FUTURE OF PHYSICAL QUANTUM ANNEALERS: IMPEDIMENTS AND HOPES

TAMEEM ALBASH1,2 AND ITAY HEN1,2

Physical quantum annealers have so far been unable to demonstrate scalable performance advantages over standard state-of-the-art optimization algorithms. Nevertheless, multiple independent efforts continue to develop the technology, with the belief that optimizers with more coherent and more densely connected quantum bits are necessary to observe a quantum speedup. In this review we discuss the role of finite temperature and thermalization in preventing these devices from functioning as scalable optimizers. In light of this, we discuss alternative potentially promising uses for such devices that may enable them to provide advantages outside of optimization.

I. Introduction

Quantum annealing1–6 is a technique that utilizes gradually decreasing quantum fluctuations to search for global minima of complicated cost functions. As an inherently quantum technique, quantum annealing holds the promise to solve certain optimization problems faster than traditional classical algorithms as originally alluded to in the seminal paper of Ray et al.7.

Quantum optimization devices implementing quantum annealing have been at the forefront of efforts to develop quantum information processing devices that can perform tasks not accessible by the largest high performance supercomputers. Already realized on various platforms8–14, these ‘quantum annealers’ solve optimization problems, that is, they find bit assignments that minimize the cost of discrete combinatorial problems or equivalently the ground states of Ising Hamiltonians, by quenching quantum fluctuations in a manner analogous to how thermal fluctuations are quenched in simulated annealing15. Driving these developments is the hope that quantum annealing devices will eventually be integrated with high performance computers and help mitigate important bottlenecks of computation.

Despite the excitement brought on by the recent technological breakthroughs in the field, which have made programmable quantum annealing optimizers consisting of thousands of quantum bits commercially available1–5, quantum adiabatic protocols have so far failed to deliver on their promise to serve as useful optimizers, i.e., to find optimum solutions of discrete combinatorial optimization problems faster than possible classically. Thus far, no examples (neither experimental nor theoretical) of practical relevance have been found to indicate a superiority of quantum adiabatic optimization over traditional methods16–24.

This absence of positive or promising results raises several pertinent questions that lie at the heart of the field. (i) Could there be fundamental limitations to quantum annealers which prevent them from exhibiting scalable quantum speedups? (ii) If there are such, can these be mitigated or overcome? (iii) Can we still expect quantum annealers to provide demonstrable quantum advantages?

In this review, we make an attempt to cover several studies set out to address the above questions and to provide partial answers. We focus on the role of temperature, specifically thermalization to the Boltzmann distribution, as an impediment to the success of a quantum
annealer as an optimizer. This review is organized as follows. In Sec. II, we provide a brief overview of the decoherence model we consider. In Sec. III, we review recent studies that reflect the detrimental role of temperature on the success of an optimizer. In Sec. IV, we present recent work that shows how operating a quantum annealer in a non-standard way can boost its success as an optimizer. In light of this, we devote Sec. V to discuss a different use of quantum annealers, namely their use as quantum simulators rather than optimizers. We analyze the projected performance of quantum annealers as quantum simulators and argue that in certain cases, quantum annealers may be expected to exhibit scaling advantages over classical state-of-the-art. We summarize the review with some discussion and conclusions in Sec. VI.

II. A Simple Model for Finite-temperature Quantum Annealers

For this review, we restrict our attention to perhaps the most innocuous model of decoherence for a quantum annealer. We assume the dissipative dynamics is described by a master equation of Lindblad form\(^{25}\), often referred to as the Markovian adiabatic master equation (AME)\(^{26-28}\). Under this model, we will refer to our quantum annealer as a finite-temperature quantum annealer. While a microscopic derivation of such a master equation is possible, the assumptions used in the derivation are not expected to hold true during the entire anneal, especially near small energy gaps. More involved descriptions in these regimes are required\(^{29-31}\), but these noise models are likely to give further pessimistic predictions of the performance of the quantum annealer.

