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

Heavy Tails in the Distribution of Time to Solution for Classical and Quantum Annealing*

Damian S. Steiger, Troels F. Rønnow, and Matthias Troyer

Phys. Rev. Lett. **115**, 230501 — Published 4 December 2015

DOI: [10.1103/PhysRevLett.115.230501](https://doi.org/10.1103/PhysRevLett.115.230501)

Heavy tails in the distribution of time-to-solution for classical and quantum annealing

Damian S. Steiger,¹ Troels F. Rønnow,^{1,2} and Matthias Troyer¹

¹*Theoretische Physik, ETH Zurich, 8093 Zurich, Switzerland*

²*Nokia Technologies, Broers Building, 21 JJ Thomson Avenue, CB3 0FA Cambridge, United Kingdom*

(Dated: November 12, 2015)

For many optimization algorithms the time-to-solution depends not only on the problem size but also on the specific problem instance and may vary by many orders of magnitude. It is then necessary to investigate the full distribution and especially its tail. Here we analyze the distributions of annealing times for simulated annealing and simulated quantum annealing (by path integral quantum Monte Carlo) for random Ising spin glass instances. We find power-law distributions with very heavy tails, corresponding to extremely hard instances, but far broader distributions – and thus worse performance for hard instances – for simulated quantum annealing than for simulated annealing. Fast, non-adiabatic, annealing schedules can improve the performance of simulated quantum annealing for very hard instances by many orders of magnitude.

Non-convex optimization problems arise in a wide range of areas, from industrial applications to scientific research. Many challenging optimization problems belong to the class of non-deterministic polynomial-time hard (NP-hard) problems. For these problems, which includes the famous traveling salesman problem, no efficient algorithm is known that scales polynomially with problem size. In the absence of efficient exact solvers, often heuristic algorithms are the best choice. For many algorithms, and especially probabilistic annealing strategies [1–5], the time to find a (near)-optimal solution may depend not just on the problem size N , but may vary greatly – by many orders of magnitude – from instance to instance. In such cases one usually quotes the *typical* performance given by the median time-to-solution.

Despite its common use, the median time-to-solution is often not the quantity that dictates efficiency, since one generally wants to solve more than half of the problem instances. In particular, in the case of broad distributions of the time-to-solutions, higher quantiles (such as the 99-th percentile) may be better indicators of the performance of an optimization algorithm. In this Letter we demonstrate that when analyzing the performance of optimization algorithms, it is very important to consider the tails of the time-to-solution distribution and not only averages or median performances.

As a specific example we consider the problem of finding the ground states of Ising spin glasses on Chimera graphs. In these models N Ising spins s_i that can take values ± 1 should be chosen to minimize the total energy,

$$H_P = - \sum_{\langle i,j \rangle} J_{ij} s_i s_j, \quad (1)$$

where the coupling constants J_{ij} take random values on the edges of the so-called Chimera graph. For more information see the Supplemental Material [7]. This quasi-two-dimensional graph with eight spins per unit cell is non-planar, which makes the problem of finding the ground state of the Ising spin glass NP-hard [8].

The Ising spin glass problem on the Chimera graph has recently received special attention, since this graph is implemented in the optimization devices built by the Canadian company D-Wave systems [9], whose performance has been controversially discussed in the recent literature [6, 10–14]. We summarize, in Fig. 1 previous results [6] of the scaling of the time-to-solution (TTS) as a function of problem size N for three different approaches: a simulated annealer (SA), a simulated quantum annealer (SQA) using path integral quantum Monte Carlo, and a D-Wave Two device (DW2). Note one striking feature of these results: there is a huge spread in TTS between easy and hard instances (the 99th percentile), which varies by three orders of magnitude for SA, and by even more for SQA and DW2. The hard instances completely dominate the TTS. Below we will quantitatively analyze the distribution of TTS for the hardest instances, and show how it can be substantially reduced for SQA by using non-adiabatic annealing schedules.

Simulated annealing (SA) is inspired by annealing, a process in metallurgy by which a material is heated up and then slowly cooled down in order to relieve internal stresses. In 1983, Kirkpatrick and coworkers [1] showed that simulating an annealing process by a Monte Carlo method can be used as a general purpose heuristic solver for optimization problems, which is widely used in many application areas [2]. By slowly decreasing the temperature the system can relax into a low-energy state, escaping local minima by thermal activation over barriers. Implementation details are discussed in the Supplemental Material [7].

