Limitations in parallel Ising machine networks: Theory and practice

Matthew X. Burns[*] and Michael C. Huang[†]
Department of Electrical and Computer Engineering, University of Rochester, USA

(Received 25 March 2025; accepted 16 July 2025; published 19 August 2025)

Analog Ising machines (IMs) occupy an increasingly prominent area of computer architecture research, offering high-quality, low-latency, and low-energy solutions to intractable computing tasks; however, IMs have a fixed capacity, with little to no utility in out-of-capacity problems. Previous works have proposed parallel, multi-IM architectures to circumvent this limitation [A. Sharma, *et al.*, in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22 (Association for Computing Machinery, New York, NY, USA, 2022), p. 508; R. Santos, *et al.*, Enhancing quantum annealing via entanglement distribution, ArXiv:2212.02465]. In this work, we theoretically and numerically investigate trade-offs in parallel IM networks to guide researchers in this burgeoning field. We propose formal models of parallel IM execution models, and we then provide theoretical guarantees for probabilistic convergence. Numerical experiments illustrate our findings and provide empirical insights into the high- and low-synchronization-frequency regimes. We also provide practical heuristics for parameter and model selection, informed by our theoretical and numerical findings.

DOI: 10.1103/gpg9-3tjj

I. INTRODUCTION

With the slowing of Moore's law, dynamics-based analog computation offers an alternative path to increased performance and power efficiency. Recent dynamical accelerator proposals include adiabatic quantum computing (AQC) [1], quantum annealing (QA) [2], analog electronics [3,4], and photonic systems [5], each offering low-latency computation for specific tasks. Furthermore, classical analog optimization (electronics and photonics) is an order of magnitude more energy efficient than traditional systems, prompting further interest in lowering the energy footprint of high-performance computing. The most prominent class relies on continuous relaxations of the Ising spin-glass Hamiltonian, earning the appellation "Ising machines" (IMs) [6,7]. Simulated behavior and hardware prototypes have demonstrated that Ising machines can effectively accelerate a number of NP-hard combinatorial tasks, including graph maximum cut [3–5], low-density parity-check decoding [8], satisfiability testing [9,10], energy-based model inference [11], and data clustering [12].

[*]Contact author: mburns13@ur.rochester.edu
[†]Contact author: michael.huang@rochester.edu

Published by the American Physical Society under the terms of the Creative Commons Attribution 4.0 International license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

Nonetheless, current Ising-machine proposals have a number of limitations, chief among which is a fixed device capacity. Once designed, an IM has a fixed number of degrees of freedom. If that limit is exceeded, the IM alone cannot solve the problem directly. Broadly speaking, there are two approaches to the problem: (1) using the IM hardware as a subsolver in an otherwise conventional algorithmic framework; or (2) orchestrating multiple IM chips to form a logically larger IM.

In the subsolver approach, conceptually, the overall problem is still solved conventionally by a digital algorithm, except IM hardware is available to accelerate subproblem solving. Note that a naive divide-and-conquer technique will generally fail [13]. The current state of the art is the D-Wave Leap Hybrid framework, in which quantum annealing "guides [a] classical heuristic" [14], such as Tabu search or simulated annealing. While some works have shown time-to-solution advantages in crafted benchmark problems [14], hybrid solvers can still be outperformed by classical algorithms [15]. Works targeting problem-specific applications such as network community detection [16] and satisfiability [10] have achieved some limited speedup; however, their approach is both highly limited by Amdahl's law and is (by design) not general purpose.

Experimental results have demonstrated the limitations of the subsolver approach. Ref. [17] examined the scaling behavior of "divide-and-conquer" hybrid solvers. Figure 1 (from Ref. [17]) shows the estimated speedup over CPU-based simulated annealing using the (then state-of-the-art)

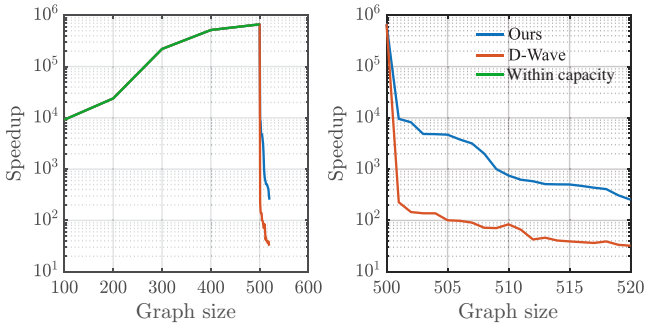


FIG. 1. From Ref. [17]. Estimated speedup degradation as solver capacity (500 spins) is exceeded. The right-hand plot shows an enlarged version of the left-hand plot. "Ours" refers to an annealing-based divide-and-conquer scheme tested in Ref. [17], while "D-Wave" refers to the QBSOLV [18] framework.

QBSOLV [18] framework ("D-Wave") and a simulated-annealing-based solver examined in Ref. [17] ("Ours"). The speedup is shown to decay precipitously as the problem size barely exceeds hardware capacity, motivating the authors to pursue hardware-only approaches. Figure 2 reinforces the conclusions of Ref. [17] using the more modern D-Wave Kerberos hybrid workflow [20]. Here, we (1) simulate a typical analog IM (BRIM [3]) and (2) run Kerberos using the D-Wave Leap Hybrid service, with problems selected from the 2000-node GSet suite graphs [19]. Both solver times are then normalized to CPU-based parallel tempering [21]. We again see orders-of-magnitude differences in performance, highlighting the significant challenge of the subsolver approach to IM scalability.

The approach using multiple IM chips attempts to preserve energy efficiency and solver speed at the cost of extra IM hardware. The problem is partitioned into disjoint subproblems, each solved on a different IM, communicating over a common network. So-called "multichip" approaches are a mainstay of traditional computer architecture, but they run into significant difficulties when applied to analog systems. As shown empirically by Sharma *et al.* in 2022 [17], communication bandwidth is the limiting factor in both speedup and solution quality. The authors then

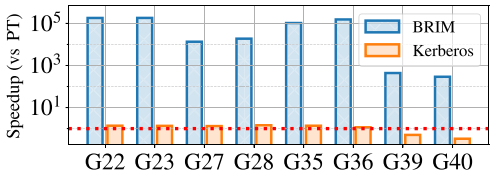


FIG. 2. Estimated IM speedup versus CPU-based parallel tempering (PT) running on eight cores on problems from the GSet suite [19]. Simulated BRIM [3] behavior represents a typical analog Ising machine, while D-Wave Kerberos [20] represents a typical hybrid platform. All solvers are run to a minimum target of 99% of the best-known solution.

explored two solution paths: *serial* and *concurrent* execution models. Serial execution adopts a pipelined execution model, which sacrifices latency for bandwidth savings, while concurrent execution provides lower-latency operation with higher communication costs. The trade-offs were empirically studied using Sherrington-Kirkpatrick spin glasses, but they were not discussed in generality. In particular, the behavior and convergence properties of concurrently executing subsystems are not well understood within the IM community.

In this work, we perform an in-depth theoretical and numerical study of parallel Ising-machine architectures, with a particular emphasis on lesser-studied concurrent execution models. We focus here on Ising machines whose behavior can be modeled by an overdamped Langevin equation. Our contributions can be summarized as follows:

(1) We present a convergence analysis for concurrent execution mode, providing lower bounds in Kullback-Leibler (KL) divergence (Proposition 1) and sufficient conditions for per-iteration contraction (Propositions 2 and 3).

(2) We provide numerical results explicitly demonstrating our theoretical findings on small, enumerable systems.

(3) We conduct energy- and time-efficiency comparisons between parallel execution models across a range of network bandwidths and energy costs in canonical MaxCut benchmarks.

(4) We propose practical, application-dependent heuristics for execution-model and parameter selection.

Section II provides an overview of relevant topics used in our analysis. Section III proposes a general framework for parallel Ising-machine analysis that is applicable to a wide swath of current proposals. Section IV provides theoretical convergence properties. Narrowing our focus to linear Ising machines [3], Sec. V discusses numerical experiments comparing the time and energy costs of each proposal in unconstrained binary optimization, and a decision tree for choosing between serial and parallel execution is then proposed. Section VI concludes by summarizing our findings.

