This is the accepted manuscript made available via CHORUS. The article has been published as:

**Uncertain fate of fair sampling in quantum annealing**

Mario S. Kötz, Guglielmo Mazzola, Andrew J. Ochoa, Helmut G. Katzgraber, and Matthias Troyer


DOI: [10.1103/PhysRevA.100.030303](https://doi.org/10.1103/PhysRevA.100.030303)
Quantum annealing (QA) [1–9] is a heuristic designed to harness the advantages of quantum mechanics to solve optimization problems. The performance of QA and, in particular, QA machines such as the D-Wave Systems Inc. devices are controversial to date [10–32]. Most studies have focused on finding the minimum value of a binary quadratic cost function (problem Hamiltonian), yet less on the variety of solutions obtained when repeating the optimization procedure multiple times. Important applications that rely on sampling, such as SAT-based probabilistic membership filters [33–36], propositional model counting and related problems [37–39], or machine learning [40, 41] rely on ideally uncorrelated states. This sought-after fair sampling ability of an algorithm, i.e., the ability to find (ideally all) states associated with a cost function with (ideally) the same probability, is thus of importance for a variety of applications. Moreover, the ability of an algorithm to sample ground states with similar probability is directly related its ergodicity which strongly influences the efficiency of optimization and sampling techniques.

Following small-scale studies [42], Ref. [43] recently performed systematic experiments on the D-Wave 2X annealer. The results demonstrated that quantum annealers using a transverse-field driver are biased samplers, an effect also observed in previous studies [13, 14, 44]. Matsuda et al. [42] conjectured that more complex drivers might alleviate this bias, something we test in this work.

Binary optimization problems can be mapped onto k-local spin Hamiltonians. Without loss of generality we study problem Hamiltonians with N degrees of freedom in a 2-site basis of the form

\[ \mathcal{H}_V = - \sum_{i,j=1}^{N} J_{ij} \sigma_i^z \sigma_j^z, \]  

where \( \sigma_i^z \) is the z-component of the Pauli operator acting on site i. Note that local biases can also act on the variables. For such a problem Hamiltonian, in principle, a driver of the form

\[ \mathcal{H}_{x,n} = \sum_{n=1}^{N} \Gamma^{x,n} [\otimes \sigma^x]^n \]  

would induce transitions between all states and therefore ensure a fair sampling, provided the anneal is performed slow enough. Unfortunately, such driver is hard to engineer and, at best, one can expect drivers of the form \( \mathcal{H}_{x,N} = - \sum_j \Gamma^x \sigma_j^x + \sum_{j,k} K_{j,k} \sigma_j^z \sigma_k^z \). Quantum fluctuations are induced by the driver and then reduced to sample states from the problem Hamiltonian, i.e., \( \mathcal{H}(t) = (1 - t/T) \mathcal{H}_{x,n} + (t/T) \mathcal{H}_V \), where \( t \in [0, T] \), T the annealing time, and n the order of the interactions in the driver. For an infinitely-slow anneal, the adiabatic theorem [2, 45] ensures that for \( t = T \) a (ground) state of the problem Hamiltonian is reached. It is therefore desirable to know if after an infinite amount of repetitions, the process results in all minimizing states, i.e., fair sampling.

