Efficiency of quantum vs. classical annealing in nonconvex learning problems

Carlo Baldassi[a,b,1,2] and Riccardo Zecchina[a,c,1,2]

[a]Bocconi Institute for Data Science and Analytics, Bocconi University, 20136 Milan, Italy; [b]Istituto Nazionale di Fisica Nucleare, Sezione di Torino, 10125 Turin, Italy; and [c]Condensed Matter and Statistical Physics Group, International Centre for Theoretical Physics, 34151 Trieste, Italy

Edited by William Bialek, Princeton University, Princeton, NJ, and approved January 2, 2018 (received for review June 26, 2017)

Quantum annealers aim at solving nonconvex optimization problems by exploiting cooperative tunneling effects to escape local minima. The underlying idea consists of designing a classical energy function whose ground states are the sought optimal solutions of the original optimization problem and add a controllable quantum transverse field to generate tunneling processes. A key challenge is to identify classes of nonconvex optimization problems for which quantum annealing remains efficient while thermal annealing fails. We show that this happens for a wide class of problems which are central to machine learning. Their energy landscapes are dominated by local minima that cause exponential slowdown of classical thermal annealers while simulated quantum annealing converges efficiently to rare dense regions of optimal solutions.

nonconvex optimization | machine learning | quantum annealing | neural networks | statistical physics

Quantum annealing (QA) aims at finding low-energy configurations of nonconvex optimization problems by a controlled quantum adiabatic evolution, where a time-dependent many-body quantum system which encodes for the optimization problem evolves toward its ground states so as to escape local minima through multiple tunneling events (1–5). Classical simulated annealing (SA) uses thermal fluctuations for the same computational purpose, and Markov chains based on this principle are among the most widespread optimization techniques across science (6). Quantum fluctuations are qualitatively different from thermal fluctuations, and in principle, QA algorithms could lead to extremely powerful alternative computational devices.

In the QA approach, a time-dependent quantum transverse field is added to the classical energy function leading to an interpolating Hamiltonian that may take advantage of correlated fluctuations mediated by tunneling. Starting with a high transverse field, the quantum model system can be initialized in its ground state, i.e., all spins aligned in the direction of the field. The adiabatic theorem then ensures that by slowly reducing the transverse field, the system remains in the ground state of the interpolating Hamiltonian. At the end of the process, the transverse field vanishes, and the systems ends up in the sought ground state of the classical energy function. The original optimization problem would then be solved if the overall process could take place in a time bounded by some low-degree polynomial in the size of the problem. Unfortunately, the adiabatic process can become extremely slow. The adiabatic theorem requires the rate of change of the Hamiltonian to be smaller than the square of the gap between the ground state and the first excited state (7–9). For small gaps, the process can thus become inefficient. Exponentially small gaps are not only possible in worst-case scenarios, but have also been found to exist in typical random systems where comparative studies between quantum and classical annealing have so far failed in displaying quantum exponential speed-up, e.g., at first-order phase transition in quantum spin glasses (10, 11) or 2D spin-glass systems (12–14). More positive results have been found for ad hoc energy functions in which global minima are planted in such a way that tunneling cascades can become more efficient than thermal fluctuations (4, 15). As far as the physical implementations of quantum annealers is concerned, studies have been focused on discriminating the presence of quantum effects rather than on their computational effectiveness (16–18).

Consequently, a key open question is to identify classes of relevant optimization problems for which QA can be shown to be exponentially faster than its classical thermal counterpart.

Here, we give an answer to this question by providing analytic and simulation evidence of exponential speed-up of quantum vs. classical SA for a representative class of random nonconvex optimization problems of basic interest in machine learning. The simplest example of this class is the problem of training binary neural networks (described in detail below): Very schematically, the variables of the problem are the (binary) connection weights, while the energy measures the training error over a given dataset.

These problems have been very recently found to possess a rather distinctive geometrical structure of ground states (19–22): The free-energy landscape has been shown to be characterized by the existence of an exponentially large number of metastable states and isolated ground states and a few regions where the ground states are dense. These dense regions, which had previously escaped the equilibrium statistical physics analysis (23, 24), are exponentially rare, but still possess a very high local internal

Significance

Quantum annealers are physical quantum devices designed to solve optimization problems by finding low-energy configurations of an appropriate energy function by exploiting cooperative tunneling effects to escape local minima. Classical annealers use thermal fluctuations for the same computational purpose, and Markov chains based on this principle are among the most widespread optimization techniques. The fundamental mechanism underlying quantum annealing consists of exploiting a controllable quantum perturbation to generate tunneling processes. The computational potentialities of quantum annealers are still under debate, since few ad hoc positive results are known. Here, we identify a wide class of large-scale nonconvex optimization problems for which quantum annealing is efficient while classical annealing gets stuck. These problems are of central interest to machine learning.

Author contributions: C.B. and R.Z. designed research, performed research, contributed new reagents/analytic tools, analyzed data, and wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

[1]C.B. and R.Z. contributed equally to this work.

[2]To whom correspondence may be addressed. Email: carlo.baldassi@unibocconi.it or riccardo.zecchina@unibocconi.it.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1711456115/-/DCSupplemental.

entropy: They are composed of ground states that are surrounded, at extensive but relatively small distances, by exponentially many other ground states. Under these circumstances, classical SA (as any Markov chain satisfying detailed balance) gets trapped in the metastable states, suffering ergodicity breaking and exponential slowing down toward the low-energy configurations. These problems have been considered to be intractable for decades and display deep similarities with disordered spin-glass models, which are known to never reach equilibrium.