II. PRELIMINARIES

A. Combinatorial optimization by spin-glass sampling

We begin by recalling the general Ising spin-glass Hamiltonian $\mathcal{H}(\boldsymbol{s})$:

$$\mathcal{H}(\boldsymbol{s}) = -\sum_{i\neq j} J_{ij}\, s_i s_j - \sum_{i=1}^{N} h_i s_i, \tag{1}$$

where $\boldsymbol{s} \in \{-1, 1\}^N$ is a vector of one-dimensional spins, $J \in \mathcal{S}(\mathbb{R})^{N\times N}$ is an $N \times N$ symmetric matrix describing spin-spin couplings, and $h \in \mathbb{R}^N$ is an external field term.

Spin-glass systems are typically studied within the canonical ensemble with equilibrium statistics described by the Gibbs-Boltzmann distribution $\pi(\boldsymbol{s})$:

$$\pi(\boldsymbol{s}) = \frac{1}{Z}\exp[-\beta\mathcal{H}(\boldsymbol{s})], \tag{2}$$

where $\beta = T^{-1}$ is the inverse temperature of the heat bath, with the Boltzmann constant set to unity for convenience. In the limit $\beta \to \infty$, $\pi(\boldsymbol{s})$ concentrates around the ground states of $\mathcal{H}(\boldsymbol{s})$.

Finding the ground state of a spin-glass system with nonplanar coupling graphs [22] is NP-hard [23,24], meaning that efficient spin-glass optimizers can be used to solve any problem in the NP complexity class. Moreover, several real-world NP-hard optimization problems have equivalent Ising formulations [25], including integer linear programming, production scheduling, and graph coloring (to name a few).

The classic unconstrained combinatorial optimization problem is the (weighted) maximum cut (MaxCut) problem. Given a graph $G = (V, E)$ with weighted adjacency matrix W, the MaxCut problem seeks to partition the graph into two sets, S and T, such that the weight of edges crossing the partition is maximized. MaxCut has a trivial representation as an Ising Hamiltonian with $J = -W$ with spin values denoting set membership. Given the energy $\mathcal{H}(\boldsymbol{s}) = -\frac{1}{2}\boldsymbol{s}^T J\boldsymbol{s}$, the cut value $C(S, T)$ can be computed as

$$C(S, T) = \frac{1}{2}\left(\sum_{(u,v)\in E} W_{uv} - \mathcal{H}(\boldsymbol{s})\right). \tag{3}$$

MaxCut is a widely used benchmark task for Ising machines; however, real-world problems typically have constraints defining feasible solutions. Constrained optimization is generally accomplished by introducing a penalty term $\mathcal{H}_{\text{Con}}(\boldsymbol{s})$ with multiplier γ:

$$\mathcal{H}(\boldsymbol{s}) = \lambda\mathcal{H}_{\text{Con}}(\boldsymbol{s}) + \mathcal{H}_{\text{Obj}}(\boldsymbol{s}), \tag{4}$$

where $\mathcal{H}_{\text{Obj}}$ represents the optimization objective, free of any constraints, while $\mathcal{H}_{\text{Con}}(\boldsymbol{s})$ penalizes infeasible states. A canonical constrained combinatorial problem is the traveling-salesman problem (TSP). Given N cities with an $N \times N$ distance matrix D, the TSP seeks a Hamiltonian path [26] between these cities that minimizes the total distance traveled. Here, λ must be chosen sufficiently large to ensure that it is generally unfavorable to violate problem constraints to improve the objective function. For the best performance, it is typically tuned per problem; however, some problem-agnostic heuristics have successfully recovered high-quality solutions [27].

B. Comparing probability measures

Instead of directly finding the ground state, we can alternatively attempt to define a random process whose sample distribution converges to π. Accordingly, we need to define notions of "distance" or "divergence" between two distributions. Note that throughout this work, we abuse notation by identifying a probability distribution with its density with respect to the Lebesgue measure.

Comparisons between two probability measures μ and ν often arise in the study of stochastic processes; hence, a number of functionals have been developed to quantify their proximity in distribution. In this work, we make use of two. The first is the p-Wasserstein distance W_p,

$$W_p(\mu, \nu) = \inf_{\gamma\in\mathcal{C}(\mu,\nu)} (\mathbb{E}_\gamma ||x - y||_p^p)^{1/p}, \tag{5}$$

where γ is a coupling [28] over the measures μ and ν and $||\cdot||_p$ is the p-norm. The Wasserstein distance measures the minimal amount of "work" needed to move probability mass from μ to ν, hence the term "optimal transport." Note that the Wasserstein distance defines a metric over probability distributions, meaning that the measure satisfies a triangle inequality (a property that we will exploit).

Wasserstein distances have seen extensive use in the analysis of sampling algorithms, particularly for Langevin Monte Carlo methods [29–31]. Most works use the Euclidean W_2, while some analyses focusing on nonconvex landscapes have favored W_1 due to the relative simplicity of its triangle inequality [31]. Recent works in stochastic thermodynamics have made extensive use of optimal transport to bound stochastic entropy production and define optimal control functions [32–36].

The second measure we will use is the Kullback-Leibler divergence

$$\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) = \mathbb{E}_\mu\left[\log\frac{d\mu}{d\nu}\right], \tag{6}$$

where $d\mu/d\nu$ is the Radon-Nikodym derivative of μ with respect to ν. Unlike the Wasserstein distance, $\mathrm{D}_{\mathrm{KL}}(\mu\|\nu)$ is not a distance, as it is not symmetric and does not satisfy the triangle inequality. Nevertheless, it provides a useful measure of convergence between two distributions, as $\mathrm{D}_{\mathrm{KL}}(\mu\|\nu) \geq 0$ with equality iff $\mu = \nu$.

In a discrete state space Ω, the expectation takes the form of a sum of the possible system states

$$W_p(\mu, \nu) = \inf_{\gamma\in\mathcal{C}(\mu,\nu)}\left(\sum_{x,y\in\Omega}\gamma(x, y)||x - y||_p^p\right)^{1/p}, \tag{7}$$

$$\mathrm{D_{KL}}(\mu\|\nu) = \sum_{x\in\Omega} \mu(x) \log \frac{d\mu}{d\nu}(x) \tag{8}$$

and integrals in a continuous state space $\mathcal{D} \subseteq \mathbb{R}^d$

$$W_p(\mu,\nu) = \inf_{\gamma\in\mathcal{C}(\mu,\nu)} \left(\int_{x,y\in\mathcal{D}} ||x-y||_p^p \gamma(dx,dy) \right)^{1/p}, \tag{9}$$

$$\mathrm{D_{KL}}(\mu\|\nu) = \int_{\mathcal{D}} \mu(x) \log \frac{d\mu}{d\nu}(x) dx. \tag{10}$$

III. PROBLEM PARTITIONING

When the number of spins N exceeds the IM capacity K, we need to decompose the problem into interacting subsystems, each of which is evolved separately. Without loss of generality, we assume that N is an integer multiple of K so that the number of chips needed, $B = N/K$, is an integer; otherwise, we can simply pad with noninteracting spins.

We denote the global problem state $X \in \mathbb{R}^N$, the local subsystems as $X_p \in \mathbb{R}^{N_p}$, and the individual variables as $x_i \in \mathbb{R}$. Hence, $X_p \subseteq X$ and $x_i \in X_p$. By $\nabla_{X_p} f(X)$, we mean the gradient of function f with respect to system X_p, which may have dependencies on other subsystems. By $X_{\setminus p}$, we mean all degrees of freedom *not* in subsystem X_p, i.e., $X_{\setminus p} = (X_1,\ldots,X_{p-1},X_{p+1},\ldots,X_B)$.

Two primary approaches have been proposed for solving out-of-capacity problems: *serial* and *concurrent*, each of which is illustrated in Fig. 3.

Serial operation can be likened to block Gibbs sampling, where the IM separately evolves each subsystem conditioned on the rest of the problem state:

$$\text{Serial} \begin{cases} X_1^1 \sim \pi(X_1|X_2^0,\ldots,X_B^0), \\ X_2^1 \sim \pi(X_2|X_1^1,X_3^0,\ldots,X_B^0), \\ \quad\vdots \\ X_B^1 \sim \pi(X_B|X_1^1,X_2^1,\ldots,X_{B-1}^1); \end{cases} \tag{11}$$

however, actual IM systems will have some history dependence or bias in their sample distributions, and (11) is therefore simply an idealization.

