figure 6.

1103/PhysRevA.98.042326.
[38] T. Graß, Phys. Rev. Lett. **123**, 120501 (2019), URL https://link.aps.org/doi/10.1103/PhysRevLett.123.120501.
[39] D. Sels and A. Polkovnikov, Proceedings of the National Academy of Sciences **114**, E3909 (2017).
[40] S. Bao, S. Kleer, R. Wang, and A. Rahmani, Physical Review A **97**, 062343 (2018).
[41] Z.-C. Yang, A. Rahmani, A. Shabani, H. Neven, and C. Chamon, Physical Review X **7**, 021027 (2017).
[42] I. Čepaitė, A. Polkovnikov, A. J. Daley, and C. W. Duncan, PRX Quantum **4**, 010312 (2023), URL https://link.aps.org/doi/10.1103/PRXQuantum.4.010312.
[43] A. Montanari, SIAM Journal on Computing pp. FOCS19–1 (2021).
[44] E. Farhi, J. Goldstone, S. Gutmann, and L. Zhou, Quantum **6**, 759 (2022).
[45] S. Boulebnane and A. Montanaro, arXiv preprint arXiv:2110.10685 (2021).
[46] J. Basso, E. Farhi, K. Marwaha, B. Villalonga, and L. Zhou, arXiv preprint arXiv:2110.14206 (2021).
[47] B. Altshuler, H. Krovi, and J. Roland, Proceedings of the National Academy of Sciences **107**, 12446 (2010).
[48] S. Knysh, Nature communications **7** (2016).
[49] Z. Tang and E. Kapit, Physical Review A **103**, 032612 (2021).
[50] E. Farhi, J. Goldstone, D. Gosset, S. Gutmann, and P. Shor, arXiv preprint arXiv:1010.0009 (2010).
[51] S. Sahni and T. Gonzalez, Journal of the ACM (JACM) **23**, 555 (1976).
[52] J. Håstad, Journal of the ACM (JACM) **48**, 798 (2001).
[53] M. Mézard, G. Parisi, and R. Zecchina, Science **297**, 812 (2002).
[54] M. Mézard, T. Mora, and R. Zecchina, Physical Review Letters **94**, 197205 (2005).
[55] A. K. Hartmann and M. Weigt, *Phase transitions in combinatorial optimization problems: basics, algorithms and statistical mechanics* (John Wiley & Sons, 2006).
[56] F. Krzakala and J. Kurchan, Physical Review E **76**, 021122 (2007).
[57] F. Altarelli, R. Monasson, and F. Zamponi, in *Journal of Physics: Conference Series* (IOP Publishing, 2008), vol. 95, p. 012013.
[58] M. Ibrahimi, Y. Kanoria, M. Kraning, and A. Monta-