Inspired by the phenomenon of quantum tunneling, quantum annealing (QA) uses quantum fluctuations instead of thermal fluctuations to explore the configuration space [3, 15–18]. It employs a time-dependent Hamiltonian,

$$H(t) = A(t)H_D + B(t)H_P, \quad (2)$$

where $H_D = - \sum_i \sigma_i^x$ is a driver Hamiltonian implementing quantum dynamics, and the problem Hamiltonian

FIG. 1. Scaling of the annealing time to find the ground state with 99% probability for a bimodal Ising spin glass with couplings ± 1 on Chimera graphs with N spins using (a) a simulated annealer (SA), (b) a simulated quantum annealer (SQA) and (c) a D-Wave Two device (DW2). Results (with standard errors) are taken from Ref. [6]. We observe an enormous spread in the scaling of different percentiles, ranging from easy (1%) to hard (99%) problem instances.

$H_P = -\sum_{\langle i,j \rangle} J_{ij} \sigma_i^z \sigma_j^z$ implements the Ising spin glass Eq. (1) using quantum spin-1/2 particles. σ^x and σ^z are Pauli matrices. At the beginning of the annealing schedule when $A(0) = 1$ and $B(0) = 0$ we start with only H_D . The spins are initialized in the ground state aligning along the x-axis. During the annealing process we decrease $A(t)$ and increase $B(t)$, so that at the end of the annealing at time $t = t_a$, we end up with the Ising spin glass of Eq. (1). The Hamiltonian is changed slowly enough such that the system stays in (or close to) the ground state at all times. This process takes place at a constant finite temperature T and therefore constant inverse temperature $\beta = 1/(k_B T)$, where k_B is the Boltzmann constant. The D-Wave devices have been designed to implement quantum annealing using superconducting flux qubits [9] in a physical device.

Quantum annealing can also be implemented as simu-

FIG. 2. Distribution of mean number of repetitions τ required to find the ground state for 20000 problems with 200 spins on Chimera topology for SA and SQA (with $\beta = 10$). The tails of these distribution functions are decaying polynomially slow and there is a qualitative difference between SA and SQA: For SA the tail is decaying faster and the overall distribution function is more compact.

lated quantum annealing (SQA) on a classical computer using a path integral quantum Monte Carlo simulation [4, 5, 19]. Furthermore, by replacing the quantum spins in Eq. (2) with two-dimensional classical rotor magnets a mean-field version (MFA) [11] can be obtained as a semi-classical version of SQA [20]. Details of the SQA and MFA algorithms and their implementations are presented in the Supplemental Material [7]. Note that annealing times t_a for SA, SQA and MFA are measured in numbers of sweeps, where one sweep is defined as one attempted update for each spin.

For each problem size N we created 20000 different random problem instances with couplings $J_{ij} = \pm 1$. For each instance, after determining the exact ground state energy using an exhaustive search algorithm we repeatedly performed SA, SQA and MFA simulations until we found a ground state at least 100 times for each algorithm. For SQA with $N = 288$ spins and annealing time $t_a = 10^4$ sweeps we never found the ground state of the hardest instances despite billions of attempts. We thus focus on the case of $N = 200$ spins for our quantitative analysis, where all algorithms found the ground states of all problem instances with reasonable effort. From the number of repetitions required to find the ground state one hundred times we then calculated the single-run success probability s and the mean number of repetitions $\tau = 1/s$ required to find the ground state.

Plotting histograms of τ for 20000 instances with $N = 200$ we arrive at the distributions shown in Fig. 2. We see very slowly decaying power-law tails, consistent with the wide spread of quantiles in Fig. 1. We also see that while the median times are very similar ($\tau = 2.1$ for SA and $\tau = 2.2$ for SQA), there is a substantial quantitative difference between the distributions. For SA tails extend up to large values of $\tau \approx 100$, but for SQA to values of $\tau \approx 10^7$ which is more than 5 orders of magnitude

FIG. 3. Generalized Pareto probability density function $w_{\xi,0,\sigma}(x) = \frac{d}{dx}W_{\xi,0,\sigma}(x)$ plotted on a linear-log scale. The qualitative tail behavior depends on the shape parameter ξ . For $\xi < 0$ the probability density is bounded, for $\xi = 0$ it decays exponentially and for $\xi > 0$ the tail is heavy and decays as a power-law.

larger. The median times are thus neither indicative of the time needed to solve hard problem instances, nor of the relative performance of the two algorithms!