Serial methods have also been referred to as "alternating minimization" in optimization literature [37–39], "large-neighborhood local search" in IM literature [14,18], and "generalized heat-bath" or "partial-resampling" methods in computational physics [40]. Prominent examples include hybrid D-Wave Leap algorithms, which alternate subproblem sampling with classical postprocessing. If each subsolver converges to its conditional distribution, serial operation guarantees asymptotic convergence to the stationary distribution [41]. Serial algorithms have also achieved high solution quality in a number of combinatorial applications, particularly when combined with classical postprocessing; however, serial algorithms necessarily increase latency by sacrificing the parallelism afforded by dynamical system optimization.

In contrast, *concurrent operation* evolves all subsystems simultaneously, conditioned on the previous problem state:

$$\text{In parallel} \begin{cases} X_1^1 \sim \pi(X_1|X_2^0,\ldots,X_B^0), \\ X_2^1 \sim \pi(X_2|X_1^0,X_3^0,\ldots,X_B^0), \\ \quad\vdots \\ X_B^1 \sim \pi(X_B|X_1^0,X_2^0,\ldots,X_{B-1}^0). \end{cases} \tag{12}$$

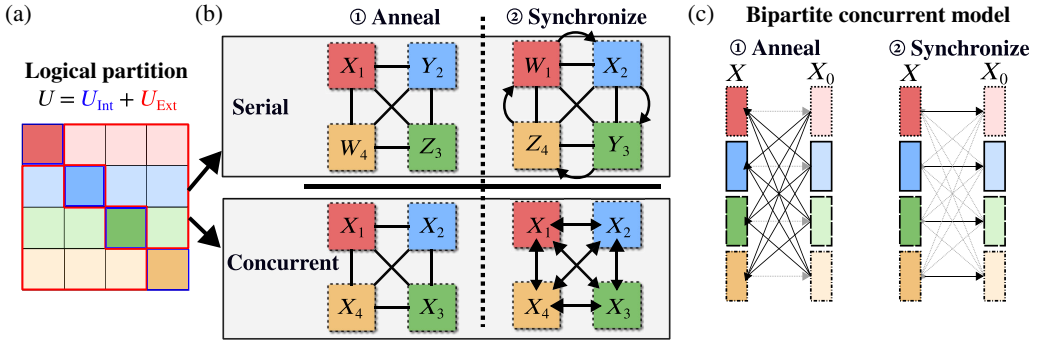


FIG. 3. Illustration of the two primary parallel IM execution models. (a) A graphical illustration of a logical partition of the target potential U into U_{Int} and U_{Ext}. Here, $U(X)$ is depicted as a quadratic function $U(X) = \frac{1}{2}X^T JX$ for some symmetric matrix J. The internal interactions comprising U_{Int} have a block-diagonal structure (outlined in blue) while the external interactions lie on the off-diagonal elements (highlighted in red). (b) The operation of the two execution models: serial and concurrent. During the annealing phase, serial execution simultaneously optimizes multiple independent replicas X, Y, Z, and V, while concurrent operation optimizes a single replica X. During a synchronization step, serial execution transfers replicas between chips to conditionally optimize the next subspace in a block Gibbs fashion and concurrent operation globally synchronizes the local spin representations of all subsystems. (c) A bipartite model representing the concurrent model. During an annealing stage, the X subsystems are independent given X_0. During synchronization, X_0 is sampled from X, and the process begins again in the next epoch.

Parallel updates have long been used in computational statistical physics to accelerate simulations of classical and quantum systems. Typically, highly parallel spin updates are limited to lattice-based systems [42,43]. Regular lattices enable researchers to apply specific domain decompositions that yield a factorized system [44]. The factorized subsystems can then be sampled in parallel without any sampling bias. If the domain decomposition does not produce a factorized distribution, then the resulting sample distribution will differ (perhaps significantly) from the target π.

Spin glasses derived from combinatorial optimization problems do not have known factorizing decompositions; however, the pursuit of high-performance optimization has driven further proposals for parallel updates in spite of their inherent inaccuracies. Examples include proposed analog IM architectures [17] and more recently in an AQC system [45]. Notice that such parallel physical systems differ from other parallel *computational* IMs [46,47] in subtle but important ways.

A computational IM can have multiple subunits and thus execute a parallel algorithm. Yet that parallel algorithm can still follow the serial dynamics of an ideal, monolithic dynamical system [46]. Physical IMs, on the other hand, are different; their dynamics are dictated by how the components are connected. While spins on the same chip can impact each other practically instantaneously, spins from different chips cannot: the continuous nature of both the time and state of spins, along with the limited communication bandwidth (relative to the ultrafast dynamics of physical IMs), dictate that the impacts of remote spins are necessarily delayed and approximate. This approximation changes the actual Hamiltonians of the parallel machines. Slowing down the dynamics can obviously reduce approximation error, but it is also clearly suboptimal. In this paper, we focus on the impact of this approximation error in classical IMs. Similar effects in quantum systems have been studied using Trotterization and Baker-Campbell-Hausdorff truncation [45].

In a realistic parallel IM implementation with limited communication bandwidth, partitions can only exchange spin states at a frequency allowed by the communication fabric. Thus, the state of a spin from an external partition is only updated periodically and remains frozen until the next update. The partitioned spin-glass problem can be described using a bipartite system model, as illustrated in Fig. 3(c). We distinguish between active degrees of freedom ($X(t)$) and latent spin copies $X_0 \triangleq X(\lfloor t/\tau \rfloor \tau)$ encoding the system state at the last synchronization epoch. In practice, the latent states may either be explicitly represented as hardware copies [17] or summarized into a linear bias term [18,48].

In either case, the partitioned solver acts as a block Gibbs sampler, alternating between sampling $X \sim \pi(Y|X_0)$ and $X_0 \sim \pi(Y|X)$. The X sampling phase occurs in continuous time on each dynamical system, while sampling X_0 occurs during a discrete synchronization step. When viewed as a bipartite system, "concurrent mode" is an instance of partial resampling [40] with exact conditional distributions.

Moreover, the conditional distribution $\pi(X|X_0)$ factorizes over partitions into $\pi(X|X_0) = \prod_{p=1}^{B} \pi_b(X_p|X_0)$, with $\pi_b(X_b|X_0)$ being the stationary distribution of subsystem p. Therefore the *conditional* can be exactly sampled using parallel subsystems; however, the primary problem is immediately apparent. Unless the target distribution π permits the same factorization, concurrent-mode dynamics will not have π as the stationary target. A key aim of this work is to determine how closely the factorized model *approximates* π.

A. System dynamics

We consider a system with a Lyapunov function

$$H(X) = -\frac{\xi}{2}U(x) + \sum_{i=1}^{d} f_i(x_i), \tag{13}$$

where ξ is a system-dependent constant, $U(x)$ is the system potential implementing pairwise interactions, and $\{f_i(X) : \mathbb{R}^N \to \mathbb{R}\}$ are a set of functions corresponding to self-interaction terms (e.g., subharmonic injection locking) and external fields. Notably, we assume that U is a *separable* potential.

Definition 1 (Separable potential). A function $U : \mathbb{R}^N \to \mathbb{R}$ is a *separable* potential if it can be expressed as

$$U(x) = \sum_{i \neq j} u_{ij}(x_i, x_j), \tag{14}$$

where $\{u_{ij}\}$ is a set of continuously differentiable functions.

The IM models considered here include a Brownian noise term dW_t in their system dynamics to model analog noise and/or implement thermal annealing [4,9]. We abuse notation for the sake of simplicity by denoting each vector or scalar Brownian term as dW_t, with the dimension clear from context.

Example 1 (Linear systems). The analog electronic architecture proposed in Refs. [3,9,49] uses capacitor voltage as a continuous approximation of spin. The system can be modeled using the stochastic differential equation (SDE)

$$dV = \frac{1}{RC} JV dt + \sqrt{2\beta(t)^{-1}} dW_t, \tag{15}$$

where $V \in \mathbb{R}^N$ describes the voltages across N resistively coupled capacitors with system time constant RC, J is

a symmetric coupling matrix, and $\sqrt{2\beta(t)^{-1}}$ is a time-dependent drift term. In practice, there are added potential terms confining V to [GND, VDD]; however, Eq. (15) is a useful approximation within the normal operating region. The Hamiltonian is correspondingly

$$H(V) = -\frac{1}{2RC}V^T JV, \tag{16}$$

which is linear and therefore separable. For numerical simulations, we set $C = 50$ fF and $R = 310$ kΩ and integrate (15) using an Euler-Maruyama discretization and a 10^{-12} time step.