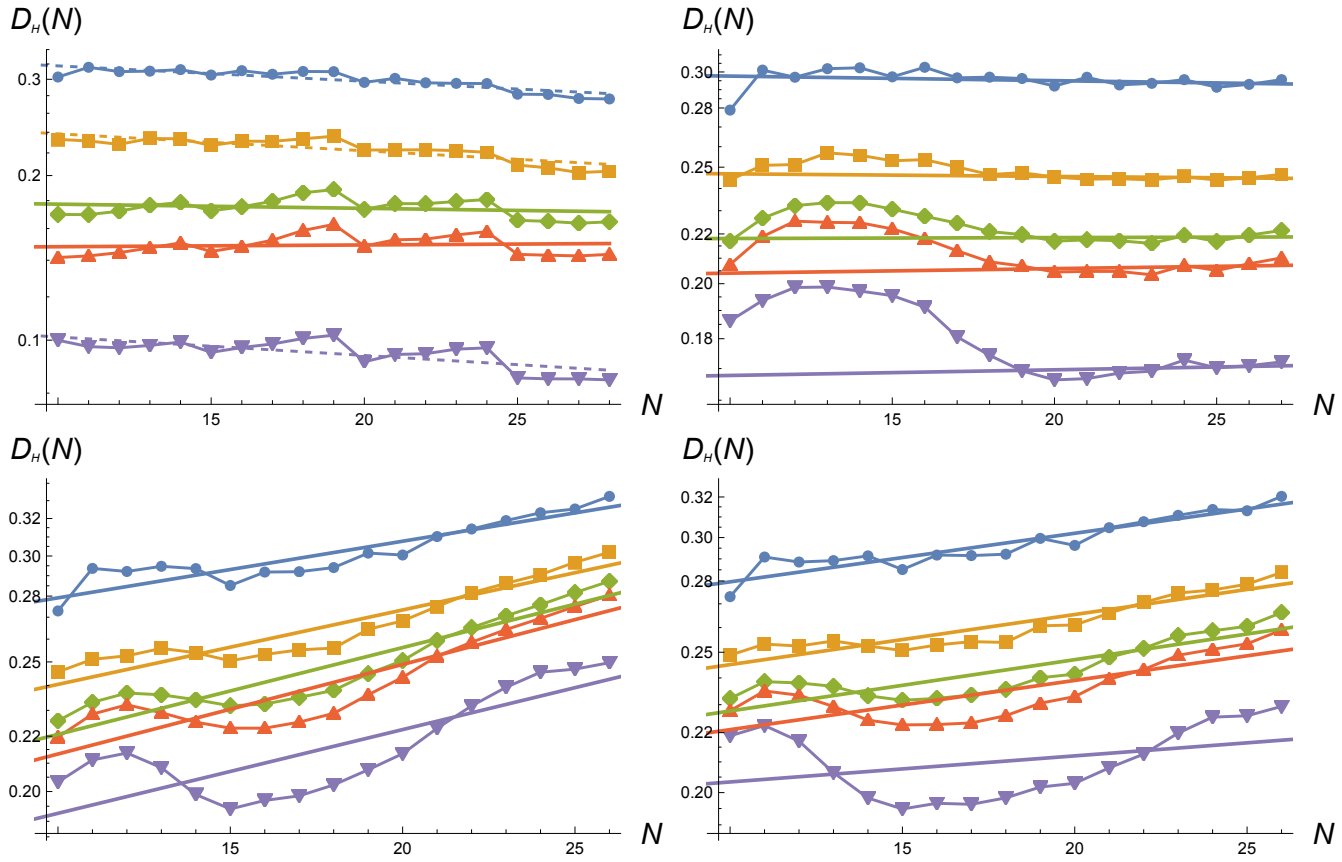

FIG. 12: Probabilities of returning states within Hamming distance $D_H = \{2/5, 1/3, 1/4, 1/5, 1/8\}\,N$ flips from the ground state G for the AQC formulation of spectral folding in figure 7 (top to bottom curves), for $N_C = \left\{2, 4, 6, 3\sqrt{N}/2\right\} N$ (clockwise from top left) and application parameters in text. For $N_C = 2N$ some probabilities decay slowly with system size, due to competition with other minima (though non-monotonicity makes the fitting somewhat ambiguous here); for all other cases they are constant or increase toward some large N saturation value, indicating that this variation is finding states near the global minimum with constant probability. The probability of finding G is not plotted, as for spectrally folded optimization it's essentially zero.

nari, in *Proceedings of the twenty-third annual ACM-SIAM symposium on Discrete Algorithms* (SIAM, 2012), pp. 760–779.

[59] J. Basso, D. Gamarnik, S. Mei, and L. Zhou, in *2022 IEEE 63rd Annual Symposium on Foundations of Computer Science (FOCS)* (IEEE, 2022), pp. 335–343.

[60] S. Boulebnane and A. Montanaro, arXiv preprint arXiv:2208.06909 (2022).

[61] A. Anshu and T. Metger, Quantum **7**, 999 (2023).

[62] N. Benchasattabuse, A. Bärtschi, L. P. García-Pintos, J. Golden, N. Lemons, and S. Eidenbenz, arXiv preprint arXiv:2308.15442 (2023).

[63] J. Roland and N. J. Cerf, Phys. Rev. A **65**, 042308 (2002).

[64] O. Dubois and J. Mandler, Comptes Rendus Mathematique **335**, 963 (2002).

[65] J. Håstad and S. Venkatesh, in *Proceedings of the thirty-fourth annual ACM symposium on Theory of computing* (2002), pp. 43–52.

[66] M. Kowalsky, T. Albash, I. Hen, and D. A. Lidar, arXiv preprint arXiv:2103.08464 (2021).

[67] L. Zhu, H. L. Tang, G. S. Barron, F. Calderon-Vargas, N. J. Mayhall, E. Barnes, and S. E. Economou, Physical Review Research **4**, 033029 (2022).

[68] E. Kapit and V. Oganesyan, Quantum Science and Technology **6**, 025013 (2021).