Power-law tails occur in many areas of physics, geoscience and economics [21–23]. While the origin of such tail behavior is not fully understood, they play an important role in complex systems. They have previously been observed in the context of spin glasses, see for example Ref. [24–30]. To quantitatively analyze such power-law distributions, we will use extreme value theory (EVT) to obtain a good estimate for the tail distribution of τ given only a limited data set and specifically to calculate the so-called shape parameter ξ , which gives information about its power-law tail. Let us denote by $F(x) = \Pr\{X \leq x\}$ the unknown distribution function of a random variable X , e.g. the distribution of τ for a specific algorithm. If F satisfies certain conditions (described in the Supplemental Material [7]) then the Balkema-de Haan-Pickands (BdHP) theorem [31, 32] states that the tail of the distribution function, $F(x)$ for x larger than a high threshold u , is approximated by a generalized Pareto (GP) distribution function $W_{\xi,\tilde{u}_u,\tilde{\sigma}_u}(x)$, i.e.

$$F(x) \approx W_{\xi,\tilde{u}_u,\tilde{\sigma}_u}(x) \quad \text{for } x \geq u. \quad (3)$$

The GP distribution function with threshold/location parameter u , shape parameter ξ , scale parameter $\sigma_u > 0$ is given by

$$W_{\xi,u,\sigma_u}(x) := 1 - \left(1 + \xi \left(\frac{x-u}{\sigma_u}\right)\right)^{-1/\xi}, \quad (4)$$

which is defined on $\{x : x \geq u \text{ and } (1 + \xi(x-u)/\sigma_u) \geq 0\}$ [21, 33, 34]. For $\xi = 0$ one takes the limit $\xi \rightarrow 0$,

$$W_{0,u,\sigma_u}(x) = 1 - \exp(-(x-u)/\sigma_u), \quad x \geq u. \quad (5)$$

The qualitative behavior of the GP distribution function is given by its shape parameter ξ , which turns out to

FIG. 4. Distribution of mean number of repetitions τ required to find the ground state for 20000 problem instances with $N = 200$ spins for (a) SQA with $\beta = 10$ and (b) SA with either the optimal number or 10^4 sweeps. While the distribution changes only slightly for SA, surprisingly the tails in SQA decrease much more rapidly when annealing faster.

be very useful when benchmarking different algorithms. The shape parameter is independent of the threshold u and of any linear transformations on X , e.g. considering the number of spin flips to find the ground state instead of τ . The probability density has three distinct tail behaviors, shown in Fig. 3, depending on the sign of ξ . It is bounded for $\xi < 0$, exponentially decaying for $\xi = 0$ and slowly decaying as a power-law for $\xi > 0$, which we also refer to as having a heavy tail. Note that the k -th moment of $W_{\xi,u,\sigma_u}(x)$ diverges for $k \geq 1/\xi$ [35].

If a heuristic algorithm has $\xi > 0$, then hard instances dominate the average time-to-solution. If the shape parameter $\xi \geq 1$ the average run-time of an instance is infinite. In our case of discrete couplings ± 1 , there is only a finite set of instances for each N . The conditions of the BdHP theorem are thus not satisfied because the distribution of τ is discrete with only a finite number of points in its support [36, p. 217]. Nevertheless we can still use the GP distribution to analyze our data, and obtain excellent fits (see Supplemental Material [7]). The main consequence of a finite set of instances is that $F(\tau)$ differs from the GP distribution function for the hardest instances, and thus for example the mean τ is not infinite but dominated by the hardest instances. In the Supplemental Material [7] we show that the mean is not converging for $\xi \geq 1$ using 20000 problem instances.