Example 2 (Kuramoto model). A network of N Kuramoto oscillators with phases $\{\theta_i\}_{i=1}^N$ with a subharmonic injection locking term obey a system of coupled SDEs [4]

$$\begin{aligned} d\theta_i = &-K_J(t)\sum_{j\neq i} J_{ij}\sin\big(\theta_i-\theta_j\big)dt \\ &+ K_S(t)\sin(2\theta_i)dt + \sqrt{2\beta(t)^{-1}}dW_t, \end{aligned} \tag{17}$$

where $K_J(t)$ and $K_S(t)$ are time-dependent coefficients defining the energy landscape and J_{ij} are the elements of a symmetric coupling matrix. The corresponding Hamiltonian is

$$H(\theta) = -K_J(t)\sum_{j\neq i} J_{ij}\cos\big(\theta_i-\theta_j\big) + \frac{K_S(t)}{2}\sin(2\theta_i)dt, \tag{18}$$

which is obviously separable.

Note that this framework does not handle nonlinear activation functions as proposed in Ref. [50], indicating a direction for future work.

By restricting ourselves to separable systems, we can perform block decompositions $X = (X_1, \ldots, X_B)$ with separable gradients

$$\frac{\partial U}{\partial x_i} = \sum_{p=1}^{B}\sum_{x_j\in X_p}\frac{\partial u_{ij}}{\partial x_i}, \tag{19}$$

allowing us to express the system SDE

$$\begin{aligned} dX_p = &\Bigg(\overbrace{\nabla_{X_p}\Bigg[\sum_{x_i\in X_p} f_i(x_i) + \sum_{\substack{i\neq j\\ x_i,x_j\in X_p}} u_{ij}(x_i,x_j)\Bigg]}^{\nabla_{X_p}U_{\text{Int}}(X(t))} \\ &+ \overbrace{\nabla_{X_p}\Bigg[\sum_{\substack{x_i\in X_p\\ x_j\notin X_p}} u_{ij}(x_i,x_j)\Bigg]}^{\nabla_{X_p}U_{\text{Ext}}(X(t))}\Bigg)dt + \sqrt{2\beta^{-1}(t)}dW_t. \end{aligned} \tag{20}$$

We define $\nabla_{X_p}U_{\text{Int}}(X(t))$ as "Internal" interactions occurring within the variable block and $\nabla_{X_p}U_{\text{Ext}}(X(t))$ as "External" interactions occurring between distinct blocks. Viewing the system as a network of interacting nodes, the external interactions are the set of "cut" edges formed by the partition.

Suppose that we condition the dynamics of block X_p on all external blocks by fixing their states at time $s \leq t$. Then the current

$$\nabla_{X_p}U_{\text{Ext}}(X_p(t), X_{\setminus p}(s)) \tag{21}$$

can be regarded as either a fixed external field (for the linear system) or a self-interaction term (in the case of the Kuramoto model), conditioned on the states of other blocks.

The system dynamics for each block are then given by

$$\begin{aligned} dX_p = &\big[\nabla_{X_p}U_{\text{Int}}(X_p(t)) + \nabla_{X_p}U_{\text{Ext}}(X_p(t), X_{\setminus p}(s))\big]dt \\ &+ \sqrt{2\beta^{-1}}dW_t. \end{aligned} \tag{22}$$

We define the *synchronization epoch* as $\tau = \max(t-s)$. Clearly, Eq. (22) recovers Eq. (20) in the $\tau \to 0$ (infinite bandwidth) limit; however, the finite bandwidth fundamentally results in some approximation. For a given realization of $X_0 \triangleq X(s)$, the dynamics described by Eq. (22) will in general have different fixed points than (20). Taking the example of a deterministic linear model, we have

$$dX = J_{\text{Int}}X(t)dt + J_{\text{Ext}}X_0 dt. \tag{23}$$

For sufficiently small synchronization epochs τ, (23) is a good approximation to the full system dynamics; however, as τ increases, we have increasing divergence from the ideal system dynamics. The second term provides a linear bias aligning $X(t)$ parallel to $J_{\text{Ext}}X_0$. When the spectral radius of J_{Ext} is significantly greater than J_{Int}, the external linear term can become large enough to dominate the

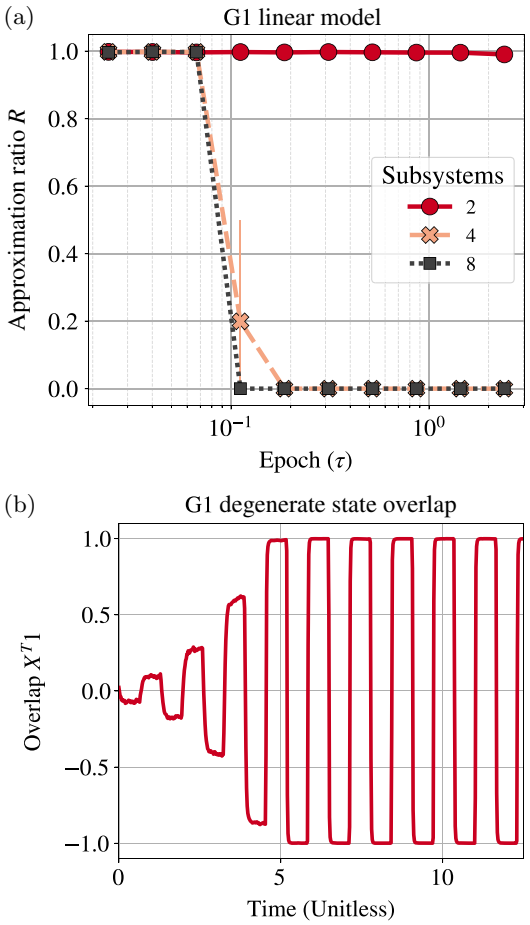


FIG. 4. Simulated linear model behavior on graph G1 versus unitless time. (a) Concurrent mode cut value versus synchronization epoch τ for the unweighted graph G1 using two, four, and eight subsystems. A cut value of 0 indicates that the system diverged to the degenerate $\pm 1^N$ state. (b) State overlap $\langle X, 1^N \rangle$ versus time for four-subsystem concurrent execution with a $\tau = 0.6$ synchronization epoch.

time evolution of X_t. Furthermore, when the maximum-magnitude eigenvalue of J_{Ext} is negative, sufficiently large values of τ result in oscillatory behavior between epochs, as we now demonstrate.

Figure 4 shows the behavior of a linear process with increasing synchronization period τ on the GSet [19] benchmark graph G1. Figure 4(a) shows the approximation ratio R for several block counts B with increasing synchronization epoch. While the system performs well with large τ when $B = 2$, increasing the number of blocks results in a sharp decline in solution quality. The graph has purely antiferromagnetic -1 couplings, resulting in a dominant eigenvector with strictly positive elements and a negative eigenvalue for both J_{Int} and J_{Ext}.

For $B = 2$, the spectral radii $\rho(J_{\text{Int}})$ and $\rho(J_{\text{Ext}})$ are comparable (25.1 and 24.9, respectively). Increasing to $B = 4$ results in J_{Int} being dominated by J_{Ext} ($\rho(J_{\text{Int}}) = 13.1, \rho(J_{\text{Ext}}) = 37.0$), with $B = 8$ further exacerbating the difference ($\rho(J_{\text{Int}}) = 7.5, \rho(J_{\text{Ext}}) = 43.0$). The minimum energy state of the approximate Lyapunov function within the hypercube $[-1, 1]^N$ then becomes the degenerate $\{-1\}$ and $\{+1\}$ states, depending on X_0. For a random initialization, we typically have $X_0^T \mathbf{1} > 0$. The dominant eigenvector of J_{Ext} then tends to amplify the magnitude of $X(t)^T \mathbf{1}$ over time, as shown in Fig. 4(b).

As a result, the system oscillates between the two poles on each synchronization epoch. These oscillations are purely an artifact of using a bounded $X \in [-1, 1]^N$ domain, the true problem is that *the factorized stationary distribution no longer faithfully approximates the highly coupled target distribution*.

We find that this behavior is limited to fully antiferromagnetic models with dominant spectral modes, and we do not observe it in benchmarks with mixed ferromagnetic and antiferromagnetic couplings or in problems with nonunit weights. Nevertheless, parallel execution models still introduce limits on convergence, as shown theoretically and numerically in Secs. IV and V.