Here we analyze the behavior of more complex drivers of the form \( \mathcal{H}_{x,n} \) (\( n > 1 \)) on the fair sampling abilities of QA. Following Ref. [42] we first study small systems where the Schrödinger equation can be integrated using QuTiP [46]. We have exhaustively analyzed all possible graphs with up to \( N = 6 \) with both ferromagnetic and antiferromagnetic interactions and show in Fig. 1 paradigmatic examples that illustrate different scenarios using drivers with \( n \leq 2 \). Even for some of these small instances, in some cases the inclusion of higher-order driver terms does not remove the bias. If we
anneal adiabatically, i.e., $T$ large enough, the instantaneous ground states is never left, which means towards the end of annealing at $T - \lambda$ (for a small $\lambda > 0$) the system is in the ground state of $\mathcal{H}(T - \lambda)$. This observation is key to predicting the sampling probabilities for different degenerate ground states. These probabilities are given by squaring the amplitudes of the lowest eigenvector of $\mathcal{H}(T - \lambda)$, assuming for now the small contribution from the driver lifts the degeneracies. Because $\mathcal{H}(T - \lambda)$ can be viewed as $\mathcal{H}_D$ perturbed by $\mathcal{H}_{\alpha,n}$, we analyze fair sampling using a perturbative approach [47]. To better quantify the fair-sampling behavior of a given system, we introduce the term “hard suppression” (i.e., total suppression) if the sampling probability is 0 for a particular ground-state configuration at the end of the anneal and the term “soft suppression” if a particular state is undersampled by a certain finite fraction in comparison to other minimizing configurations. Finally, we complement these studies with quantum Monte Carlo simulations for large two-dimensional Ising spin-glass problems following Ref. [43] and discuss the effects of higher-order drivers. Our results show that QA is not well suited for sampling applications, unless post-processing techniques are implemented [48].

**Perturbation theory.**— In the following, we show how to determine the sampling probabilities, as well as the influence the driver has on it. In short, if we apply $\mathcal{H}_D$ as a perturbation of strength $\lambda$ to $\mathcal{H}_P$, some degeneracies will be lifted, i.e., the perturbed ground-state-space is smaller. The ground-state-space is never left during an adiabatic anneal, hence it will not be possible to reach the entire ground-state-space of the unperturbed Hamiltonian by annealing in the generic case. This analysis holds for any driver Hamiltonian $\mathcal{H}_D$, not just the stoquastic $\mathcal{H}_{\alpha,n}$-type drivers we use in this work. In nondegenerate perturbation theory, the first-order corrected wave function $|n\rangle$ is given by $|n\rangle = |n^0\rangle + \lambda \sum_{m \neq n} \frac{\langle m^0|H_D|n^0\rangle}{E_{m^0} - E_{n^0}} |m^0\rangle$, where $|n^0\rangle$ are the eigenstates and $E_{n^0}^0$ the eigenvalues of the unperturbed Hamiltonian $\mathcal{H}_P$. If states $m \neq n$ are degenerate, i.e., $E_{m}^0 = E_{n}^0$, there is a singularity. To avoid it, degenerate perturbation theory requires linear combinations $|\alpha^0\rangle$ which satisfy $\langle \alpha^0|H_D|\beta^0\rangle \sim \delta_{\alpha,\beta}$ in every degenerate subspace. This ensures that the corrected wave function does not diverge due to singularities. We focus on the ground-state subspace, but the procedure is identical for any subspace. Given $k$ ground-states $|n_{gs}^0\rangle$ of $\mathcal{H}_D$ with energy $E_{gs}^0$, we need to form the $k \times k$ subspace matrix $V_{n,m} = \langle n_{gs}^0|H_D|n_{gs}^0\rangle$. Because $\mathcal{H}_D$ is hermitian, $V$ is too. Every hermitian matrix can be diagonalized by a unitary transformation $(U^{-1}VU = D)$ and we find the correct linear combinations $|\alpha^0\rangle$ in the columns of $U$. It satisfies $\langle \alpha_{gs}|V|\beta_{gs}\rangle \sim \delta_{\alpha,\beta}$ since $D$ is diagonal. The diagonal entries of $D$ are the eigenvalues of $V$ and also the first order energy corrections $E_{\alpha}^1$. We need to pick the lowest eigenvalue $E_{\alpha,\text{low}}$ and find the corrected ground state energy $E_{\text{GS}} = E_{gs}^0 + E_{\alpha,\text{low}}$. The corresponding $l$ eigenvectors $|\alpha_{gs}^0\rangle$ will now determine the sampling behavior, since the annealing state will be in their span. The following scenarios can occur:

(i) $l = 1$ — In this case $p_l = \langle n_{gs}^0|\alpha_{gs,\text{low}}^0\rangle^2$, because there is a single state $|\alpha_{gs,\text{low}}^0\rangle$. If sampling is fair, it will remain fair, regardless of how much the higher energy eigenvalues of $\mathcal{H}_P$ change during the adiabatic anneal. If certain states have $p_l = 0$, $|\alpha_{gs}^0\rangle$ will never be available at the end of the anneal.