We start, by investigating the effects of annealing time t_a on the distribution. In Fig. 4 we compare the distributions for $t_a = 10^4$ sweeps (which for SQA with $\beta = 10$ was found to give good correlations with a D-Wave One device [10]) to simulations run with the (size-dependent) optimal t_a^{opt} that minimizes the total effort $t_a \tau$ (see Ref. 6 and the Supplemental Material citeSM). The surprising result is that performing SQA with a fast schedule of only $t_a^{\text{opt}} = 150$ sweeps leads to a more compact distribution

FIG. 5. Correlation plot of (a) SQA with $\beta = 10$ and (b) SA for all 20000 instances with $N = 200$ spins using t_a^{opt} and $t_a = 10^4$ sweeps respectively. Points below the white line (τ, τ) are those where the success probability s increases upon annealing faster. This is the case for 2312 instances for SQA but only 11 for SA. This increase in s for the hard instances explains the more rapid decrease of the tail in SQA. The grey line indicates points with equal total effort $t_a \tau$. Instances below the grey line (all but one for SQA and all but 51 for SA) are those which get solved more efficiently with a faster annealing time.

with much faster decaying tails. While easy instances may need a few more repetitions, hard instances need far less repetitions to find the ground state.

As the correlation plot in Fig. 5 shows, the relative hardness of an instance does not change significantly when changing t_a , but the single-run success probability s of SQA increases for more than 10% of the instances when SQA is run faster. This is opposite to SA where for almost all instances s decreases when annealing faster. In both cases, however, the total effort τt_a decreases for almost all instances. Even though more repetitions τ are needed, this is more than compensated by the faster annealing time. For SQA, the combined effect of needing 688 times fewer repetitions and performing only 150 instead of 10^4 sweeps per repetition reduces the total effort for the hardest of our instances by a factor 45866!

To quantify the difference in the tail distribution functions we fit them to GP distribution functions and plot

FIG. 6. Shown are the fitted shape parameters ξ with standard errors for SA, MFA and SQA. The different versions of SQA have different inverse temperature β and t_a , with $t_a^{\text{opt}}(N)$ varying from 20 to 400 sweeps depending on N and β . For $\xi > 0$ (indicated by a dashed line) the tail of the distribution function is heavy and for $\xi \geq 1$ (indicated by a dotted line) the mean diverges. The k th moment diverges for $k \geq 1/\xi$.

the size-dependence of the shape parameter ξ for SA, SQA and MFA in Fig. 6. For SQA we use the two schedules discussed above and additionally a third schedule, with $t_a^{\text{opt}}(N)$ sweeps but a higher temperature $\beta = 4$, which is the optimal temperature for MFA (see Supplemental Material [7]). All algorithms have heavy tails, $\xi > 0$, for problems with size $N \geq 72$, and ξ is increasing with system size. For SQA using $\beta = 10$ and $t_a = 10^4$ we find $\xi \geq 1$ already for $N \geq 72$, indicating that the mean time-to-solution is already divergent and dominated by the hardest instances. As we already saw above for $N = 200$, the tail behavior of SQA improves when using faster annealing schedules, and further small improvements are obtained by raising the temperature. While SQA with the best setting and MFA have the same ξ , SA has smaller ξ , indicating a faster decaying tail and better performance on hard instances.

Our results show that the tail behavior is the important measure of performance when considering the performance of an optimization algorithm since the hardness of an instance is a priori unknown. Therefore one has to choose a large enough run-time such that a high percentile of random instances are solved. Considering just the typical (median) performance of an optimization algorithm can be misleading in the presence of broad heavy-tailed distributions. The median effort misses the important fact that already at moderate problem sizes the mean time-to-solution can diverge, indicating that the hardest instances dominate the average. As one typically wants to solve much more than half of the problem instances, reducing the tails of the distribution is crucial to optimize the performance of an optimization algorithm.

One surprising result we obtained here is that for sim-

ulated quantum annealing, fast non-adiabatic schedules show far superior performance than slow adiabatic annealing schedules that would naïvely seem superior. This observation is consistent with results obtained in Refs. [19, 37] which showed that for some instances fast annealing schedules can help a (simulated) quantum annealer escape local minima. The superiority of fast annealing schedules can be understood by realizing that for hard instances, the annealing is stuck in local minima which are far from any ground state. Much slower annealing would increase the probability to find a ground state at the cost of substantially extended annealing times. In these hard instances it is, however, more efficient to introduce a perturbation (by running faster), which will sometimes kick the system out of these local minima in which they are trapped. This will increase the success probability for hard instances while it does not affect the easy instances by a lot and hence lead to a much more compact distribution of TTS.