IV. CONVERGENCE PROPERTIES

In this section, we discuss the weak convergence properties of the serial and concurrent execution models. We begin by briefly reviewing existing results for serial execution models, which are better studied in the literature. We then present results concerning the weak convergence of concurrent execution models. We first demonstrate finite lower bounds for asymptotic KL divergence between the sample and target distributions, and we then provide sufficient conditions for contraction in 1-Wasserstein distance in each epoch.

The latter results are not *directly* usable, as Wasserstein distances are currently intractable to estimate for general probability distributions [51]; however, they will provide insights into the behavior of the concurrent processes, from which we will formulate empirically usable heuristics in later sections. Moreover, our results may be of independent interest in stochastic thermodynamics, where Wasserstein distances have recently been applied [34–36].

A. Serial execution

As explored in Refs. [41,48,52], conditional sampling leaves the stationary distribution invariant. Therefore, we have no finite convergence lower bound for a serial execution model [Eq. (11)]. Note, however, that the *rate* of convergence may decay significantly.

For instance, the (relatively) well-behaved functions examined in Ref. [41] incurred a linear slowdown in $D_{\text{KL}}(\mu \| \pi)$ convergence. Spin-glass optimization is far from well behaved. We strongly conjecture that conditional sampling on highly nonconvex potentials incurs

superpolynomial slowdowns, though we leave exact characterization to future work.

B. Concurrent execution

1. Lower bounds

It is clear that the concurrent execution model only approximates full Ising-machine behavior. One method of comparing the approximation accuracy is to consider the difference between the sample distribution of the concurrent IM network and that of a monolithic IM. The distance (or divergence) between these distributions (Sec. II B) in the limit $t \to \infty$ is the *bias* of the approximation.

We consider two processes that approximate real-world machine construction. For clarity, we denote the state of the first process as $Y(t) \triangleq Y_t$ and that of the second as $X(t) \triangleq X_t$.

The first is an ideal overdamped Langevin process Y_t with instantaneous interaction between components:

$$dY_t = \big(\nabla U_{\text{Int}}(Y_t) + \nabla U_{\text{Ext}}(Y_0)\big)dt + \sqrt{2\beta^{-1}}dW_t. \quad (24)$$

This is an approximation of a "monolithic" physical Ising machine, where the on-chip spin interaction is fast enough to be modeled as instantaneous. The second process models multiple independent chips with a finite communication bandwidth. In such a system, practical implementation may not allow all spin-state changes on one chip to be seen by other chips (let alone instantaneously). One possibility is the periodic exchange of state updates among chips [17]. We approximate such a design with a concurrent-mode process $X(t) \triangleq X_t$,

$$dX_t = \big(\nabla U_{\text{Int}}(X_t) + \nabla U_{\text{Ext}}(X_0)\big)dt + \sqrt{2\beta^{-1}}dW_t, \quad (25)$$

where $t \in [0, \tau)$, X_0 is the system state at the last synchronization time point, and X_τ is a Markov process obeying Eq. (12), thereby fitting the "concurrent execution" model defined in the previous section. By definition, the processes X_t and Y_t are synchronously coupled, as they share the same Brownian noise source. The X_t process represents the networked analog system, while Y_t gives the dynamics in the $\tau \to 0$ limit. Let ν_t and μ_t be the distributions underlying processes Y_t and X_t, respectively, at time t.

Given suitable regularity conditions on U (satisfied by a relaxed Ising Hamiltonian), ν_t (24) is guaranteed to converge to the stationary distribution π (albeit in time exponential in the problem dimension) [53]; however, the same is *not* true for μ_t, as it is asymptotically biased. We can intuitively see why: μ_t factorizes across partitions, whereas π (in general) does not; however, we would still like to know *how close* the two distributions are, and what parameters affect the asymptotic bias.

Suppose we start both processes from the same initial point $X_0 = Y_0$ and then observe the difference between their sample paths. Recall from Sec. II that the KL divergence (also known as the relative entropy) is given by

$$\mathrm{D}_{\text{KL}}(\mu \| \nu) = \int_{\mathcal{D}} \mu(x) \log \frac{d\mu}{d\nu}(x) dx. \quad (26)$$

We can express the relative entropy between the concurrent and ideal systems using the gradient error in the following proposition (proved in Appendix A).

Proposition 1 (KL divergence of the approximated process). Let μ_τ and ν_τ be the probability distributions of the concurrent process and full-system processes, respectively, at time $\tau \geq 0$. Suppose

$$\tau \langle \|\delta \nabla U\|^2 \rangle = \int_0^\tau \|\nabla U_{\text{Ext}}(X_t) - \nabla U_{\text{Ext}}(X_0)\|^2 \, dt \quad (27)$$

is the accumulated gradient error from process X up to time τ. The KL divergence between the two measures is then

$$\mathrm{D}_{\text{KL}}(\mu_\tau \| \nu_\tau) = \frac{\beta}{4} \mathbb{E}_{\mu_\tau}[\tau \langle \|\delta \nabla U\|^2 \rangle]. \quad (28)$$

Equation (28) makes intuitive sense: the distance between our "ideal" process Y_t and our approximated process (X_t) is a function of the accumulated gradient error. This can also be understood using existing notions of information flow from the literature. The Liang-Kleeman information flow [54], which measures the effect of one subsystem on the entropy of another, also depends on instantaneous partial derivatives of the potential function with respect to subsystem parameters. By conditioning the external parameters, we impose a bottleneck on system information exchange, thereby limiting the change in the entropy of each subsystem.

Similarly, the relative Fisher information $\mathrm{FI}(\mu_\tau \| \nu_\tau)$ between μ_τ and ν_τ can be expressed as

$$\mathrm{FI}(\mu_\tau \| \nu_\tau) = \mathbb{E}_{\mu_\tau}[\big\| \tau \langle (\nabla^2 \delta U) \delta \nabla U \rangle \big\|^2], \quad (29)$$

where

$$\nabla^2 \delta U = \nabla^2 U_{\text{Ext}}(X_s) - \nabla^2 U_{\text{Ext}}(X_0),$$

using the notation $\nabla^2 f(x)$ as the Hessian of $f(x)$. The Fisher information captures the local differential structure affecting the differences between the measures. Therefore, our "gradient error" framework also captures the local curvature of measure space. A corollary is that linear models obeying Eq. (15) have no curvature in measure space, while nonlinear Ising-oscillator-based Ising machines navigate a richer landscape. We leave a full comparison between linear and nonlinear models to future work.

Note that by Pinsker's inequality, we can restate (28) as an upper bound on the more familiar total variation distance δ:

$$\delta(\mu,\nu) \le \sqrt{\frac{\beta}{8}\mathbb{E}_{\mu_\tau}[\tau\langle\|\delta\nabla U\|^2\rangle]}. \tag{30}$$

2. *Sufficient conditions for convergence*

Equation (28) provides useful theoretical lower bounds; however, it does not provide much direct insight into the problem parameters that may impact convergence, or when convergence to π will stall. In this section, we use concepts from optimal transport [55] to provide *sufficient* conditions for the concurrent process to decrease the distance to π within an iteration.

Recall from Sec. II that the 1-Wasserstein distance is given by

$$W_1(\mu_t,\nu_t) = \inf_{\gamma\in\mathcal{C}(\mu_t,\nu_t)} \mathbb{E}_\gamma\left[\|X_t - Y_t\|_1\right]. \tag{31}$$

The 1-Wasserstein distance satisfies the triangle inequality

$$W_1(\mu_t,\pi) \le W_1(\pi,\nu_t) + W_1(\nu_t,\mu_t), \tag{32}$$

where π is our target distribution.

For the ideal Langevin process, $W_1(\pi,\nu_t) \le C(t)W_1(\pi,\nu_0)$, where $C(t):\mathbb{R}\to(0,1)$ is a decreasing function of t [30,53,56,57]. For general nonconvex problems, convergence rates are often logarithmic (see Ref. [58] for convergence in total variation distance, for instance), making convergence essentially asymptotic. Since $\mu_0 = \nu_0$, if we have

$$\begin{aligned} W_1(\mu_t,\nu_t) + C(t)W_1(\pi,\mu_0) &< W(\pi,\mu_0) \\ \frac{W_1(\mu_t,\nu_t)}{1-C(t)} &< W(\pi,\mu_0), \end{aligned} \tag{33}$$

then we have some guaranteed contraction

$$W_1(\pi,\mu_t) < W(\pi,\mu_0). \tag{34}$$

The key, then, is to estimate $W_1(\mu_t,\nu_t)$. For any functional $G(\mu_t,\nu_t) \ge W_1(\mu_t,\nu_t)$, we can take the stronger condition

$$\frac{G(\mu_t,\nu_t)}{1-C(t)} < W(\pi,\mu_0) \tag{35}$$

as sufficient for contraction in W_1. In this section, we provide two such functionals upper-bounding $W_1(\mu_t,\nu_t)$, both proved in Appendix B.