The large deviation analysis that has unveiled the existence of the rare dense regions has led to several novel algorithms, including a Monte Carlo scheme defined over an appropriate objective function (20) that bears close similarities with a quantum Monte Carlo (QMC) technique based on the Suzuki–Trotter transformation (5). Motivated by this analytical mapping and by the geometrical structure of the dense and degenerate ground states which is expected to favor zero-temperature kinetic processes (25, 26), we have conducted a full analytical and numerical statistical physics study of the QA problem, reaching the conclusion that in the quantum limit, the QMC process, i.e., simulated QA (SQA), can equilibrate efficiently, while the classical SA gets stuck in high-energy metastable states. These results generalize to multilayered networks.

While it is known that other quasioptimal classical algorithms for the same problems exist (20, 27, 28), here, we focus on the physical speed-up that a QA approach could provide in finding rare regions of ground states. We provide physical arguments and numerical results supporting the conjecture that the real-time QA dynamics behaves similarly to SQA.

As far as machine learning is concerned, dense regions of low-energy configurations (i.e., quasiflat minima over macroscopic length scales) are of fundamental interest, as they are particularly well-suited for making predictions given the learned data: On the one hand, these regions are by definition robust with respect to fluctuations in a sizable fraction of the weight configurations and, as such, are less prone to fit the noise. On the other hand, an optimal Bayesian estimate, resulting from a weighted consensus vote on all configurations, would receive a major contribution from one of such regions, compared with a narrow minimum; the centroid of the region (computed according to any reasonable metric which correlates the distance between configurations with the network outcomes) would act as a representative of the region as a whole (29). In this respect, it is worth mentioning that in deep learning (30), all of the learning algorithms which lead to good prediction performance always include effects of a systematically injected noise in the learning phase, a fact that makes the equilibrium Gibbs measure not the stationary measure of the learning protocols and drives the systems toward wide minima. We expect that these results can be generalized to many other classes of nonconvex optimization problems where local entropy plays a role, ranging from robust optimization to physical disordered systems.

Quantum gate-based algorithms for machine learning exist; however, the possibility of a physical implementation remains a critical issue (31).

Energy Functions

As a working example, we first consider the problem of learning random patterns in single-layer neural network with binary weights, the so-called binary perceptron problem (23). This network maps vectors of N inputs $\xi \in \{-1,+1\}^N$ to binary outputs $\tau = \pm 1$ through the nonlinear function $\tau = \mathrm{sgn}\,(\sigma \cdot \xi)$, where $\sigma \in \{-1,+1\}^N$ is the vector of synaptic weights. Given αN input patterns $\{\xi^\mu\}_{\mu=1}^{\alpha N}$ with $\mu = 1, ..., \alpha N$ and their corresponding desired outputs $\{\tau^\mu\}_{\mu=1}^{\alpha N}$, the learning problem consists in finding σ such that all input patterns are simultaneously classified correctly, i.e., $\mathrm{sgn}\,(\sigma \cdot \xi^\mu) = \tau^\mu$ for all μ. Both the components of the input vectors ξ_i^μ and the outputs τ^μ are independent identically distributed unbiased random variables ($P(x) = \frac{1}{2}\delta(x-1) + \frac{1}{2}\delta(x+1)$). In the binary framework, the procedure for writing a spin Hamiltonian whose ground states are the sought optimal solutions of the original optimization problem is well known (32). The energy E of the binary perceptron is proportional to the number of classification errors and can be written as

$$E(\{\sigma_j\}) = \sum_{\mu=1}^{\alpha N} \Delta_\mu^n \Theta(-\Delta_\mu), \quad \Delta_\mu \doteq \frac{\tau^\mu}{\sqrt{N}} \sum_{j=1}^{N} \xi_j^\mu \sigma_j \qquad [1]$$

where $\Theta(x)$ is the Heaviside step function: $\Theta(x) = 1$ if $x > 0$, $\Theta(x) = 0$ otherwise. When the argument of the Θ function is positive, the perceptron is implementing the wrong input–output mapping. The exponent $n \in \{0, 1\}$ defines two different forms of the energy functions which have the same zero-energy ground states and different structures of local minima. The equilibrium analysis of the binary perceptron problem shows that in the large size limit, and for $\alpha < \alpha_c \simeq 0.83$ (23), the energy landscape is dominated by an exponential number of local minima and of zero-energy ground states that are typically geometrically isolated (33), i.e., they have extensive mutual Hamming distances. For both choices of n, the problem is computationally hard for SA processes (34): In the large N limit, a detailed balanced stochastic search process gets stuck in metastable states at energy levels of order $O(N)$ above the ground states.