As good agreement was found between SQA and the D-Wave devices, we expect that similarly faster annealing schedules can help the performance not only of simulated quantum annealers but also of physical quantum annealing devices.

The tail behavior is also important for extrapolating the performance of annealers to larger sizes. The tails are indicative of harder problems that will dominate for larger problem sizes. We thus expect scaling differences between different annealers to first show up in the tails.

We acknowledge discussions with Sergio Boixo, Paul Embrechts, Helmut G. Katzgraber, Hartmut Neven and Diethelm Würtz and thank Ethan Brown for proofreading and Coal Ila for inspiration. This work has been supported by the Swiss National Science Foundation through the National Competence Center in Research QSIT and is based upon work supported in part by IARPA via MIT Lincoln Laboratory Air Force Contract No. FA8721-05-C-0002.

-
- [1] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi, *Science* **220**, 671 (1983).
 - [2] B. Suman and P. Kumar, *J. Oper. Res. Soc.* **57**, 1143 (2006).
 - [3] T. Kadowaki and H. Nishimori, *Physical Review E* **58**, 5355 (1998).
 - [4] G. E. Santoro, R. Martonak, E. Tosatti, and R. Car, *Science* **295**, 2427 (2002).
 - [5] G. E. Santoro and E. Tosatti, *J. Phys. A: Math. Gen.* **39**, R392 (2006).
 - [6] T. F. Rønnow, Z. Wang, J. Job, S. Boixo, S. V. Isakov, D. Wecker, J. M. Martinis, D. A. Lidar, and M. Troyer, *Science* **345**, 420 (2014).
 - [7] See Supplemental Material [URL], which includes Refs. [38–47].
 - [8] F. Barahona, *J. Phys. A: Math. Gen.* **15**, 3241 (1982).
 - [9] M. W. Johnson, M. H. S. Amin, S. Gildert, T. Lanting, F. Hamze, N. Dickson, R. Harris, A. J. Berkley, J. Johansson, P. Bunyk, E. M. Chapple, C. Enderud, J. P. Hilton, K. Karimi, E. Ladizinsky, N. Ladizinsky, T. Oh, I. Perminov, C. Rich, M. C. Thom, E. Tolkacheva, C. J. S. Truncik, S. Uchaikin, J. Wang, B. Wilson, and G. Rose, *Nature* **473**, 194 (2011).
 - [10] S. Boixo, T. F. Rønnow, S. V. Isakov, Z. Wang, D. Wecker, D. A. Lidar, J. M. Martinis, and M. Troyer, *Nat. Phys.* **10**, 218 (2014).
 - [11] S. W. Shin, G. Smith, J. A. Smolin, and U. Vazirani, (2014), [arXiv:1401.7087v2 \[quant-ph\]](#).
 - [12] C. C. McGeoch and C. Wang, in *Proceedings of the 2013 ACM Conference on Computing Frontiers* (2013).
 - [13] S. Dash, (2013), [arXiv:1306.1202v2 \[math\]](#).
 - [14] H. G. Katzgraber, F. Hamze, and R. S. Andrist, *Physical Review X* **4**, 021008 (2014).
 - [15] P. Ray, B. K. Chakrabarti, and A. Chakrabarti, *Physical Review B* **39**, 11828 (1989).
 - [16] A. B. Fennila, M. A. Gomez, C. Sebenik, C. Stenson, and J. D. Doll, *Chemical Physics Letters* **219**, 343 (1994).
 - [17] E. Farhi, J. Goldstone, S. Gutmann, J. Lapan, A. Lundgren, and D. Preda, *Science* **292**, 472 (2001).
 - [18] A. Das and B. K. Chakrabarti, *Rev. Mod. Phys.* **80**, 1061 (2008).
 - [19] B. Heim, T. F. Rønnow, S. V. Isakov, and M. Troyer, *Science* (2015), [arXiv:1411.5693](#).
 - [20] T. Albash, T. Rønnow, M. Troyer, and D. Lidar, (2014), [arXiv:1409.3827 \[quant-ph\]](#).
 - [21] S. Coles, *An Introduction to Statistical Modeling of Extreme Values*, Springer Series in Statistics (Springer, 2001).
 - [22] P. Embrechts, C. Klüppelberg, and T. Mikosch, *Modelling Extremal Events: for Insurance and Finance*, corrected fourth printing 2003 ed. (Springer, 1997).
 - [23] L. de Haan, *Statistica Neerlandica* **44**, 45 (1990).
 - [24] H. G. Katzgraber, M. Körner, F. Liers, M. Jünger, and A. K. Hartmann, *Phys. Rev. B* **72**, 094421 (2005).
 - [25] M. Körner, H. G. Katzgraber, and A. K. Hartmann, *J. Stat. Mech: Theory Exp.* **2006**, P04005 (2006).
 - [26] H. G. Katzgraber, D. Würtz, and G. Blatter, *Phys. Rev. B* **75**, 214511 (2007).
 - [27] H. G. Katzgraber, S. Trebst, D. A. Huse, and M. Troyer, *J. Stat. Mech: Theory Exp.* **2006**, P03018 (2006).
 - [28] P. Dayal, S. Trebst, S. Wessel, D. Würtz, M. Troyer, S. Sabhapandit, and S. N. Coppersmith, *Phys. Rev. Lett.* **92**, 097201 (2004).
 - [29] S. Alder, S. Trebst, A. K. Hartmann, and T. Matthias, *J. Stat. Mech.*, P07008 (2004).
 - [30] M. Castellana, *Phys. Rev. Lett.* **112**, 215701 (2014).
 - [31] A. A. Balkema and L. de Haan, *Ann. Probab.* **2**, 792 (1974).
 - [32] J. Pickands III, *Ann. Stat.* **3**, 119 (1975).
 - [33] R.-D. Reiss and M. Thomas, *Statistical Analysis of Extreme Values: with Applications to Insurance, Finance, Hydrology and Other Fields*, 3rd ed. (Birkhäuser, 2007).
 - [34] M. Falk, J. Hüsler, and R.-D. Reiss, *Laws of Small Numbers: Extremes and Rare Events*, third, revised and extended edition ed. (Birkhäuser, 2011).
 - [35] A. J. McNeil, R. Frey, and P. Embrechts, *Quantitative Risk Management: Concepts, Techniques, and Tools* (Princeton University Press, 2005).
 - [36] B. C. Arnold, N. Balakrishnan, and H. N. Nagaraja, *A First Course in Order Statistics* (Society for Industrial