(ii) $l > 1$ — Let $A$ be the $k \times m$ matrix consisting of all $l$ $|\alpha_{gs,\text{low}}^0\rangle$. If there is a vector $x$ such that $Ax = y$ and $y_1, y_2 = 1$...
for all $i$, then fair sampling is potentially possible according to first order. If there exists an $i$ such that $y_i = 0$ for all $x$, then that ground-state is never found. The same argument can be made for biased sampling where there is no suppression but certain states are over-sampled.

(iii) $V$ is zero – All eigenvalues $E^1_{\alpha} = 0$ and the sampling probabilities are determined by second-order perturbation, i.e., the probabilities depend on higher eigenvalues of $H_{\parallel}$ (see Fig. 3).

$$H_{x,1} = \begin{pmatrix}
0 & -1 & -1 & 0 & -1 & 0 & 0 & 0 \\
-1 & 0 & 0 & -1 & 0 & -1 & 0 & 0 \\
0 & -1 & -1 & 0 & 0 & 0 & 0 & -1 \\
-1 & 0 & 0 & -1 & 0 & 0 & -1 & 0 \\
0 & -1 & 0 & 0 & -1 & 0 & 0 & -1 \\
0 & 0 & -1 & 0 & -1 & -1 & 0 & -1 \\
0 & 0 & 0 & -1 & 0 & -1 & -1 & 0 \\
0 & 0 & 0 & 0 & -1 & 0 & -1 & -1 \\
\end{pmatrix}$$

$$V = \begin{pmatrix}
0 & 0 & 0 \\
0 & 0 & -1 \\
0 & -1 & 0 \\
\end{pmatrix} \Rightarrow k_{\text{low}} = \frac{1}{\sqrt{2}} \begin{pmatrix}0 \\ 1 \\ 1 \end{pmatrix}$$

Figure 2. To obtain the sampling probabilities, the ground-state eigenvectors $|gs\rangle$ need to be known, represented here as shaded rows and columns in the matrix $H_{x,1}$, since the solution of the diagonal $H_{\parallel}$ is a classical one. One then needs to analyze the subspace $V$ that is formed by restricting the driver (here $H_{x,1}$) to space spanned by the ground-states of $H_{\parallel}$. The lowest eigenvector(s) determine the sampling probabilities. In this example, there is one lowest eigenvector and the first ground state corresponding to the top column (first row) is suppressed (all spins up) and will never be sampled in an adiabatic anneal.

The second-order perturbation terms only play a relevant role if $V$ is trivial. If $l > 0$, the sampling behavior is determined by $V$ which does not depend on $H_{\parallel}$. This means that the sampling behavior is purely a property of the driver Hamiltonian $H_{x,n}$ and the ground-state eigenvectors of $H_{\parallel}$. We have verified this on numerous small systems, as well as structured and random-coupling systems with direct integration and were always able to predict the sampling probabilities that correspond to the state found after the anneal.

