Difficulty Analysis of Posiform-Planted QUBO Benchmarks via Energy and Hamming Metrics

Yideun Kim
Department of Computer Engineering
Kangwon National University
Chuncheon, Republic of Korea
Email: rladlems1031@kangwon.ac.kr

Dongjae Lee†
Department of Convergence Security
Kangwon National University
Chuncheon, Republic of Korea
Email: dongjae.lee@kangwon.ac.kr

***Abstract*—Benchmarking quantum annealers and classical heuristics needs problem instances whose optimal solution is known in advance but that stay hard to solve. Posiform planting generates Quadratic Unconstrained Binary Optimization (QUBO) instances with one unique planted optimum. A hardened variant raises their difficulty. It adds random spin-glass terms and scales the planted component by a coefficient α. A practical obstacle is measurement. Hardness is usually graded by the ground-state success rate. But across the whole hard regime this rate is 0%, so it cannot distinguish instances, solvers, or values of α. This poster presents a measurement protocol for the zero-success regime. Instead of asking whether a run succeeded, we ask how close the failing runs come to the planted optimum. We use two quantities, computed over failures only and on a fixed set of the hardest instances. The first is the energy gap to the optimum. The second is the Hamming distance from the planted answer. Together they turn the flat 0% rate into a continuous, rankable difficulty curve. On a D-Wave Pegasus graph ($n = 5612$ qubits), the curve is non-monotone in α. It rises to a peak and then declines. A single supporting run on a D-Wave Advantage annealer reproduces the same shape, more weakly. This is consistent with the structure reflecting the instances rather than simulated annealing alone. The protocol gives a finer way to grade planted QUBO benchmarks.**

***Index Terms*—QUBO benchmarking, quantum annealing, simulated annealing, posiform planting, performance evaluation, NISQ, quantum optimization.**

Poster Relevance to IEEE Quantum Week (QCE 2026)

This work addresses the benchmarking, performance evaluation, and verification of programmable quantum annealers. These are core themes of IEEE Quantum Week. Its subject is the QUBO problem, the native input format of D-Wave quantum annealers and a central model for quantum and quantum-inspired optimization. The proposed failure-conditioned metrics target the NISQ-era regime, where solvers seldom reach the true ground state and a binary success rate is uninformative. They still rank instances, solvers, and a hardware-relevant difficulty parameter. The study spans both classical (simulated annealing) and quantum (D-Wave Advantage) solvers under a common measure, which supports cross-platform comparison. It maps directly to two QCE topics: *adiabatic quantum computation & quantum annealing* (Quantum Computing), the model that D-Wave devices implement; and *benchmarks* (Quantum Systems Software), since the failure-conditioned protocol and open instance set form a benchmark for the zero-success regime. All code and benchmark instances are released openly (`github.com/YIDEUNKIM/qubo_dataset`) to support reproducibility.

†Corresponding author and representative contact: Dongjae Lee (`dongjae.lee@kangwon.ac.kr`).

Difficulty Analysis of Posiform-Planted QUBO Benchmarks via Energy and Hamming Metrics

Yideun Kim
Department of Computer Engineering
Kangwon National University
Chuncheon, Republic of Korea
Email: rladlems1031@kangwon.ac.kr

Dongjae Lee†
Department of Convergence Security
Kangwon National University
Chuncheon, Republic of Korea
Email: dongjae.lee@kangwon.ac.kr