Following the standard SQA approach, we identify the binary variables σ with one of the components of physical quantum spins, say, σ^z, and we introduce the Hamiltonian operator of a model of N quantum spins with the perceptron term of Eq. **1** acting in the longitudinal direction z and a magnetic field Γ acting in the transverse direction x. The interpolating Hamiltonian reads:

$$\hat{H} = E\left(\{\hat{\sigma}_j^z\}\right) - \Gamma \sum_{j=1}^{N} \hat{\sigma}_j^x \qquad [2]$$

where $\hat{\sigma}_j^z$ and $\hat{\sigma}_j{}^x$ are the spin operators (Pauli matrices) in the z and x directions. For $\Gamma = 0$, one recovers the classical optimization problem. The QA procedure consists of initializing the system at large β and Γ, and slowly decreasing Γ to 0. To analyze the low-temperature phase diagram of the model, we need to study the average of the logarithm of the partition function $Z = \mathrm{Tr}\left(e^{-\beta \hat{H}}\right)$. This can be done by using the Suzuki–Trotter transformation, which leads to the study of a classical effective Hamiltonian acting on a system of y interacting Trotter replicas of the original classical system coupled in an extra dimension:

$$H_{\mathrm{eff}}\left(\{\sigma_j^a\}_{j,a}\right) = \frac{1}{y}\sum_{a=1}^{y} E\left(\{\sigma_j^a\}_j\right) - \frac{\gamma}{\beta}\sum_{a=1}^{y}\sum_{j=1}^{N} \sigma_j^a \sigma_j^{a+1} - \frac{NK}{\beta} \qquad [3]$$

where the $\sigma_j^a = \pm 1$ are Ising spins, $a \in \{1, \ldots, y\}$ is a replica index with periodic boundary conditions $\sigma_j^{y+1} \equiv \sigma_j^1$, $\gamma = \frac{1}{2}\log \coth\left(\frac{\beta\Gamma}{y}\right)$ and $K = \frac{1}{2} y \log\left(\frac{1}{2}\sinh\left(2\frac{\beta\Gamma}{y}\right)\right)$.

The replicated system needs to be studied in the limit $y \to \infty$ to recover the so-called path integral continuous quantum limit and to make the connection with the behavior of quantum devices (14). The SQA dynamical process samples configurations from an equilibrium distribution, and it is not necessarily equivalent to the real-time Schrödinger equation evolution of the system. A particularly dangerous situation occurs if the ground states of the system encounter first-order phase transitions which are associated with exponentially small gaps (10, 35, 36) at finite

N. As discussed below, this appears not to be the case for the class of models we are considering.

Connection with the Local Entropy Measure

The effective Hamiltonian Eq. **3** can be interpreted as many replicas of the original systems coupled through one-dimensional periodic chains, one for each original spin (Fig. 1*B*). Note that the interaction term γ diverges as the transverse field Γ goes to 0. This geometrical structure is very similar to that of the robust ensemble (RE) formalism (20), where a probability measure that gives higher weight to rare dense regions of low-energy states is introduced. There, the main idea is to maximize $\Phi(\sigma^\star) = \log \sum_{\{\sigma\}} e^{-\beta E(\sigma) - \lambda \sum_{j=1}^{N} \sigma_j \sigma_j^\star}$, i.e., a "local free entropy" where λ is a Lagrange parameter that controls the extensive size of the region around a reference configuration $\sigma^\star$. One can then build a new Gibbs distribution $P(\sigma^\star) \propto e^{y\Phi(\sigma^\star)}$, where $-\Phi$ has the role of an energy and y of an inverse temperature: In the limit of large y, this distribution concentrates on the maxima of Φ.

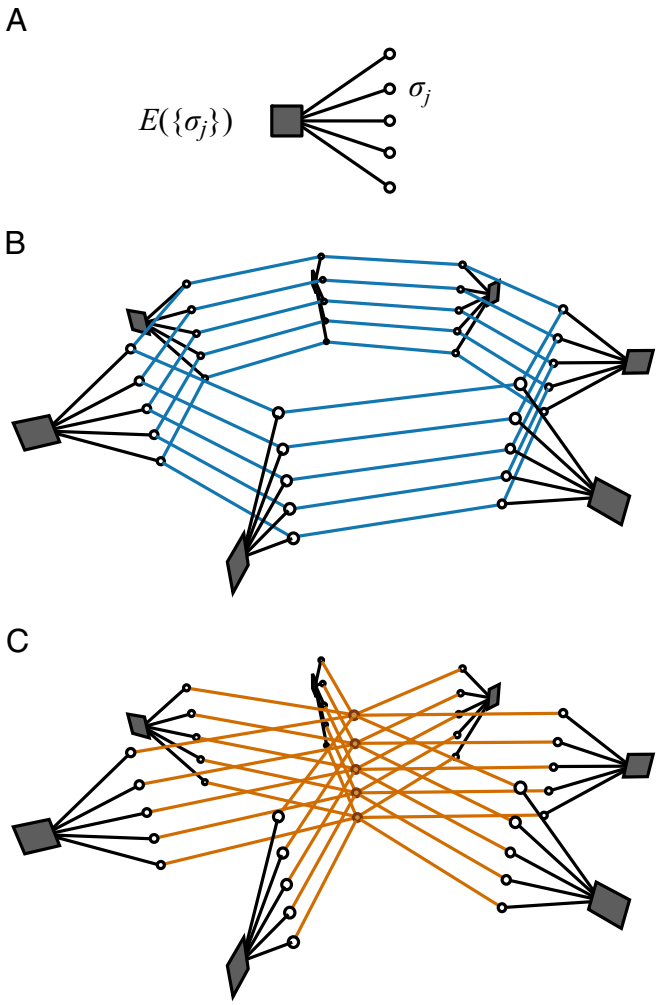

Fig. 1. Topology of the Suzuki–Trotter vs. robust ensemble (RE) representations. (*A*) The classical objective function we wish to optimize which depends on N discrete variables $\{\sigma_j\}$ ($N = 5$ in the picture). (*B*) Suzuki–Trotter interaction topology: y replicas of the classical system ($y = 7$ in the picture) are coupled by periodic one-dimensional chains, one for each classical spin. (*C*) RE interaction topology: y replicas are coupled through a centroid configuration. In the limit of large N and large y (quantum limit) and for strong interaction couplings, all replicas are forced to be close, and the behavior of the two effective models is expected to be similar.