Figure 1(a) is the example studied in Ref. [42], where $H_{x,2}$ lead to fair sampling. There are 6 degenerate ground states, 2 of which are suppressed. With a driver of the form $H_{x,1}$ we obtain $l > 1$, meaning that there are multiple $|gs\rangle$ states that determine the sampling. However, the suppressed states have $p_i = 0$. In Fig. 1(b) we show a more complex example – the smallest problem we were able to find that has $l = 1$ and one state where $p_i = 0$. It is a 12-fold degenerate system with two states fully suppressed when $H_{x,1}$ is used as a driver. The fact that $l = 1$ could be a reason why the suppression sets in earlier during the anneal. This case is problematic for annealing schedules that are fast quenches, because there is a much smaller window during the anneal where the total ground-state probability is approximately unity and the suppressed state has not yet reached zero probability. Using $H_{x,2}$ results in fair sampling. Figure 1(c) shows a system that has 6 ground states with 2 ground states in total suppression. Using $H_{x,1}$ as a driver, we obtain $l = 1$ and a unique $|gs\rangle$ with 2 totally suppressed states with zero probability. For $H_{x,2}$, $l > 1$ we obtain multiple $|gs\rangle$ states. However, two states are totally suppressed. Using $H_{x,3}$ as a driver results in $l = 1$ and a unique $|gs\rangle$. However, there is a soft suppression of two ground states (not shown). Finally, using $H_{x,4}$ we obtain $l = 1$ and fair sampling. The case shown in Fig. 1(d) reveals the undersampled states when $H_{x,1}$ is replaced by $H_{x,2}$. More precisely, it changes from $l > 1$ with 4 soft suppressed states to $l = 1$ with the previously 2 oversampled state now being undersampled. Using a driver $H_{x,3}$ results in $l = 1$ and fair sampling.

Figure 3 shows a problem where by changing the strength of $J_{3,4}$ one can change the sampling bias arbitrarily. Note that changing $J_{3,4} = -1.2$ to $-1.8$ does shift the relative energies of the ground state and the various low exited states, but does not change their order. In terms of perturbation theory, $V$ is trivial, and second-order perturbations dictate the behavior of the system. Because there are terms $\propto 1/(E_i - E_{gs})$ with $E_{gs}$ the ground state energy, shifting the energy levels $E_i$ will influence the sampling.

Quantum Monte Carlo results.—To corroborate our results with larger systems, we perform a fair-sampling study analogous to the one done in Ref. [43] for two-dimensional Ising spin glasses on a square lattice with periodic boundary conditions. The couplers are chosen from $J_{i,j} \in \left\{\pm 1, \pm 2, \pm 4\right\}$. This ensures that degeneracies are small. The coupler-configuration space is mined for specific degeneracies as done in Ref. [43]. Figure 4 shows representative rank-ordered probabilities to find different minimizing configurations using simulated annealing (SA) [49–51], as well as transverse-field simulated quantum annealing (SQA-$H_{x,1}$) [5, 52–56] and simulated quantum annealing with a stoquastic two-spins driver (SQA-$H_{x,2}$) [57] [58]. The data are averaged over 100 disorder realizations. While the data for SA for this particular problem show a fair sampling of all minimizing configurations, neither a transverse-field $H_{x,1}$ nor a more complex $H_{x,2}$ driver can remove the bias. This suggests that even if QA ma-
machines with more complex $H_{x,2}$ drivers are constructed, sampling will remain unfair unless post-processing is applied [48].

The close connection between SQA and QA performances is discussed in Refs. [53, 54].

Figure 4. Rank-ordered probability $p_{GS}$ to find different degenerate states for a two-dimensional Ising spin glass with $N = 8^2 = 64$ and a ground-state degeneracy of 32. The data are averaged over 100 disorder realizations. For each instance, 500 independent runs are performed and the probability to find a given ground-state configuration computed. While simulated annealing (SA) samples close to fair, both SQA-$H_{x,1}$ and SQA-$H_{x,2}$ show a clear bias in the sampling. In particular, there is no notable improvement of using a driver of the form $H_{x,2}$ over a transverse-field driver $H_{s,1}$. We have also simulated systems with up to $N = 12^2 = 144$ variables and ground states with up to 96-fold degeneracy obtaining similar results. Note that the bias becomes more pronounced for increasing system size $N$ (not shown).