As in the previous section, the processes X_t (concurrent) and Y_t (ideal) are synchronously coupled and begin from the same initial state. First suppose we have access to both processes. We can then upper-bound $W_1(\mu_t,\nu_t)$ using the following result:

Proposition 2 (Upper bound (I)). Let μ_t and ν_t be the distributions of X_t and Y_t within a concurrent execution epoch. Then,

$$\begin{aligned} W_1(\mu_t,\nu_t) \le& \mathbb{E}_{\nu_t\times\mu_t}\|\nabla U(Y_t) - \nabla U(X_t)\|_1 \\ &+ \mathbb{E}_{\mu_t}\|\nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_t)\|_1. \end{aligned} \tag{36}$$

Substituting (36) into (35), we obtain the contraction condition for W_1:

Corollary 1. Let μ_t be the distribution of X_t within a concurrent execution epoch. Then, the condition

$$\begin{aligned} &(1-C(t))^{-1}\left[\mathbb{E}_{\nu_t\times\mu_t}\|\nabla U(Y_t) - \nabla U(X_t)\|_1\right. \\ &\qquad \left. + \mathbb{E}_{\mu_t}\|\nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_t)\|_1\right] \\ &< W_1(\mu_0,\pi) \end{aligned} \tag{37}$$

is sufficient to guarantee $W_1(\mu_t,\pi) < W(\mu_0,\pi)$.

Nonetheless, outside of simulations, we do not have access to a synchronously coupled, ideal Langevin process. With some additional assumptions, we can express a contraction condition that only depends on the problem structure and the trajectory X_t.

Let $\mathcal{D}$ be the domain of the Hamiltonian H [e.g., the unit hypercube for the linear model in Eq. (15)]. We assume that $\nabla U(X)$ is L-Lipschitz continuous in $\mathcal{D}$, that is, there exists a constant $L \ge 0$ such that for all $X, Y \in \mathcal{D}$,

$$\|\nabla U(X) - \nabla U(Y)\|_p \le L\,\|X - Y\|_p \tag{38}$$

for $p \in \{1,2\}$, where we use the same constants for the ℓ_1 and ℓ_2 norms for simplicity. We only assume Lipschitz continuity on $\mathcal{D}$, allowing for the possibility of nonglobally Lipschitz confining potentials, such as cubic or quartic polynomials, if $\mathcal{D}$ is compact. In all analysis, we therefore assume that the processes X_t and Y_t are confined to $\mathcal{D}$.

Proposition 3 (Sufficient condition for convergence (II)). Let μ_t and ν_t be the distributions of X_t and Y_t within a concurrent execution epoch. Assume that $0 \le tL < 1$ and that $\mathbb{E}_\gamma\|X_t - Y_t\|_1$ is nondecreasing with time, where γ is the optimal coupling between μ_t and ν_t. Then,

$$W_1(\mu_t,\nu_t) \le \frac{\mathbb{E}_{\mu_t}\|\nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_t)\|_1}{1-tL} \tag{39}$$

is sufficient to guarantee $W_1(\mu_t,\pi) < W(\mu_0,\pi)$.

As before, we immediately obtain a contraction condition for $W_1(\mu_t,\pi)$.

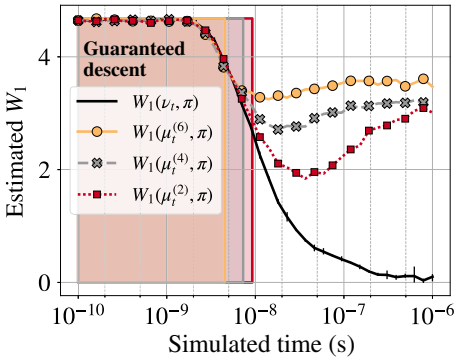


FIG. 5. Simulated W_1 convergence on a 12-spin Ising lattice for a parallel, concurrent process and an ideal Langevin process. The number of subsystems B for the concurrent process was either two, four, or six, and is denoted by the superscript $\mu_t^{(B)}$. Each solver started from a uniform random distribution. The error bars show the effect of sampling error (primarily visible on the $W_1(\nu_t, \pi)$ curve).

Corollary 2. Let μ_t be the distribution of X_t within a concurrent execution epoch. Then, the condition

$$(1 - C(t))^{-1} \frac{\mathbb{E}_{\mu_t} \|\nabla U_{\text{Ext}}(X_0) - \nabla U_{\text{Ext}}(X_t)\|_1}{1 - tL} < W_1(\mu_0, \pi) \tag{40}$$

is sufficient to guarantee $W_1(\mu_t, \pi) < W(\mu_0, \pi)$.

To numerically evaluate our bounds, we simulated ideal Y_t and concurrent X_t processes on a 12-spin ferromagnetic lattice with two, four, and six subsystems with fixed inverse temperature $\beta = 10$. For computational simplicity, we focus on the linear IM model [Eq. (15)]. For each process, we compute a discrete empirical distribution and compare this to the true Gibbs distribution. Appendix C contains further experimental details.

Figure 5 shows the estimated W_1 for the concurrent processes (with the subsystem count B denoted by the superscript $\mu_t^{(B)}$) compared against the ideal processes. For $\tau < 10^{-8}$, the concurrent processes are approximately equal to the ideal Y_t process in Wasserstein distance; however, past that point, the processes begin to deviate. As expected, the ideal process approximately converges to the true distribution. In contrast, the concurrent processes have a nonzero minimal Wasserstein distance, after which $W_1(\mu_t^{(B)}, \pi)$ begins increasing.

We also compute $\mathbb{E}_{\nu_t \times \mu_t} \|\nabla U(Y_t) - \nabla U(X_t)\|_1$ and $\mathbb{E}_{\mu_t} \|\nabla U_{\text{Ext}}(X_0) - \nabla U_{\text{Ext}}(X_t)\|_1$ during the simulation process. With these values, we compute the maximum τ with guaranteed W_1 contraction, shown as the shaded regions in Fig. 5 (with different vertical lines marking the region boundaries, color-matched to the corresponding curves). While the τ upper bounds are conservative, we note

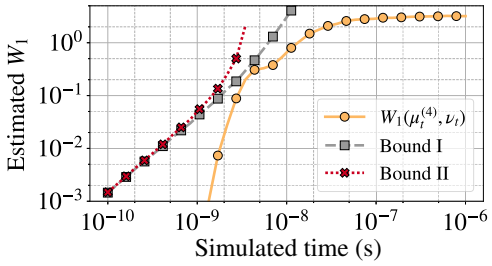


FIG. 6. Comparison between the two upper bounds derived in the main text ("Bound I" and "Bound II") and the estimated distance $W_1(\mu_t^{(4)}, \nu_t)$ for the $B = 4$ process.

that they are within an order of magnitude of the minimal W_1 for each block count. The Bound-II results (not shown) were overly conservative and were not satisfied beyond 10^{-9}.

Figure 6 compares Bounds I and II against the computed value $W_1(\mu_t, \nu_t)$. The behavior was similar across block counts; therefore, we only show the $B = 4$ case for clarity. Each bound is relatively tight for $\tau < 10^{-9}$, with each bound able to guarantee some contraction; however, Bound II is no longer valid for $\tau > 1/L$ (approximately 3.9×10^{-9} s), and Bound I becomes too loose for practical use past 10^{-8}. The latter value also roughly corresponds to the region where we observe energy divergence in larger problems, as we now show.

To demonstrate the effects of increasing τ on larger-scale problems, we compare the concurrent and serial execution models on $N = 2000$ Sherrington-Kirkpatrick (SK) spin glasses [59]. Each instance has the Hamiltonian

$$H(s) = \sum_{i<j} J_{ij} s_i s_j, \tag{41}$$

where J_{ij} are normally distributed with variance $1/N$. We simulate system behavior for 20 μs across three problems with 50 trials each. Due to the lack of cross-partition communication, the $B = 1$ configuration produced consistent results across τ values. Therefore, we report the u value relative to the mean $B = 1$ energy per spin ($\overline{u}_1 \sim -0.756$) averaged across all trials.