The key feature of the AME for us is that, if the annealer were to be paused at a given point along the interpolation, denoted by the dimensionless interpolation parameter \(s\), the steady state of the dynamics (barring pathological situations) would be the Gibbs state (or Boltzmann distribution) associated with \(H(s)\) at the temperature set by the thermal environment, 
\[
\rho_G(s) = e^{-\beta H(s)} \mathcal{F} e^{-\beta H(s)}.
\]
The dissipative dynamics can therefore be thought of as trying to always push the system towards the instantaneous Gibbs state \(\rho_G(s)\), with a time constant that depends on the effective strength of the dissipation at that point in the anneal\(^2\). For example, the excitation and relaxation dynamics between the \(i\)-th and \(j\)-th energy eigenstates are slowed down when the overlap 
\[
|\langle \psi_i(s) | A_a | \psi_j(s) \rangle|
\]
becomes small, where \(A_a\) is the system operator from the system-bath interaction, and \(\{|\psi_i(s)\}\) is the instantaneous \(i\)-th energy eigenstate of \(H(s)\). For an independent ‘dephasing’ bath, i.e. \(A_a = \sigma_a^z\), this happens late in the anneal when the Hamiltonian \(H(s)\) is dominated by its diagonal in the computational basis component.

In the closed-system case, quantum annealers are guaranteed in the adiabatic limit to find with high probability a minimizing configuration of the cost function, or the ground state of the final Hamiltonian they are to solve\(^{33-41}\). For a finite temperature quantum annealer, there is no guarantee of reaching the ground state with high probability. Instead the open-system adiabatic theorem only gives us a guarantee of reaching the Boltzmann distribution of the final Hamiltonian at the annealer temperature\(^{42}\). In this regard, temperature can play a critical role in the success of the quantum annealer.

In order to illustrate this more concretely, we consider a classical algorithm that effectively realizes the dynamics above: simulated quantum annealing (SQA)\(^{43}\), also referred to as path-integral Monte Carlo along the quantum annealing schedule (PIMC-QA)\(^{44}\). For any given \(s\), the quantum partition function is mapped to the dual classical system, on which Monte Carlo updates are performed. The state after the updates is fed into the dual classical system at \(s + \Delta s\). Because quantum Monte Carlo at every \(s\) attempts to sample from the Gibbs state associated with \(s\), we can think of the Monte Carlo updates as helping the system ‘follow’ the instantaneous Gibbs state in a manner analogous to the dynamics of the AME. In fact, SQA has been used with remarkable success to reproduce the output characteristics of the D-Wave quantum annealing processors\(^{24,45}\).

![Figure 1. Computational effort measured in terms of TTS for SQA operating at different inverse-temperatures \(\beta\). Results are for the ‘logical-planted’ class of instances studied in Ref. [24]. SQA simulations used a fixed 64 Trotter slices and the annealing schedule of a D-Wave processor. The latter sets the energy scale, such that \(\beta = 0.5\) approximately corresponds to a temperature of 15mK.](image-url)
In Ref.\textsuperscript{24}, SQA was used to demonstrate the detrimental role an insufficiently low temperature can have on the performance of a quantum annealer as an optimizer. We reproduce in Fig. 1 the time-to-solution (TTS), a measure of the computational effort to find the ground state at least once with 99% probability, for SQA on a specific class of Ising instances defined on the Chimera connectivity\textsuperscript{46}. The results show that for this class of instances, the temperature can have a dramatic effect on performance, resulting in both an absolute and scaling with problem size difference.

**III. Fundamental Limitations due to Temperature to Scalable Optimizations**

In what follows, we make the assumption that the open-system dynamics is sufficiently strong that crossing the minimum gap, even when the closed system adiabatic condition is satisfied, depopulates the ground state to the point that it can only be repopulated by thermal relaxation after the minimum gap. In this case, the success of the quantum annealer for finding the ground state is bounded by the ground state population in the thermal state. In Ref.\textsuperscript{47}, we showed that even instantly-thermalizing finite-temperature quantum annealers are severely limited as optimizers.