[69] B. Barak, A. Moitra, R. O'Donnell, P. Raghavendra, O. Regev, D. Steurer, L. Trevisan, A. Vijayaraghavan, D. Witmer, and J. Wright, arXiv preprint arXiv:1505.03424 (2015).

[70] A. Anshu, D. Gosset, K. J. M. Korol, and M. Soleimanifar, Physical Review Letters **127**, 250502 (2021).

[71] K. Marwaha and S. Hadfield, Quantum **6**, 757 (2022).

[72] L. W. Wang and A. Zunger, The Journal of Physical Chemistry **98**, 2158 (1994).

[73] L.-W. Wang and A. Zunger, The Journal of Chemical Physics **100**, 2394 (1994).

[74] J. R. McClean, J. Romero, R. Babbush, and A. Aspuru-Guzik, New Journal of Physics **18**, 023023 (2016).

[75] R. Santagati, J. Wang, A. A. Gentile, S. Paesani, N. Wiebe, J. R. McClean, S. Morley-Short, P. J. Shadbolt, D. Bonneau, J. W. Silverstone, et al., Science advances **4**, eaap9646 (2018).

[76] F. Zhang, N. Gomes, Y. Yao, P. P. Orth, and T. Iadecola, Physical Review B **104**, 075159 (2021).