Figure 7 shows the "energy error" $u - \overline{u}_1$ for the serial and concurrent processes with varying block configurations and increasing epochs. For $\tau < 10^{-8}$, the concurrent processes attain a consistently lower energy error than the serial processes with the same total annealing time. For higher τ values, the energy estimates begin to diverge from the monolithic system results. As in the toy ferromagnetic problem, configurations with more subsystems diverge more rapidly, with the estimated energy becoming positive for the four- and eight-system configurations. The $B = 2$ system remains close to the serial estimate; however, it still demonstrates a significant loss in accuracy with increasing τ.

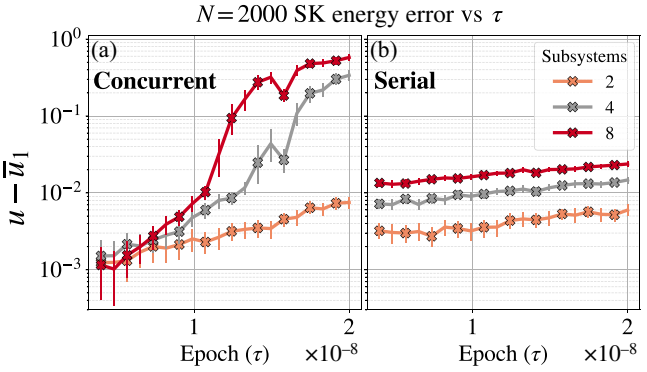


FIG. 7. Deviation of the estimated energy per spin relative to the $B = 1$ case for 2000-node SK spin-glass instances. Each data point represents three different SK instances with 50 trials each.

In contrast, the serial processes experience near-linear degradation in the estimated energy with increasing τ. Unlike the concurrent system, where τ fundamentally alters system dynamics, the serial system simply requires more iterations to regain accuracy, as observed in Ref. [17]. Serial execution models are therefore more robust to low-synchronization-frequency environments. We therefore conjecture that serial models are also more robust in the presence of asynchronous updates, though we leave that comparison to future work.

V. EMPIRICAL ANALYSIS

Having established theoretical bounds for general separable dynamical systems, we narrow our focus to linear Ising machines evolving according to Eq. (15). We begin by deriving a heuristic parameter choice based on Proposition 3, and we then empirically evaluate the time and energy efficiency of serial and concurrent execution models for unconstrained binary optimization. We conclude by offering a decision-tree heuristic informed by our findings.

A. Heuristic parameter selection

While we numerically show that Eqs. (37) and (40) hold in toy problems, real-world users lack access to the W_1 contraction factor $C(t)$ as well as the starting $W_1(\mu_0, \pi)$. In contrast, we do have access to the problem structure in the coupling matrix J. Proposition 3 suggests that the spectrum of J is a limiting factor in convergence. In this section, we heuristically derive a similar relation for Ising machines by analyzing the flip time of a single spin.

Consider a single degree of freedom $x_i \in [-1, 1]$ in subsystem X_p. We have

$$dx_i = \left[\sum_{x_j \in X_p} J_{ij} x_j + \sum_{x_j \in X_{\setminus p}} J_{ij} x_j \right] dt + \sqrt{2T} dW_t. \tag{42}$$

From Itô's lemma, we have

$$d(x_i^2) = \left[2x_i \cdot \left(\sum_{x_j \in X_p} J_{ij} x_j + \sum_{x_j \in X_{\setminus p}} J_{ij} x_j \right) + 2T \right] dt + 2x_i \sqrt{2T} dW_t, \tag{43}$$

hence,

$$\mathbb{E}[x_{i,\tau}^2 - x_{i,0}^2] = \int_0^\tau \left[\mathbb{E}\left(2x_i \sum_{x_j \in X_p} J_{ij} x_j + 2x_i \sum_{x_j \in X_{\setminus p}} J_{ij} x_j \right) + 2T \right] dt. \tag{44}$$

Assuming that the gradient

$$J_i \triangleq \sum_{x_j \in X_p} J_{ij} x_j + \sum_{x_j \in X_{\setminus p}} J_{ij} x_j \tag{45}$$

remains (relatively) constant during the annealing epoch, we have

$$\mathbb{E}[x_{i,\tau}^2 - x_{i,0}^2] \sim \mathbb{E}[(x_{i,\tau}^2 - x_{i,0}^2) J_i + 2T]\tau. \tag{46}$$

For combinatorial problems, the quantity of interest between $x_{i,\tau}$ and $x_{i,0}$ is the *Hamming distance* [60], not the ℓ_1 or ℓ_2 distances. Therefore, we can focus our attention on *spin flips* rather than the entire trajectory. Assuming that the spin flips within the epoch ($\mathbb{E}[x_{i,\tau}^2] = 0$) and the spins began the epoch in a bifurcated state ($\mathbb{E}[x_{i,\tau}^2] = 1$), we can solve for τ_{Flip} as

$$\tau_{\text{Flip}} \sim \frac{1}{(J_i - 2T)}. \tag{47}$$

In the low-temperature limit $J_i \gg 2T$, we further approximate as

$$\tau_{\text{Flip}} \sim \frac{1}{J_i} \geq \frac{1}{\rho(J)}, \tag{48}$$

where $\rho(J) = \max_i |\lambda_i|$ is the spectral radius of the coupling matrix J. Selecting $\tau < \tau_{\text{Flip}}$ results in relatively few spin flips occurring each epoch, decreasing the Hamming distance between X_0 and X_τ. As we show in the next subsection, we find that using the spectral radius is overly conservative, and in practice, using the mean eigenvalue magnitude $\overline{|\lambda|}(J) = 1/N \sum_{i=1} |\lambda_i|$ is sufficient to maintain high-quality solutions without unnecessarily high bandwidth.

B. Optimization performance

In this section, we explore how parallel execution configurations impact performance in unconstrained optimization problems. We focus on six $N = 2000$ graph instances from the GSet [19] MaxCut suite. We specifically focus on Erdős-Rényi (ER) uniform random graphs [61] and Barabási-Albert (BA) scale-free graphs [62]. We test three graphs of each type ($\{G27, G28, G29\}$ and $\{G39, G40, G41\}$, respectively), each with bimodal $\{+1, -1\}$ couplings. We omit testing purely antiferromagnetic problems due to the unique spectral behavior explored on smaller instances in previous sections.

Figure 8 shows the average solution quality versus synchronization frequency for several graphs from the GSet suite for serial and concurrent execution. We simulate each solver for 20 μs, with each data point representing 40 trials. The y axis shows the "cut error" (lower is better) $1 - \text{Cut/BKS} \triangleq 1 - R$, where BKS is the best-known solution for a given problem and R is the "approximation ratio." We indicate the frequencies selected by the spectral radius $\rho(J)$ and average eigenvalue magnitude $\overline{|\lambda|}(J)$ with vertical lines.

For low frequencies (<100 MHz), both the serial and concurrent models perform poorly (high cut errors); however, for the serial mode, the cut error appears to be a smooth function of frequency, while the concurrent mode exhibits the sharp discontinuity seen in Fig. 7. As the frequency increases, both models see solution quality gains

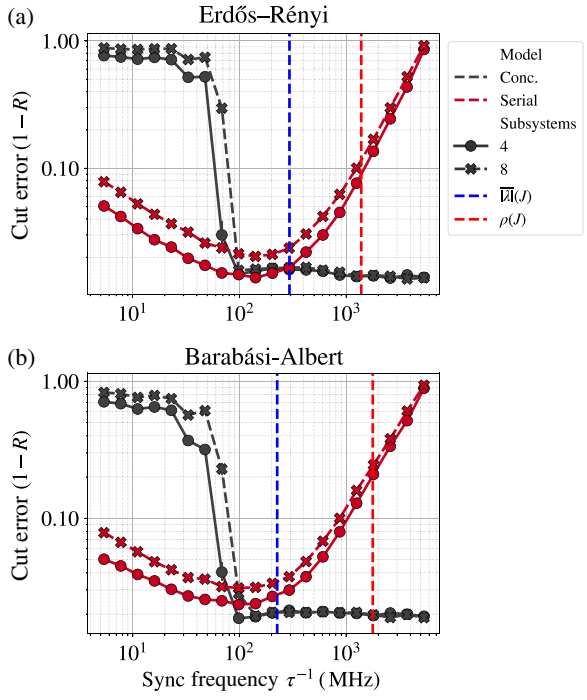


FIG. 8. Cut error $1 - R$ versus the synchronization frequency for six $N = 2000$ (a) Erdős-Rényi (ER) and (b) Barabási-Albert (BA) graphs from the GSet [19] suite.

up to the "crossover point" at approximately 100–200 MHz. Serial execution achieves higher-quality solutions at the crossover point for $B = 4$ on the ER graphs, while concurrent execution achieves lower average error rates for $B = 8$ on both the ER and BA graphs, as well as for $B = 4$ on the BA graph.