Upon restricting the values of y to be integer (and large), $P(\sigma^\star)$ takes a factorized form yielding a replicated probability measure $P_{\rm RE}(\sigma^\star, \sigma^1, \ldots, \sigma^y) \propto e^{-\beta H_{\rm eff}^{\rm RE}(\sigma^\star, \{\sigma_j^a\})}$ where the effective energy is given by

$$H_{\rm eff}^{\rm RE}\left(\sigma^\star, \{\sigma_j^a\}_{j,a}\right) = \sum_{a=1}^{y} E\left(\{\sigma_j^a\}_j\right) - \frac{\lambda}{\beta} \sum_{a=1}^{y} \sum_{j=1}^{N} \sigma_j^a \sigma_j^\star \qquad [4]$$

As in the Suzuki–Trotter formalism, $H_{\rm eff}^{\rm RE}\left(\sigma^\star, \{\sigma_j^a\}_{j,a}\right)$ corresponds to a system with an overall energy given by the sum of y individual "real replica energies" plus a geometric coupling term; in this case, however, the replicas interact with the "reference" configurations $\sigma^\star$ rather than among themselves (Fig. 1*C*).

The Suzuki–Trotter representation and the RE formalism differ in the topology of the interactions between replicas and in the scaling of the interactions, but for both cases, there is a classical limit, $\Gamma \to 0$ and $\lambda \to \infty$, respectively, in which the replicated systems are forced to correlate and eventually coalesce in identical configurations. For nonconvex problems, these will not in general correspond to configuration dominating the original classical Gibbs measure.

For the sake of clarity, we should remind that in the classical limit and for $\alpha < \alpha_c$, our model presents an exponential number of far-apart isolated ground states which dominate the Gibbs measure. At the same time, there exist rare clusters of ground states with a density close to its maximum possible value (high local entropy) for small but still macroscopic cluster sizes (19). This fact has several consequences: No further subdivision of the clusters into states is possible; the ground states are typically $O(1)$ spin flip connected (19); and a trade-off between tunneling events and exponential number of destination states within the cluster is possible.

Phase Diagram: Analytical and Numerical Results

Thanks to the mean field nature of the energetic part of the system, Eq. **3**, we can resort to the replica method for calculating analytically the phase diagram. As discussed in *SI Appendix*, this can be done under the so-called static approximation, which consists of using a single-parameter q_1 to represent the overlaps along the Trotter dimension, $q_1^{ab} = \left\langle \frac{1}{N} \sum_{j=1}^{N} \sigma_j^a \sigma_j^b \right\rangle \approx q_1$. Although this approximation crudely neglects the dependency of q_1^{ab} from $|a-b|$, the resulting predictions show a remarkable agreement with numerical simulations.

In Fig. 2, we report the analytical predictions for the average classical component of the energy of the quantum model as a function of the transverse field Γ. We compare the results with the outcome of extensive simulations performed with the reduced-rejection-rate (RRR) Monte Carlo method (37), in which Γ is initialized at 2.5 and gradually brought down to 0 in regular small steps, at constant temperature, and fixing the total simulation time to $\tau N y \cdot 10^4$ (as to keep constant the number of Monte Carlo sweeps when varying N and y). Additional details are reported in *Materials and Methods* and *SI Appendix*. The size of the systems, the number of samples, and the number of Trotter replicas are scaled up to large values, so that both finite size effects and the quantum limit are kept under control. A key point is to observe that the results do not degrade with the number of Trotter replicas: The average ground-state energy approaches a limiting value, close to the theoretical prediction, in the large y quantum limit. The results appear to be rather insensitive to both N and the simulation time-scaling parameter τ. This indicates that Monte Carlo appears to be able to equilibrate efficiently, in a constant (or almost constant) number of sweeps, at each Γ. The analytical prediction for the classical energy only appears to display a relatively small systematic offset (due to the static