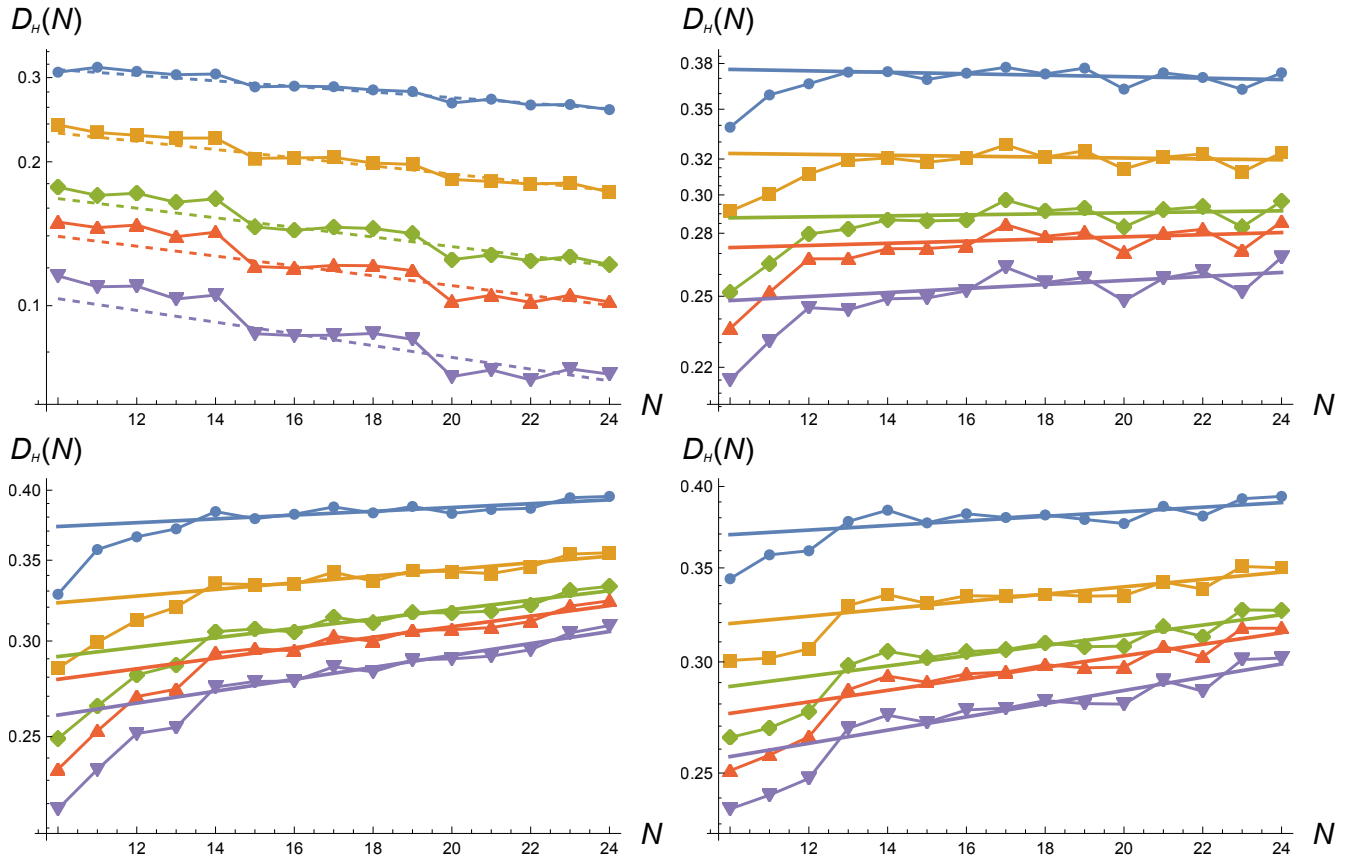

FIG. 13: Probabilities of returning states within Hamming distance $D_H = \{2/5, 1/3, 1/4, 1/5, 1/8\}\,N$ flips from the ground state G for the TMA formulation of spectral folding in figure 8 (top to bottom curves), for $N_C = \left\{2, 4, 6, 3\sqrt{N}/2\right\} N$ (clockwise from top left) and application parameters in text. For $N_C = 2N$ the probabilities decay slowly with system size, due to competition with other minima; for all other cases they are constant or increase toward some large N saturation value, indicating that this variation is finding states near the global minimum with constant probability. The probability of finding G is not plotted, as for spectrally folded optimization it's essentially zero.

[77] J. Tilly, H. Chen, S. Cao, D. Picozzi, K. Setia, Y. Li, E. Grant, L. Wossnig, I. Rungger, G. H. Booth, et al., Physics Reports **986**, 1 (2022).

[78] L. C. Tazi and A. J. Thom, arXiv preprint arXiv:2305.04783 (2023).

[79] E. Kapit, *Systems and methods for accelerated quantum optimization* (2021), uS Patent App. 17/027,146.

[80] S. Bravyi, D. P. DiVincenzo, and D. Loss, Annals of physics **326**, 2793 (2011).

[81] F. Pietracaprina, V. Ros, and A. Scardicchio, Physical Review B **93**, 054201 (2016).

[82] C. Baldwin, C. Laumann, A. Pal, and A. Scardicchio, Physical Review B **93**, 024202 (2016).

[83] C. Baldwin, C. Laumann, A. Pal, and A. Scardicchio, Physical Review Letters **118**, 127201 (2017).

[84] A. Scardicchio and T. Thiery, arXiv preprint arXiv:1710.01234 (2017).

[85] B. Derrida, Physical Review Letters **45**, 79 (1980).

[86] T. Jörg, F. Krzakala, J. Kurchan, A. C. Maggs, and J. Pujos, EPL (Europhysics Letters) **89**, 40004 (2010).

[87] B. F. Schiffer, D. S. Wild, N. Maskara, M. Cain, M. D. Lukin, and R. Samajdar, arXiv preprint arXiv:2306.13131 (2023).

[88] E. Kapit, *Systems and methods for passive quantum error correction* (2021), uS Patent 10,956,267.

[89] G. Mossi, V. Oganesyan, and E. Kapit, arXiv preprint arXiv:2306.10632 (2023).

[90] Y. Suzuki, Y. Kawase, Y. Masumura, Y. Hiraga, M. Nakadai, J. Chen, K. M. Nakanishi, K. Mitarai, R. Imai, S. Tamiya, et al., Quantum **5**, 559 (2021).

[91] S. V. Isakov, G. Mazzola, V. N. Smelyanskiy, Z. Jiang, S. Boixo, H. Neven, and M. Troyer, Physical review letters **117**, 180402 (2016).

[92] E. Andriyash and M. H. Amin, arXiv preprint arXiv:1703.09277 (2017).

[93] Z. Jiang, V. N. Smelyanskiy, S. V. Isakov, S. Boixo, G. Mazzola, M. Troyer, and H. Neven, Physical Review A **95**, 012322 (2017).

[94] Z. Jiang, V. N. Smelyanskiy, S. Boixo, and H. Neven, Physical Review A **96**, 042330 (2017).

[95] A. D. King, J. Raymond, T. Lanting, S. V. Isakov, M. Mohseni, G. Poulin-Lamarre, S. Ejtemaee, W. Bernoudy, I. Ozfidan, A. Y. Smirnov, et al., arXiv preprint arXiv:1911.03446 (2019).