Under plausible assumptions for a scalable architecture (bounded connectivity graph with discretized couplings) of qubits and the behavior of the free energy, it follows that the ground state energies of the problem Hamiltonian, denoted $E_0$, of any given problem class, scale linearly with increasing $n$ (i.e., the energy is an extensive property as is generically expected from physical systems) while the minimal gap $\Delta = E_1 - E_0$ remains fixed.

It then follows\textsuperscript{47} that the probability density of finding a state with intensive energy $e = E/n$ is approximately Gaussian centered around a value $e^* > E_0/n$. The value $e^*$ is the most probable intensive energy given by the maximum of the free energy density. Crucially, the variance of the Gaussian scales inversely with $n$, which means that the probability of finding by Boltzmann-sampling any intensive energy $e < e^*$ (equivalently, $E < e^* n$) is exponentially suppressed in $n$. Ideal fixed temperature quantum annealers that thermalize instantaneously to the Gibbs state of the classical Hamiltonian are therefore exponentially unlikely to find the ground state since $e^* > e_0 = E_0/n$.

In light of the above, instantly thermalizing quantum annealers at a fixed temperature are doomed to fail as optimizers as problem sizes increase. However, as was shown in Ref.\textsuperscript{57}, the performance level can be maintained as $n$ increases if the temperature is appropriately scaled with $n$. An estimate for the required scaling can be had by demanding that the ground state energy falls within the variation of the mean energy (recall from the discussion above that the distribution is Gaussian). The scaling is then controlled by the specific heat $c_\beta = \partial^2 \langle e \rangle / \partial \beta^2$. For a power-law specific heat $-c_\beta \sim \beta^{-\alpha - 2}$, it follows that the inverse-temperature must scaling as $\beta \sim n^{1/\alpha}$. If on the other hand $c_\beta$ vanishes exponentially in $\beta$, the inverse-temperature scaling will be milder, of the form $\beta \sim \log n$.

To illustrate the above points, let us consider the simple case of non-interacting qubits in a global magnetic field, $H = -\sum_{i=1}^{n} \sigma_i^z$. The partition function is given by:

$$Z = \sum_{k=0}^{n} \binom{n}{k} e^{-\beta(k-n/2)} = \left[2 \cosh(\beta/2)\right]^n$$  \hspace{2cm} (1)

Note that each energy level has a degeneracy that grows polynomially with $n$. The ground state probability on a thermal state is then given by $p_{GS} = e^{\beta E_0/2}/Z$, which we can then invert to write the inverse-temperature as:

$$\beta = -\ln\left(1 - p_{GS}^{-1/n}\right)$$  \hspace{2cm} (2)

If we pick $p_{GS}$ to be some small but fixed (independent of system size) number and take the large $n$ limit, we find that

$$\beta = \ln(n) - \ln(-\ln p_{GS}) + \frac{1}{2n} \ln p_{GS} + ...$$  \hspace{2cm} (3)

Therefore, we find that even for this simple problem, in order to maintain a constant ground state probability while the system size grows, we must scale the inverse-temperature logarithmically with system size.

A Grover search problem\textsuperscript{48,49} on the other hand yields the worst case scaling. In this case, we take a single state to have energy $-n$, while the remaining have energy $-n + 1$. The partition function is given by:

$$Z = e^{\beta n} \left[1 + (2^n - 1)e^{-\beta}\right].$$  \hspace{2cm} (4)

The ground state probability is given by $p_{GS} = 1/(1 + (2^n - 1)e^{-\beta})$. Inverting this for $\beta$, we find:

$$\beta = \ln\left(2^n - 1\right) - \ln\left(p_{GS}^{-1} - 1\right).$$  \hspace{2cm} (5)
Again, for a fixed and small $p_{GS}$, expanding for large $n$, we get:

$$\beta = n \ln 2 - \ln \left(p_{GS}^{-1} - 1\right) - 2^{-n} + \ldots$$

(6)

Therefore, in this case, $\beta$ must grow linearly with $n$ in order to maintain a constant $p_{GS}$. Note of course that the Grover Hamiltonian contains $n$-body interaction terms and so does not even satisfy our plausible assumptions for a scalable architecture from above.