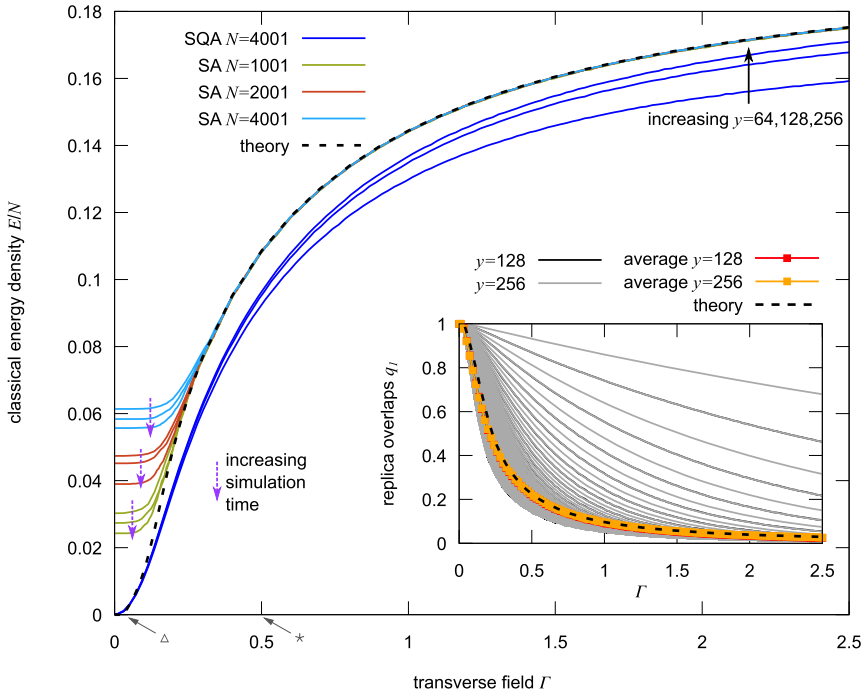

Fig. 2. Classical energy density (i.e., longitudinal component of the energy, divided by *N*) as a function of the transverse field Γ (single layer problems with $\alpha = 0.4$ and $n = 0$, 15 independent samples per curve). The SQA simulations at $\beta = 20$ approach the theoretical prediction as y increases (cf black arrow). The results do not change significantly when varying *N* (the curves with $N = 1001$ or $N = 2001$ are indistinguishable from the ones displayed at this level of detail) or the overall simulation time. All SA simulations instead got stuck and failed to equilibrate at low enough temperatures (small equivalent Γ). The results are noticeably worse for larger *N*, and doubling or quadrupling the simulation time does not help much (cf purple arrows). (*Inset*) Trotter replicas overlap q_1^{ab} (same data as for the main figure). The theoretical prediction is in remarkably good agreement with the average value measured from the simulations (the $y = 128$ curve is barely visible under the $y = 256$ one). The gray curves show the overlaps at varying distances along the Trotter dimension: The topmost one is the overlap between neighboring replicas $q_1^{a(a+1)}$, then there is the overlap between second-neighbors $q_1^{a(a+2)}$, and so on (cf Fig. 1). The $y = 128$ curves are essentially hidden under the $y = 256$ ones and can only be seen from their darker shade, following an alternating pattern.

approximation) at intermediate values of Γ, while it is very precise at both large and small Γ; the expectation of the total Hamiltonian, on the other hand, is in excellent agreement with the simulations (*SI Appendix*).

In the same plot, we display the behavior of classical SA simulated with a standard Metropolis–Hastings scheme, under an annealing protocol in β that would follow the same theoretical curve as SQA if the system were able to equilibrate (*Materials and Methods* and *SI Appendix*): As expected (34), SA gets trapped at very high energies (increasing with problem size; in the thermodynamic limit, it is expected that SA would remain stuck at the initial value $0.5N$ of the energy for times which scale exponentially with N). Alternative annealing protocols yield analogous results; the exponential scaling with N of SA on binary perceptron models had also been observed experimentally in previous results, e.g., in refs. 21 and 38.

In Fig. 2 *Inset*, we report the analytical prediction for the transverse overlap parameter q_1, which quite remarkably reproduces fairly well the average overlap as measured from simulations.

In Fig. 3, we provide the profiles of the classical energy minima found for different values of Γ in the case of SQA and different temperatures for SA. These results are computed analytically by the cavity method (see *Materials and Methods* and *SI Appendix* for details) by evaluating which is the most probable energy found at a normalized Hamming distance d from a given configuration. As it turns out, throughout the annealing process, SQA follows a path corresponding to wide valleys, while SA gets stuck in steep metastable states. The quantum fluctuations reproduced by the SQA process drive the system to converge toward wide flat regions, despite the fact that they are exponentially rare compared with the narrow minima.

The physical interpretation of these results is that quantum fluctuations lower the energy of a cluster proportionally to its size or, in other words, that quantum fluctuations allow the system to lower its kinetic energy by delocalizing; see refs. 25, 26, and 39 for related results. Along the process of reduction of the transverse field, we do not observe any phase transition which could induce a critical slowing down of the QA process, and we expect SQA and QA to behave similarly (11, 36).

This is in agreement with the results of a direct comparison between the real-time quantum dynamics and the SQA on small systems ($N = 21$): As reported in *SI Appendix*, we have performed extensive numerical studies of properly selected small instances of the binary perceptron problem, comparing the results of SQA and QA and analyzing the results of the QA process and the properties of the Hamiltonian. To reproduce the conditions that are known to exist at large values of N, we have selected instances for which a fast annealing schedule SA gets trapped at some positive fraction of violated constraints, and yet the problems display a sufficiently high number of solutions. We found that the agreement between SQA and QA on each sample is excellent. The measurements on the final configurations reached by QA qualitatively confirm the scenario described above, that QA is attracted toward dense, low-energy regions without getting stuck during the annealing process. Finally, the analysis of the gap between the ground state of the system and the first excited state as Γ decreases shows no signs of the kind of phenomena which would typically hamper the performance of QA in other models: There are no vanishingly small gaps at finite Γ (compare discussion in the introduction). We benchmarked all these results with "randomized" versions of the same samples, in which we randomly permuted the classical energies associated