***Abstract*—Benchmarking QUBO solvers requires instances with a known optimum that stay hard. Posiform planting builds such instances with a unique optimum. A recent variant adds random spin-glass terms and scales the planted term by a coefficient α to tune hardness. Prior work characterizes this hardness only empirically, via the success rate, which shows that difficulty varies with the parameters but not why, and cannot compare instances at 0% success. We instead measure how close failing runs come to the optimum. Over failures only, on a fixed hardest-instance set, we record the energy gap and Hamming distance (HD) to the planted answer. These resolve the flat 0% into a continuous difficulty curve with a clear structure: $\Delta E = \Delta R + \alpha \Delta P$, a local-minimum floor plus a growing planted penalty. Under simulated annealing (SA) on a D-Wave Pegasus graph (n=5612) it is non-monotone in α: ΔE peaks then declines while HD falls throughout, with the peak where the rising penalty is overtaken by failures drifting toward the optimum. A D-Wave Advantage run reproduces the shape, suggesting an instance property, not SA alone. Reporting both quantities grades planted QUBO benchmarks even when no run succeeds.**

***Index Terms*—QUBO benchmarking, posiform planting, quantum annealing, simulated annealing, Hamming distance, energy gap, α-difficulty.**

I. Introduction

Benchmarking quantum annealers and classical heuristics on QUBO problems needs instances whose optima are known but hard to find. Posiform planting [1] produces QUBOs with a unique ground state, but they are easy for simulated annealing (SA) [2]. Pelofske et al. [3] made them harder. They add random discrete-coefficient spin-glass QUBOs R_i to a posiform-planted QUBO P with unique minimum $x^\star$, scaled by a coefficient α:

$$Q_\alpha(x) \;=\; \sum_i R_i(x) \;+\; \alpha\, P(x). \tag{1}$$

That study tuned hardness mainly through the random-QUBO size and the annealing time. It fixed $\alpha \in \{0.01, 0.1\}$ and reported the ground-state success rate. But this rate is 0% across the whole small-α regime. It cannot rank instances, solvers, or α values. Pelofske et al. conjecture that the hardness comes from local minima whose energy is very close to the true ground state.

We make that conjecture quantitative. ΔE measures local-minimum depth. But over all samples it mixes depth with the success rate. We therefore condition on failure and use the 100 hardest instances, held fixed as α varies. Measured this way, failure-conditioned ΔE and Hamming distance (HD) reveal a $\Delta R + \alpha \Delta P$ structure with a non-monotone peak. Neither quantity is new. Our contribution is the protocol. Failure-conditioning on a fixed hardest set turns the uninformative 0% rate into a rankable curve. A supporting quantum-annealer run is consistent with this structure coming from the instances, not SA alone.

II. Methodology

Instances. We generate 500 posiform-planted instances on the D-Wave Pegasus P16 graph [4] (n=5612 qubits). Each coefficient R_{ij} is -1 or $+1$. We split the graph into blocks of at most 16 qubits by recursive bisection and solve each block R_i exactly. The planted target $x^\star$ is the concatenation of the block ground states, so it is the global minimum of $\sum_i R_i$. P is the posiform of a planted 2-SAT whose only solution is $x^\star$. We rank the 500 instances by their total SA failing-reads across the sweep and fix the hardest 100 as a constant set. This set is chosen from SA alone, before any QPU run, so the cross-solver agreement is not circular.

Metrics. For each sample x we record $\mathrm{HD}(x, x^\star)$ [5] and the energy gap $\Delta E = E(x) - E(x^\star) \geq 0$ under the same Q_α. Because every R_{ij} is ± 1, each R energy is an integer. We therefore report ΔE in units of the smallest non-degenerate R gap (Δ_R=1). This unit is the same across instances, as is HD/n. A read counts as a success only when it equals $x^\star$ (a bit match), so the 0% success rate and the per-read failure rate both track $x^\star$ exactly. This differs from matching the ground-state energy, which at α=0 is degenerate (Sec. III-A). At large α, correct samples ($\Delta E \approx 0$, HD=0) drive both quantities to 0. We therefore condition all statistics below on failure ($x \neq x^\star$). Means are over failing reads. The instance-level standard error stays below $0.06\,\Delta_R$, so we omit error bars.