IV. The Power of Pausing

In the previous sections, we argued that quantum annealing optimizers at a fixed temperature cannot be expected to provide scaling advantages in the limit where problem sizes go to infinity. It should be clear however that problems of practical relevance have finite sizes and so for these, asymptotic computational complexity considerations may be less relevant in practice.

One may therefore ask whether there are physical considerations that may help boost rather than impede the success probabilities of experimental optimizers. In Ref.50, it was demonstrated via experimenting with the D-Wave 2000Q quantum annealer$^{24}$ that even simple adaptions of the default annealing schedule, such as pausing, can result in a rather stark change in performance. In the adaptation to the annealing schedule considered there, the anneal is paused at an intermediate point for a certain amount of time. Typically each problem instance has is an ‘optimal pause point’ $s^*_{opt}$: a location during the anneal which a pause has a dramatic effect on the output distribution of the device. In particular, it was observed that this point is characterized with a more efficient sampling of the ground state and low lying energy levels, sometimes by orders of magnitude improvement in performance with respect to unpau sed annealing. The optimal stopping point $s^*_{opt}$ is found to occur after the location of minimum gap, as expected. The value of $s^*_{opt}$ is found to be relatively robust across a problem class; that is, the optimal pause point occurs in a similar, fairly narrow region for different problems of the same problem class, even for different sizes.

These results indicate that pausing after the minimum gap allows for the ground state to be repopulated. This boosting effect can be explained, at least qualitatively, in terms of the open system dynamics of the AME. After the minimum gap – after thermal excitations have significantly depopulated the ground state, the gap begins to grow again, and the instantaneous Gibbs state has a growing ground state population. Therefore, the AME dynamics is such that the system can begin to thermally relax back into the ground state. However, because the effective strength of the dissipative dynamics depends on the energy eigenstates of the system Hamiltonian (see Sec. II), there is a narrow region in $s$ where the dynamics is most efficient.

The results of Ref.51 provide positive evidence for the Boltzmann nature of the distributions returned from the annealer, even for the majority of problems for which a sensible freeze-out point does not exist. To obtain further insight into the nature of the distributions at the end of the anneal, two studies were performed in Ref.51. On small (12-qubit) problems for which it is possible to compute the quantum Boltzmann distribution for every point in the anneal, the final measured distribution and the projected quantum Boltzmann distributions were compared. For larger problems, it is not feasible to compute the quantum Boltzmann distributions throughout the anneal, but with the aid of recent entropic sampling techniques that enable accurate estimation of the eigenspectrum degeneracies for a class of planted solution problems52, fits to the final distribution, for multiple pause locations, to a classical Boltzmann distribution for $H_p$ with the effective temperature as a fitting parameter was performed.

The first study showed that the best performance occurs when the pause takes place after the minimum gap, confirming the qualitative picture provided by the AME. The best fit between the final distribution and the projected quantum Boltzmann distribution occurs somewhat after this pause point and is much higher than the physical temperature. The effective temperature depends on the device temperature, which is sampled only coarsely in time, and may not be the temperature at the qubits, and may also fluctuate within a single anneal. Control errors on the $J_g$ can also have significant effect on the distribution53.

The second study showed that for larger problems the best fit between the final distribution and a classical Boltzmann distribution for the problem Hamiltonian $H_p$ occurs when the pause is at the optimal pause point, indicating these samples are thermalizing more completely. In this case, an excellent fit between a classical Boltzmann distribution for $H_p$ and the final distribution occurs at all points, and even in cases in which no pause has been inserted. These results further suggest that the larger a problem is, the more likely it is to have a good fit to a classical Boltzmann distribution. Moreover, by varying the pause point, one can effectively vary the temperature at which one samples from the classical Boltzmann distribution.
The results generally find that the effective temperature is substantially higher than the quoted device temperature. As with the small problem study, discrepancies between the device temperature and the qubit temperature, fluctuations in this temperature during an anneal, and control errors masquerading as higher temperatures may all contribute to a higher effective temperature. While the device temperature is sampled only infrequently, the fluctuations in fitted temperature did fairly consistently track those of the actual device.