Effects of more complex drivers.— The following section shows that any driver (stoquastic or non-stoquastic) needs to be sufficiently dense to sample fair for generic larger systems. To predict the sampling probabilities it is sufficient to know $V$ (except when $V$ is trivial). $V$ can be constructed with only the ground-state eigenvectors of $H_{1}$ (no eigen-energies needed) and the driver $H_{x,n}$. This can be used to analyze different drivers—without specifying a concrete problem Hamiltonian $H_{1}$—by merely sampling from possible ground-state combinations. As an example consider a 2-fold degeneracy in a 5-spin system. We want to test the driver for all possible ground-state combinations, we can exhaustively generate all the ground-state pairs, i.e., $N(N-1)/2$, where $N = 2^5$ and check $V$ for each one pair, to analyze the sampling behavior. For larger $N$, we sample instead of searching exhaustively.

Figure 5 shows how probable it is for a random degeneracy and a ground-state combination to be sampled according to the following categories:

fair — All ground states have the same probability.

soft — At least one ground state is soft suppressed with a ratio smaller than $1 : 100$ (least likely vs most likely).

hard — At least one ground state is soft suppressed with a ratio larger than $1 : 100$ or not found at all, i.e., total suppression. For better visibility in Fig. 5 we combine these two cases. However, most of the time the suppression is total.

highord — The matrix $V$ is trivial. Higher-order perturbation will determine the sampling behavior. In the generic case of random couplings this leads to both soft or hard suppression.

In all cases and for $H_{x,n}$ with $n \leq 8$ we use $\Gamma^{x,n} = 1$. Using different values for the different amplitudes leads to worse sampling, because the matrix $V$ has multiple different entries. A random matrix has a unique eigenvector which is not parallel to $(1, 1, \ldots, 1, 1)$ in the generic case. Hence, introducing more variety into $V$ leads to more unique (and unfair) $|\alpha_{gs}^{(0)}\rangle$. How the ratio of soft to hard suppression is influenced by this was not investigated, since it is unfair in the generic case. Repeating multiple annealing runs with individually randomized $\Gamma^{x,n}_{i,j}$ and averaging improves the sampling but, if not dense enough, will not be able to remove all hard suppression in a generic case.

Conclusions.— We have studied the necessary ingredients needed for quantum annealing to sample ground states fairly. From Fig. 5 we surmise that a fairly dense driver is needed to obtain fair sampling. Carefully controlling the anneal with additional parameters, for example as shown in Ref. [59] might help mitigate the bias, however, this remains to be tested. We
do emphasize, however, that a $\mathcal{H}_{\text{trans}}$ driver with the typical annealing modus operandi used in current hardware will not yield a fair sampling of states and performs comparatively to a vanilla transverse-field driver.

ACKNOWLEDGMENTS

M. S. K. thanks Dominik Gresch for insightful discussions that accelerated the discovery process. The large scale simulated quantum annealing calculation where done using the Monch Cluster at ETH Zurich. H. G. K. would like to thank Salvatore Mandrà for multiple discussions and acknowledges the ARC Centre for Excellence in All-Sky Astrophysics in 3D (ASTRO 3D) East Coast Writing Retreat 2018 for support in preparing the manuscript. H. G. K. acknowledges support from the NSF (Grant No. DMR-1151387). H. G. K.'s research is based upon work supported by the Office of the Director of National Intelligence (ODNI), Intelligence Advanced Research Projects Activity (IARPA), via Interagency Umbrella Agreement No. IA1-1198. 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. We thank the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources (Stampede Cluster) and Texas A&M University for access to their Ada, and Lonestar clusters.


[55] *Simulation Parameters for SQA-H_{N,1}: H = 10, N_{trotter} = 1024, T = 0.001, steps = 1000.*

[56] *Simulation Parameters for SQA-H_{N,2}: H = H_{x} = 2, N_{trotter} = 64, T = 0.08, equilibrate = 10000, steps = 2000000.*