This work was supported in part by the Korea Internet & Security Agency (KISA) grant funded by the Korean government (Ministry of Science and ICT, MSIT) through the "Information Security Workforce Development (Information Security Specialized University Program)" in 2026, in part by the Basic Science Research Program through the National Research Foundation of Korea (NRF) funded by the Ministry of Education (No. 25411243), and in part by the "Creation of the Quantum Information Science R&D Ecosystem (Based on Human Resources)" through the NRF funded by the Korean government (RS-2023-NR057243).

†Corresponding author: Dongjae Lee.

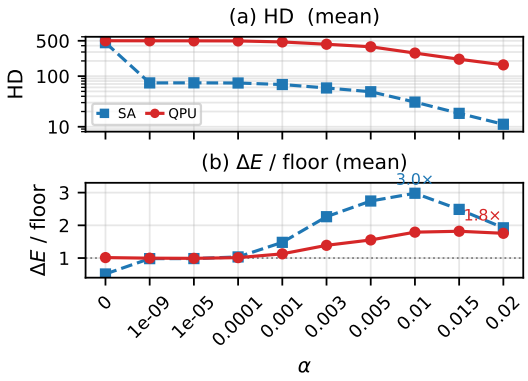


Fig. 1. Failure-conditioned difficulty over the 100 hardest instances. At each α, at least 91% of reads fail. Blue: SA. Red: a D-Wave Advantage annealer. (a) HD on a log axis falls monotonically but never reaches 0. (b) ΔE rescaled by each solver's small-α floor (SA $\approx 1.1\,\Delta_R$, QPU ≈ 24), on one axis. Both start at 1. SA rises to a peak of $3.0\times$ near $\alpha{=}10^{-2}$, then declines. The annealer shows the same shape but more weakly ($1.8\times$).

Protocol. We sweep α over 0, 10^{-9}, 10^{-5}, 10^{-4}, 10^{-3}, $3{\times}10^{-3}$, $5{\times}10^{-3}$, 10^{-2}, $1.5{\times}10^{-2}$, and $2{\times}10^{-2}$. Beyond $2{\times}10^{-2}$ the instances are increasingly solved, so no fixed failing set remains. The main sweep uses the `neal` SA package [6] at 1000 sweeps and 100 reads per (instance, α); we also run a separate 150-instance subset (not the hardest-100) at 1000 and 5000 sweeps to compare budgets. For a cross-solver check we replay the same 100 instances on a D-Wave Advantage annealer (`Advantage_system6.4`). The run is hardware-native, with auto-scaling, two-gauge averaging, and 100 reads at $20\,\mu$s. We recompute ΔE from Q_α.

III. Results

A. Floor, peak, and decline of ΔE

At $\alpha{=}0$ the random part R is degenerate. Here 56.1% of samples reach an R ground state ($\Delta E{=}0$), yet 0% equal $x^\star$, and HD is maximal (459, or 8.2% of n). So ΔE alone would read as solved, but HD shows no sample is near $x^\star$. For $\alpha{>}0$ (Fig. 1b), in the small-α region ($\alpha \in [10^{-9}, 10^{-4}]$), $\alpha\Delta P$ is negligible and ΔE stays flat at the depth of a local-minimum trap in R. We call this flat value the floor (mean 1.10, median 1). This floor of $\approx 1\,\Delta_R$ is the quantitative content of the local-minimum conjecture in [3]: in the small-α limit, failing samples sit about one R-gap above $x^\star$. HD is also flat (≈ 73). Past $\alpha{=}10^{-3}$ the planted term grows. ΔE rises to a peak of 3.35 at $\alpha{=}10^{-2}$, which is $3.0\times$ the floor, then declines to 2.16. The floor, peak, and decline are separated by far more than the instance-level SE ($< 0.06\,\Delta_R$). HD meanwhile falls monotonically ($73 \to 11$). So failing samples move closer to $x^\star$ in bits even as their energy gap grows. The peak comes from a competition. For a fixed failing bitstring, $\Delta E{=}\Delta R{+}\alpha\Delta P$ is linear in α ($\Delta R, \Delta P$ fixed per bitstring). But the failing set is not fixed: as α rises, failing samples drift toward $x^\star$ (HD falls), where ΔP is smaller. The observed mean is an envelope over this moving population: it climbs while $\alpha\Delta P$ dominates, then falls as the drift into smaller-ΔP regions wins. The peak is not a survivorship or selection artifact. The 100 instances fail at least 91% of reads at every α (0% ground-state success, as [3] report). The same shape appears over all 500 instances (floor ≈ 1.1, peak ≈ 3.0), and the hardest set only sharpens it.