[96] B. Heim, T. F. Rønnow, S. V. Isakov, and M. Troyer, Science **348**, 215 (2015).

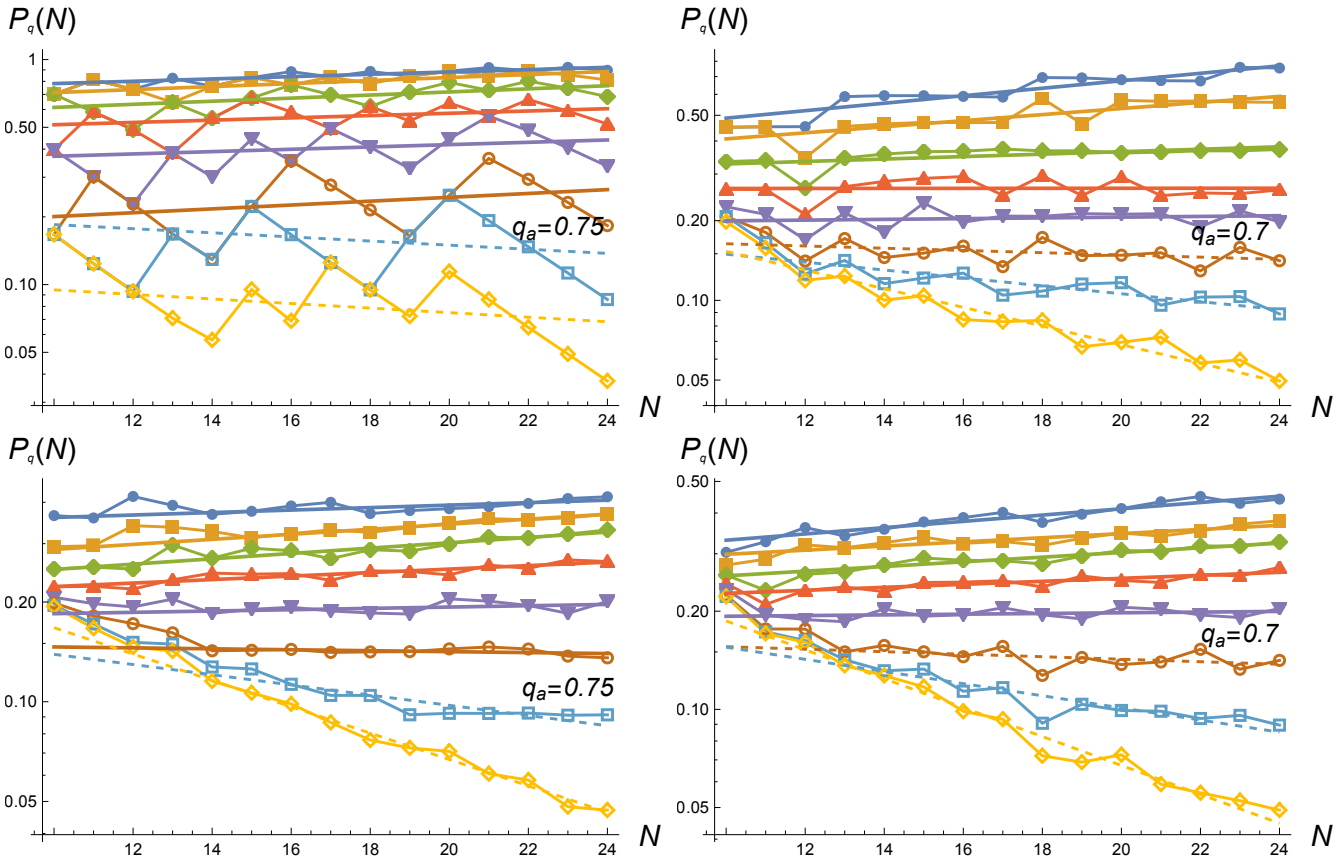

FIG. 14: Performance quadratic AQC spectral folding, for $N_C = \left\{2, 4, 6, 3\sqrt{N}/2\right\} N$ (clockwise from top left) with increased approximation target $A = 0.85$ and all other parameters equal. In comparison to the $A = 0.75$ results in figure 7, increasing A improves prefactors but does not improve scaling and in the case of $N_C/N = 6$, modestly reduces q_a. This suggests we have reached the performance ceiling for this variation on the tested PPSP classes.