The extent to which these devices can be used as effective thermal samplers remains a somewhat open question, and most likely depends on the application; on the one hand it is known that solution sampling is typically biased on these types of annealing device, but on the other hand, several works suggest that for practical applications, such as in the context of machine learning from a thermal distribution, these devices can indeed be used effectively, even if the temperature is unknown.

V. Quantum Annealers Beyond Optimization

We have so far demonstrated that a device that prepares the thermal state might be unsuccessful in sampling the ground state if the temperature is not scaled down appropriately as the system size grows. Such a device might be better suited to perform tasks associated with thermal sampling, the most obvious of which is determining the thermal expectation values of physical observables. In this way, a quantum annealer with the capability to measure at any point in the anneal (we will return to this point later) can be viewed as a quantum Monte Carlo simulator, i.e., for measuring expectation values of certain physical observables of $H(s*)$. Denoting the ground state of $H(s*)$ by $|0(s*)\rangle = \sum_{z} c_{z} |z\rangle$ where $\{ |z\rangle \}$ is the set of computational basis states, a computational-basis measurement at $s^*$ yields a single bit-string $|z\rangle$ with probability $p(z) = |c(z)|^2$. Repetitions of the above process thus allow for the estimation of ground-state averages of diagonal properties such as $\langle H_{I}\rangle$, $\langle H_{I}^2\rangle$ and $\langle M_{z}^2\rangle$ where $M_{z} = \sum_{i} \sigma_{i}^{z}$ is the magnetization along the z-direction.

We can analyze the theoretical performance of such a device as follows. The total computational effort required from a quantum annealing simulator has two components: the annealing runtime $T_{s^*}$, and the number of samples (or anneals) $N_{s^*}$ required to obtain sufficient statistics for an accurate estimation of the various diagonal quantities, for a total computational cost of $N_{s^*}T_{s^*}$. Denoting the ground state expectation value of an observable $A$ is given by $\langle A \rangle$ with variance $\langle A^2 \rangle - \langle A \rangle^2$, then by the central limit theorem the sample mean of uncorrelated measurements $\mu_{A}$ will approach $\langle A \rangle$ with a sample variance $\sigma_{A}^2 = (\langle A^2 \rangle - \langle A \rangle^2) / N_{s^*}$. Taking as a criterion for convergence that the relative variance be smaller than some fixed $p^2$, the number of samples required will be proportional to:

$$N_{s^*}^{(QAS)} = \frac{\langle A^2 \rangle - \langle A \rangle^2}{p^2 \langle A \rangle^2}.$$  \hspace{1cm} (7)

For extensive observables, away from critical points this quantity will generically converge to a constant for sufficiently large $n$. The number of anneals thus does not play a role in the scaling of the computational effort required from quantum annealing simulators.

When can we therefore expect an advantage for quantum annealing simulators in this case? The most obvious case is for non-stoquastic Hamiltonians. Let us first consider the closed-system case: assuming the absence of energy level crossings, the adiabatic theorem dictates that a quantum annealing simulator will only require runtimes scaling as a polynomial in the inverse minimum gap $\Delta_{\text{min}}^{-1}$ along its interpolating path to prepare the ground state at $s^*$. On the other hand, we can expect on general grounds for Quantum Monte Carlo (QMC) algorithms to scale exponentially in the inverse temperature to estimate ground state properties, since the inverse temperature must be taken such that $\beta\Delta_{s^*} \gg 1$ where $\Delta_{s^*}$ is the ground state gap at the target $s^*$. Therefore, as long as $\Delta_{\text{min}} \sim \Delta_{s^*}$ we may expect dramatic advantages over QMC as a function of the inverse temperature in these regimes.