B. Quantum-annealer consistency check

As supporting evidence, we ran the same 100 instances once on a D-Wave Advantage annealer (Fig. 1, red). It shows the same pattern. There is 0% success, a non-monotone ΔE peak near $\alpha{=}10^{-2}$, and a monotonic HD decrease ($501 \to 167$). Its baselines sit an order of magnitude above SA's ($\Delta E \approx 24$, HD≈ 500), and its peak is weaker in relative terms ($1.8\times$ its floor versus SA's $3.0\times$). One run cannot establish generality. But it is consistent with the structure being a property of the instances, not of SA. The integrated control error (ICE) is the annealer's control-precision limit: couplers carry a roughly fixed error of order 0.02–0.03. At the peak the planted couplers map to Ising $J \approx 0.03$, comparable to ICE, so the device only partly resolves the signal, consistent with the weaker peak above. A smaller scale would push them below ICE and should flatten it, which we leave to future work.

IV. Discussion and Conclusion

Failure-conditioned ΔE measures difficulty (local-minimum depth) and HD measures proximity to $x^\star$. Reporting both separates how deep a failure is from how close it is. The all-sample mean (the success rate) hides the peak. It falls at larger α only because more reads succeed, not because failures are shallower. ΔE also separates solvers that the success rate cannot: on the 150-instance subset at $\alpha{=}10^{-3}$, both budgets give 0% success, yet ΔE is 1.68 versus 1.00. We therefore recommend reporting failure-conditioned ΔE (floor and peak) with HD over a fixed hardest set whenever an α-like parameter is swept. This gives a failure-conditioned diagnostic of the benchmark instances rather than a single difficulty score. Absolute values still depend on the solver and schedule.

Limitations. The QPU evidence is a single run (two gauges, one $20\,\mu$s anneal). The classical side is a single SA family. The instances use one coefficient type (±1) and one topology (Pegasus).

Future work: broader classical solvers, more annealer gauges and anneal times, other coefficient types and topologies, and joint HD–ΔE analysis. Code and instances: https://github.com/YIDEUNKIM/qubo_dataset.

References

[1] G. Hahn, E. Pelofske, and H. N. Djidjev, "Posiform planting: generating QUBO instances for benchmarking," *Frontiers in Computer Science*, vol. 5, Art. no. 1275948, 2023.

[2] S. Kirkpatrick, C. D. Gelatt Jr., and M. P. Vecchi, "Optimization by simulated annealing," *Science*, vol. 220, no. 4598, pp. 671–680, 1983.

[3] E. Pelofske, G. Hahn, and H. N. Djidjev, "Increasing the hardness of posiform planting using random QUBOs for programmable quantum annealer benchmarking," *npj Unconventional Computing*, vol. 2, Art. no. 17, 2025.

[4] K. Boothby, P. Bunyk, J. Raymond, and A. Roy, "Next-generation topology of D-Wave quantum processors," arXiv:2003.00133, 2020.

[5] R. W. Hamming, "Error detecting and error correcting codes," *Bell Syst. Tech. J.*, vol. 29, no. 2, pp. 147–160, 1950.

[6] D-Wave Systems Inc., "dwave-neal: a simulated annealing sampler," GitHub, 2022. https://github.com/dwavesystems/dwave-neal.