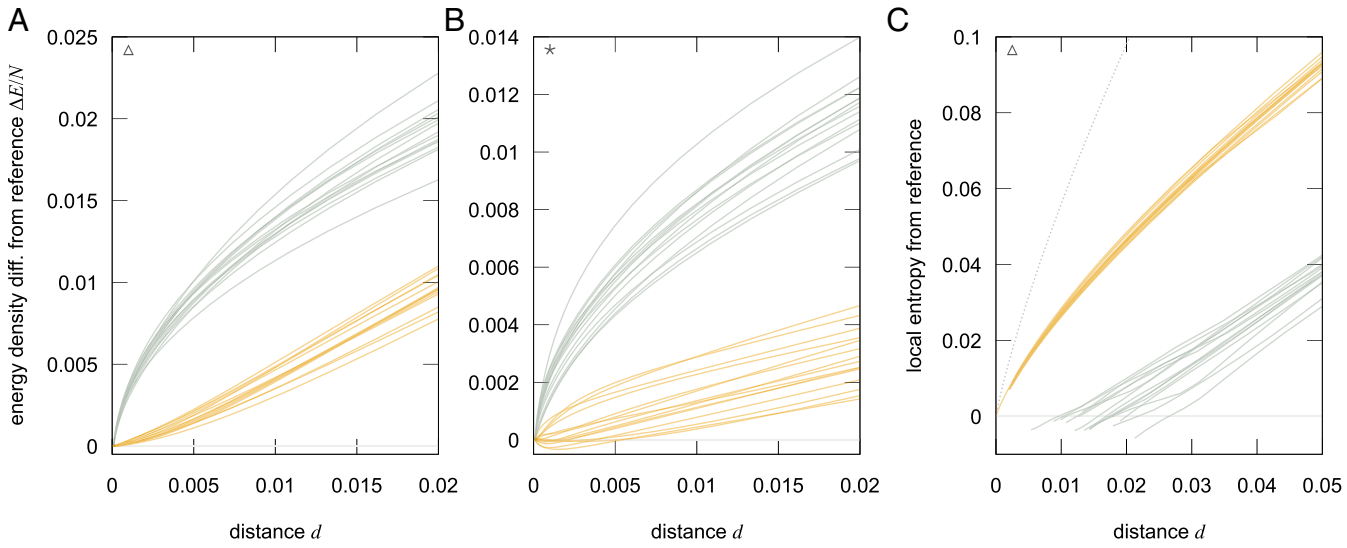

Fig. 3. (*A* and *B*) Energetic profiles (in terms of the classical energy E; Eq. **1**) around the configurations reached during the annealing process, comparing SQA (orange lower curves) with SA (gray top curves). The profiles represent the most probable value of the energy density shift $\Delta E/N$ with respect to the reference point when moving away from the reference at a given normalized Hamming distance d. The curves refer to the data shown in Fig. 2, using two different times in the annealing process, marked with $\triangle$ and $\star$. For SQA, we show the results for 15 instances with $N = 4{,}001$, $y = 256$, $\tau = 4$, using the mode of the replicas $\sigma_j^\star = \text{sgn}\left(\sum_{a=1}^{y} \sigma_j^a\right)$ as the reference point; for SA, we show 15 samples for $N = 4{,}001$ and $\tau = 16$. These results show a marked qualitative difference in the type of landscape that is typically explored by the two algorithms: The local landscape of SQA is generally much wider, while SA is typically working inside narrow regions of the landscape which tend to trap the algorithm eventually. (*C*) Local entropy, i.e., the logarithm of the number of solutions surrounding the reference point at a given distance d for the same configurations of *A*. The SQA configurations (orange curves at the top) are located in regions with exponentially many solutions surrounding them (although these regions are not maximally dense, as can be seen from the comparison with the dashed curve representing the overall number of surrounding configurations at that distance). The SA configurations (gray curves at the bottom) are far away from these exponentially dense regions (the local entropy has a gap around $d = 0$).

with each spin configuration, so as to keep the distribution of the classical energy levels while destroying the geometric structure of the states. Indeed, for these randomized samples, we found that the gaps nearly close at finite $\Gamma \simeq 0.4$, and that, correspondingly, the QA process fails to track the ground state of the system, resulting in a much-reduced probability of finding a solution to the problem.

As concluding remarks, we report that the models with $n = 0$ and $n = 1$ have phase diagrams which are qualitatively very similar (for the sake of simplicity, here we reported the $n = 0$ case only). The former presents at very small positive values of Γ a collapse of the density matrix onto the classical one, whereas the latter ends up in the classical state only at $\Gamma = 0$.

For the sake of completeness, we have checked that the performance of SQA in the $y \to \infty$ quantum limit extends to more complex architectures which include hidden layers; the details are reported in *SI Appendix*.