VI. Summary and Discussion

We discussed several fundamental limitations of physical quantum annealing devices that hinder their performance as well as one use for these devices that may be a viable path towards a demonstration of a quantum annealing speedup. First, we showed that in physical settings such as nonzero temperature, lead to an exponentially decaying success probability with system size for any fixed nonzero temperature or noise level. We further demonstrated how temperature must be scaled down to ensure that quantum annealing optimizers find nontrivial energy values with sub-exponential probabilities. These results shed light on existing benchmarking studies that have found no quantum speedups, identifying the above effects as likely culprits for their unfavorable performance.
It is worth noting however that the analysis is asymptotic in nature and is valid in the large \( n \) limit. In practice, if only a specific problem class of a certain finite size are of interest, then it would be possible in principle to engineer a device with a sufficiently low noise to mitigate the problems illustrated by our work for that problem class and size. For example, one can achieve lower effective temperatures by increasing the overall energy scale of the Hamiltonian. One way to achieve this is to develop quantum error correction techniques to effectively increase the energy scale of the Hamiltonian by coupling multiple qubits to form a single logical qubit\(^ {60-66} \). Alternatively, quantum error correction techniques can also be used to effectively decouple the system from the bath\(^ {67-70} \). That being said, it is important to note that a fault tolerant scalable approach to achieve this remains absent\(^ {71} \). Without a clear path to success and scalability, it remains to be seen whether analog quantum annealing devices will be a viable alternative to classical optimization algorithms.

In light of the peculiarities of analog quantum computing, we considered other scenarios that may be considered more promising for the future of quantum annealers. We presented general arguments that quantum annealers may very well exhibit quantum advantages if they are to be used as simulators rather than optimizers, competing against quantum Monte Carlo algorithms which encounter severe sign problems if the to-be-simulated Hamiltonian is nonstoquastic.

It should be noted however that we have not considered a complete survey of classical algorithms that can be used to study the ground state of non-stoquastic Hamiltonians. Depending on the accuracy required in the simulation, tensor network based methods\(^ {72} \) can be used. These methods are unaffected by the sign problem and are polynomially efficient in problem size but scale exponentially with the ‘bond dimension’. An interesting question is how the bond dimension needs to scale with problem size to maintain a suitable level of accuracy to accurately reproduce ground state expectation values of non-stoquastic Hamiltonians. Another candidate classical algorithm is to use an artificial neural network as an ansatz for the ground state wavefunction\(^ {73} \). How the number of nodes in the neural network and the number of steps needed to ‘train’ it to achieve the necessary accuracy for the ground state of non-stoquastic Hamiltonians is an important question that we leave for future study. We have also not addressed the practical limitations of implementing a quantum annealing simulator and how it might affect performance. It remains unclear to what extent the advantage we have suggested is maintained when decoherence effects are taken into account, such as interactions with a thermal environment\(^ {30,47,74-80} \) and implementation errors\(^ {53,58,60,74,81} \).

Efficient simulations of quantum many-body systems have far reaching implications. The sign problem is the single most important unresolved challenge in quantum many-body simulations, preventing a deep understanding of many macroscopic quantum phenomena\(^ {82,83} \). A physically meaningful way to circumvent the sign problem is therefore expected to immensely advance our understanding of the physics of a multitude of phenomena fundamental to academe, industry and government.

Acknowledgments
Compution for the work described in this paper was supported by the University of Southern California’s Center for High-Performance Computing (hpc.usc.edu) and by the Oak Ridge Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC05-00OR22725. The research is based upon work partially supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via the U.S. Army Research Office contract W911NF-17-C-0050. The views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the ODNI, IARPA, or the U.S. Government. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. This material is based on research sponsored by AFRL under agreement number FA8750-18-1-0109. The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright annotation thereon. We acknowledge the support of the Universities Space Research Association (USRA) Quantum AI Lab Research Opportunity Program.

References