- and Applied Mathematics, 2008).
- [37] E. Crosson, E. Farhi, C. Lin, H. H. Lin, and P. Shor, (2014), [arXiv:1401.7320](#).
 - [38] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, *J. Chem. Phys.* **21**, 1087 (1953).
 - [39] S. V. Isakov, I. N. Zintchenko, T. F. Rønnow, and M. Troyer, (2014), [arXiv:1401.1084v2 \[cond-mat.dis-nn\]](#).
 - [40] M. Suzuki, *Prog. Theor. Phys.* **56**, 1454 (1976).
 - [41] L. Pierre, T. Giamarchi, and H. Schulz, *Journal of Statistical Physics* **48**, 135 (1987).
 - [42] D. Dietrich, L. de Haan, and J. Hüsler, *Extremes* **5**, 71 (2002).
 - [43] J. Hüsler and D. Li, *Extremes* **9**, 69 (2006).
 - [44] H. Drees, *AStA Advances in Statistical Analysis* **96**, 225 (2012).
 - [45] E. Castillo, A. S. Hadi, N. Balakrishnan, and J. M. Sarabia, *Extreme Value and Related Models with Applications in Engineering and Science* (Wiley-Interscience, 2004).
 - [46] Original S functions written by Janet E. Heffernan with R port and R documentation provided by Alec G. Stephenson., *ismev: An Introduction to Statistical Modeling of Extreme Values*. (2012), r package version 1.39.
 - [47] B. Pfaff and A. McNeil, *evir: Extreme Values in R*. (2012), r package version 1.7-3.