Conclusions

We conclude by noticing that, at variance with other studies on spin-glass models in which the evidence for QA outperforming classical annealing was limited to finite values of y, thereby just defining a different type of classical SA algorithms, in our case the quantum limit coincides with the optimal behavior of the algorithm itself. We believe that these results could play a role in many optimization problems in which optimality of the cost function needs to also meet robustness conditions (i.e., wide minima). As far as learning problems are concerned, it is worth mentioning that for the best-performing artificial neural networks, the so-called deep networks (30), there is numerical evidence for the existence of rare flat minima (40) and that all of the effective algorithms always include effects of systematic injected noise in the learning phase (41), which implies that the equilibrium Gibbs measure is not the stationary measure of the learning protocols. For the sake of clarity, we should remark that our results are aimed to suggest that QA can equilibrate efficiently, whereas SA cannot; i.e., our notion of quantum speed-up is relative to the same algorithmic scheme that runs on classical hardware. Other classical algorithms for the same class of problems, besides the above-mentioned ones based on the RE and the SQA itself, have been discovered (27, 38, 42–44); however, all of these algorithms are qualitatively different from QA, which can provide a huge speed-up by manipulating single physical bits in parallel. Thus, the overall solving time in a physical QA implementation (neglecting any other technological considerations) would have, at worst, only a mild dependence on N.

Our results provide further evidence that learning can be achieved through different types of correlated fluctuations, among which quantum tunneling could be a relevant example for physical devices.

Materials and Methods

Simulated QA Protocol. All SQA simulations were performed by using the RRR Monte Carlo method (37). We fixed the total number of spin flip attempts at $\tau N y \cdot 10^4$ and followed a linear protocol (divided in 30τ steps) for the annealing of Γ. In Fig. 2, we show the results for $N = 4001$ and $\tau = 4$; the results for $N = 1001$, 2001 and for $\tau = 1$, 2 were essentially indistinguishable at that level of detail.

Classical SA Protocol. The results for SA presented in Fig. 2 used an annealing protocol in β designed to make a direct comparison with QA: We found analytically a curve $\beta_{\text{equiv}}(\Gamma)$ such that the classical equilibrium energy would be equal to the longitudinal component of the quantum system energy. The classical equilibrium energy was computed from the equations in ref. 23. The result is shown in *SI Appendix*, Fig. S1. The SA protocol thus consisted of setting $\beta = \beta_{\text{equiv}}(\Gamma)$ and decreasing linearly Γ from 2.5 to 0, like for the QA case. We fixed the total number of spin flip attempts at $\tau N \cdot 10^4$ and used $\tau = 4$, 8, 16; as for the QA case, the annealing process was divided in 30τ steps. If the system were able to equilibrate, it would follow

the theoretical curve (dashed black line in Fig. 2), which it does only for high temperatures.

Other more standard annealing protocols (e.g., linear, exponential, or logarithmic) yielded very similar qualitative results, as expected from the analysis of ref. 34.

Estimation of the Local Energy and Entropy Landscapes. To compute the local landscapes of the energy and the entropy around a reference configuration, Fig. 3, we used the belief propagation algorithm. We added an external field in the direction of the configuration of interest to focus on regions surrounding that configuration. The strength of the field allowed us to control the size of the region (parameter d in Fig. 3). Typical energies are computed by setting the temperature to infinity, while local entropies are computed by setting the temperature to 0. The details of the algorithm are presented in *SI Appendix*.

Real-Time QA Simulations on Small Instances. The real-time quantum dynamics simulations on small systems were performed by solving the time-dependent Schrödinger equation for the Hamiltonian of Eq. **2** by using the short iterative Lanczos method (45), which consists of computing the evolution with the Lanczos algorithm, at fixed Γ for a short time interval Δt, then lowering Γ by a small fixed amount $\Delta\Gamma$, and iterating until $\Gamma = 0$.

ACKNOWLEDGMENTS. We thank G. Santoro, B. Kappen, and F. Becca for discussions. This work was supported by Office of Naval Research Grant N00014-17-1-2569.

1. Ray P, Chakrabarti BK, Chakrabarti A (1989) Sherrington-Kirkpatrick model in a transverse field: Absence of replica symmetry breaking due to quantum fluctuations. *Phys Rev B* 39:11828–11832.
2. Finnila A, Gomez M, Sebenik C, Stenson C, Doll J (1994) Quantum annealing: A new method for minimizing multidimensional functions. *Chem Phys Lett* 219:343–348.
3. Kadowaki T, Nishimori H (1998) Quantum annealing in the transverse ising model. *Phys Rev E* 58:5355–5363.
4. Farhi E, et al. (2001) A quantum adiabatic evolution algorithm applied to random instances of an np-complete problem. *Science* 292:472–475.
5. Das A, Chakrabarti BK (2008) Colloquium: Quantum annealing and analog quantum computation. *Rev Mod Phys* 80:1061–1081.
6. Moore C, Mertens S (2011) *The Nature of Computation* (Oxford Univ Press, Oxford).
7. Born M, Fock V (1928) Beweis des adiabatensatzes. *Zeitschrift Phys A Hadrons Nuclei* 51:165–180.
8. Landau L (1932) Zur theorie der energieubertragung. II. *Phys Z Sowjetunion* 2:1–13.
9. Zener C (1932) Non-adiabatic crossing of energy levels. *Proc R Soc Lond A Math Phys Eng Sci* 137:696–702.
10. Altshuler B, Krovi H, Roland J (2010) Anderson localization makes adiabatic quantum optimization fail. *Proc Natl Acad Sci USA* 107:12446–12450.
11. Bapst V, Foini L, Krzakala F, Semerjian G, Zamponi F (2013) The quantum adiabatic algorithm applied to random optimization problems: The quantum spin glass perspective. *Phys Rep* 523:127–205.
12. Santoro GE, Martoňák R, Tosatti E, Car R (2002) Theory of quantum annealing of an ising spin glass. *Science* 295:2427–2430.
13. Martoňák R, Santoro GE, Tosatti E (2002) Quantum annealing by the path-integral Monte Carlo method: The two-dimensional random ising model. *Phys Rev B* 66:094203.
14. Heim B, Rønnow TF, Isakov SV, Troyer M (2015) Quantum versus classical annealing of ising spin glasses. *Science* 348:215–217.
15. Rønnow TF, et al. (2014) Defining and detecting quantum speedup. *Science* 345:420–424.
16. Johnson MW, et al. (2011) Quantum annealing with manufactured spins. *Nature* 473:194–198.
17. Boixo S, et al. (2014) Evidence for quantum annealing with more than one hundred qubits. *Nat Phys* 10:218–224.
18. Langbein W, et al. (2004) Control of fine-structure splitting and biexciton binding in In x Ga 1- x as quantum dots by annealing. *Phys Rev B* 69:161301.
19. Baldassi C, Ingrosso A, Lucibello C, Saglietti L, Zecchina R (2015) Subdominant dense clusters allow for simple learning and high computational performance in neural networks with discrete synapses. *Phys Rev Lett* 115:128101.
20. Baldassi C, et al. (2016) Unreasonable effectiveness of learning neural networks: From accessible states and robust ensembles to basic algorithmic schemes. *Proc Natl Acad Sci USA* 113:E7655–E7662.
21. Baldassi C, Ingrosso A, Lucibello C, Saglietti L, Zecchina R (2016) Local entropy as a measure for sampling solutions in constraint satisfaction problems. *J Stat Mech Theor Exp* 2016:P023301.
22. Baldassi C, Gerace F, Lucibello C, Saglietti L, Zecchina R (2016) Learning may need only a few bits of synaptic precision. *Phys Rev E* 93:052313.
23. Krauth W, Mézard M (1989) Storage capacity of memory networks with binary couplings. *J Phys France* 50:3057–3066.
24. Sompolinsky H, Tishby N, Seung HS (1990) Learning from examples in large neural networks. *Phys Rev Lett* 65:1683–1686.
25. Foini L, Semerjian G, Zamponi F (2010) Solvable model of quantum random optimization problems. *Phys Rev Lett* 105:167204.
26. Biroli G, Zamponi F (2012) A tentative replica theory of glassy helium 4. *J Low Temp Phys* 168:101–116.
27. Hubara I, Courbariaux M, Soudry D, El-Yaniv R, Bengio Y (2016) Quantized neural networks: Training neural networks with low precision weights and activations. arXiv:1609.07061.
28. Courbariaux M, Bengio Y, David JP (2015) BinaryConnect: Training deep neural networks with binary weights during propagations. *Advances in Neural Information Processing Systems* (Curran Associates, Red Hook, NY), Vol 28, pp 3105–3113.
29. MacKay DJ (2003) *Information Theory, Inference and Learning Algorithms* (Cambridge Univ Press, Cambridge, UK).
30. LeCun Y, Bengio Y, Hinton G (2015) Deep learning. *Nature* 521:436–444.
31. Aaronson S (2015) Read the fine print. *Nat Phys* 11:291–293.
32. Barahona F (1982) On the computational complexity of ising spin glass models. *J Phys A Math Gen* 15:3241–3253.
33. Huang H, Kabashima Y (2014) Origin of the computational hardness for learning with binary synapses. *Phys Rev E* 90:052813.
34. Horner H (1992) Dynamics of learning for the binary perceptron problem. *Zeitschrift Physik B Condens Matter* 86:291–308.
35. Bapst V, Semerjian G (2012) On quantum mean-field models and their quantum annealing. *J Stat Mech Theor Exp* 2012:P06007.
36. Bapst V, Semerjian G (2013) Thermal, quantum and simulated quantum annealing: Analytical comparisons for simple models. *J Phys Conf Ser* 473:012011.
37. Baldassi C (2017) A method to reduce the rejection rate in Monte Carlo Markov chains. *J Stat Mech Theor Exp* 2017:033301.
38. Baldassi C, Braunstein A, Brunel N, Zecchina R (2007) Efficient supervised learning in networks with binary synapses. *Proc Natl Acad Sci USA* 104:11079–11084.
39. Markland TE, et al. (2011) Quantum fluctuations can promote or inhibit glass formation. *Nat Phys* 7:134–137.
40. Keskar NS, Mudigere D, Nocedal J, Smelyanskiy M, Tang PTP (2016) On large-batch training for deep learning, 1609 large-batch training for deep learning: Generalization gap and sharp minima. arXiv:1609.04836.
41. Bottou L, Curtis FE, Nocedal J (2016) Optimization methods for large-scale machine learning. arXiv:1606.04838.
42. Braunstein A, Zecchina R (2006) Learning by message-passing in neural networks with material synapses. *Phys Rev Lett* 96:030201.
43. Baldassi C (2009) Generalization learning in a perceptron with binary synapses. *J Stat Phys* 136:902–916.
44. Baldassi C, Braunstein A (2015) A max-sum algorithm for training discrete neural networks. *J Stat Mech Theor Exp* 2015:P08008.
45. Schneider BI, Guan X, Bartschat K (2016) Chapter five-time propagation of partial differential equations using the short iterative Lanczos method and finite-element discrete variable representation. *Adv Quan Chem* 72:95–127.