Recall that the purpose of the $\rho(J)$ and $\overline{|\lambda|}(J)$ heuristics is to provide an estimate of the high-frequency, "high solution quality" region, observed to be >100 MHz. Setting $\tau = \rho(J)$ yields a conservative estimate of 1400 and 1800 MHz for the ER and BA graphs, respectively, 14–18 times greater than the actual cutoff value. In contrast, choosing $\tau = \overline{|\lambda|}(J)$ leads to significantly tighter estimates of 247 and 165 MHz for the ER and BA graphs, respectively. Using the more liberal heuristic therefore provides over 5×, and up to 11×, reductions in communication volume for the graphs tested while remaining within the region of high solution quality.

In the high-frequency region (>100 MHz), the solution quality of the concurrent model is approximately constant, remaining at around 0.015, while the serial-mode error increases rapidly. The loss of solution quality is due to 1-bit quantization. After evolving for a τ second epoch, each serial subsystem state is clamped to ± 1. Sufficiently small τ (high frequency) impedes forward progress from the subsolver, as the system is not allowed to substantially change state before being binarized. Having higher-precision spin-state representations will mitigate this problem.

The data shown in Fig. 8 represent *mean* energies, whereas optimization tasks are typically focused on finding (some approximation to) the optimal solution. Supposing we want to achieve some proportion of the best-known solution (say 98%), it is common practice to express performance in terms of "metric to target" (MTT)

$$\text{MTT} = \langle M \rangle \frac{\log(0.01)}{\log(1 - p_{\text{Success}})}.$$

The MTT measures the estimated "metric" needed to achieve a target solution quality with 99% probability. This metric is typically time; however, in this work, we also include energy as a principal cost.

We set a solution-quality target as 98% of the best-known solution (BKS), taken from Ref. [68]. Therefore, the reported time to target (TTT) or energy to target (ETT) is the time or energy required to reach 98% of the BKS. Figure 9 shows the TTT for serial and concurrent modes as a function of the update frequency. Missing values indicate that the target quality was not reached in 40 trials. Concurrent-mode simulations lasted 20 μs; however, initial serial operation testing at 20 μs only reached the solution-quality target for a narrow frequency band (shown as light and dark shaded bands for $B = 4$ and $B = 8$ respectively). As such, we also simulated the serial system for 40, 60, 80, and 100 μs and took the lowest TTT figure

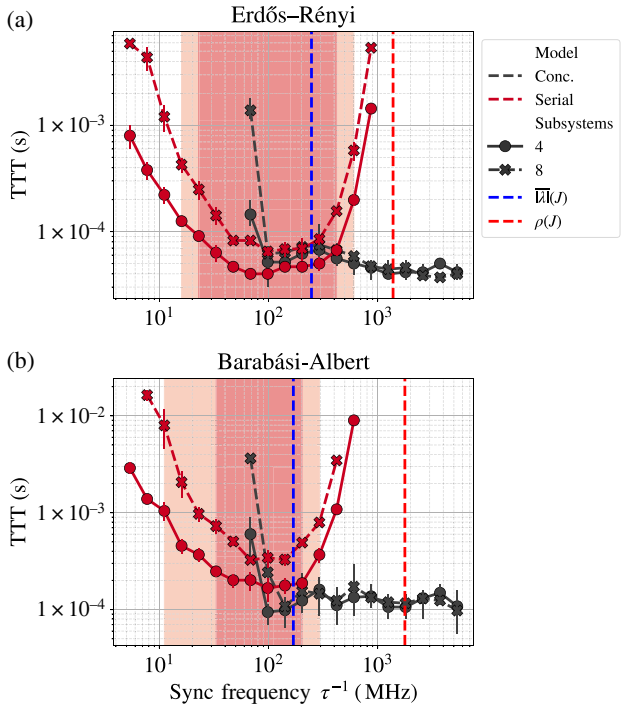


FIG. 9. Mean time to target for the serial and concurrent models with increasing frequency (decreasing epoch τ) on $N = 2000$ GSet (a) Barabási-Albert and (b) Erdős-Rényi graphs. We also show the frequencies selected with the spectral heuristics, comparing the mean eigenvalue magnitude $\overline{|\lambda|}(J)$ (blue) and the spectral radius $\rho(J)$ (red). Shaded regions denote the τ values where the serial model reached the solution quality target in 20 μs. Lighter regions denote $B = 4$ and darker regions denote $B = 8$.

for comparison. This methodology follows from the differences between each approach. In serial operation, 20 μs of *total* annealing time with eight blocks results in only 2.5 μs of annealing time per block. Therefore, increasing the annealing time follows from the necessary increase in latency required from adopting serial-mode execution.

As in the mean solution quality results, serial execution outperformed concurrent execution in ER graphs, nearly matching the high-frequency concurrent TTT with update frequencies $\leq$200 MHz; however, in BA graphs, concurrent execution has over a 2× TTT advantage for frequencies over 100 MHz. Increasing the subsystem count to $b = 8$ leads to a TTT value around 10× lower for concurrent systems compared with serial execution.

We also observe that in both the mean cut error and the TTT, selecting $\tau^{-1} = \rho(J)$ is overly conservative, requiring over 5× higher frequency compared to $\tau^{-1} = \overline{|\lambda|}(J)$. The latter was within a factor of 2 of the optimal TTT point for both ER and BA graphs, leading to a superior tradeoff between frequency and TTT. Although the increased frequency requirement does not affect the latency, it has a marked effect on energy costs.

As we use 1-bit representations for X_0, we can also quantify the energy cost stemming from communication. Figure 10 shows the estimated ETT resulting from synchronization updates with increasing sync frequency τ^{-1} and energy per bit E_{Bit}. To put the required bandwidths or energies into perspective, we annotate the plot with memory interface technologies placed according to their peak throughput and estimated energy per bit.

We compute the energy E as

$$E = 2000 E_{\text{Bit}} \left\lfloor \frac{t_{\text{Anneal}}}{\tau} \right\rfloor, \tag{49}$$

where $\lfloor t_{\text{Anneal}}/\tau \rfloor$ gives the number of synchronizations and E_{Bit} is the energy per bit. We consider expended solver energy to be negligible compared to communication overhead; moreover, the expected accelerator energy would be similar for concurrent and serial architectures in any case.

Figure 10 illustrates the hidden cost of running high-frequency updates. The left-hand subplots show serial and concurrent execution with four subsystems, while the right-hand subplots show serial and concurrent execution with eight subsystems. We plot ETT values for a range of E_{Bit} values (x axis), with markers indicating the (maximum τ^{-1}, E_{Bit}) pairs for existing memory technologies. As with the TTT, missing values indicate that the target solution quality was not reached in 40 trials.

While the time to target stays relatively constant for concurrent execution above 100 MHz, the energy cost increases linearly with frequency from the communication overhead. Accordingly, $\tau = \rho(J)$ incurs a 10× higher energy cost compared to $\tau = \overline{|\lambda|}(J)$. Below the 100-MHz threshold, the concurrent ETT abruptly diverges, while the serial ETT demonstrates a gradual upward trend.

From the contours shown, optimal concurrent ETT is attained by minimizing the operating frequency without going below the stability threshold (around 100 MHz). In contrast, optimal serial ETT occurs at lower frequencies (around 50 MHz), with a wider base of near-optimal frequencies. The ETT therefore drastically differs from the TTT, where concurrent operation has a distinct advantage (particularly in the eight-block case). The two approaches achieve similar ETT in the mid-frequency range (around 70–500 MHz), while serial operation extends the energy-efficiency region into the low-frequency (<50 MHz) range. Concurrent operation is capable of recovering high-quality solutions for >500 MHz; however, the marginal decrease in TTT is overwhelmed by the increasing energy cost of communication. The opposite is true of serial execution, where the increase in TTT at low frequencies is (somewhat) compensated by decreased communication costs, resulting in more graceful ETT degradation.

For efficient operation, Fig. 10 implies that concurrent Ising-machine architectures require an optimized synchronization schedule to achieve (near) optimal ETT,

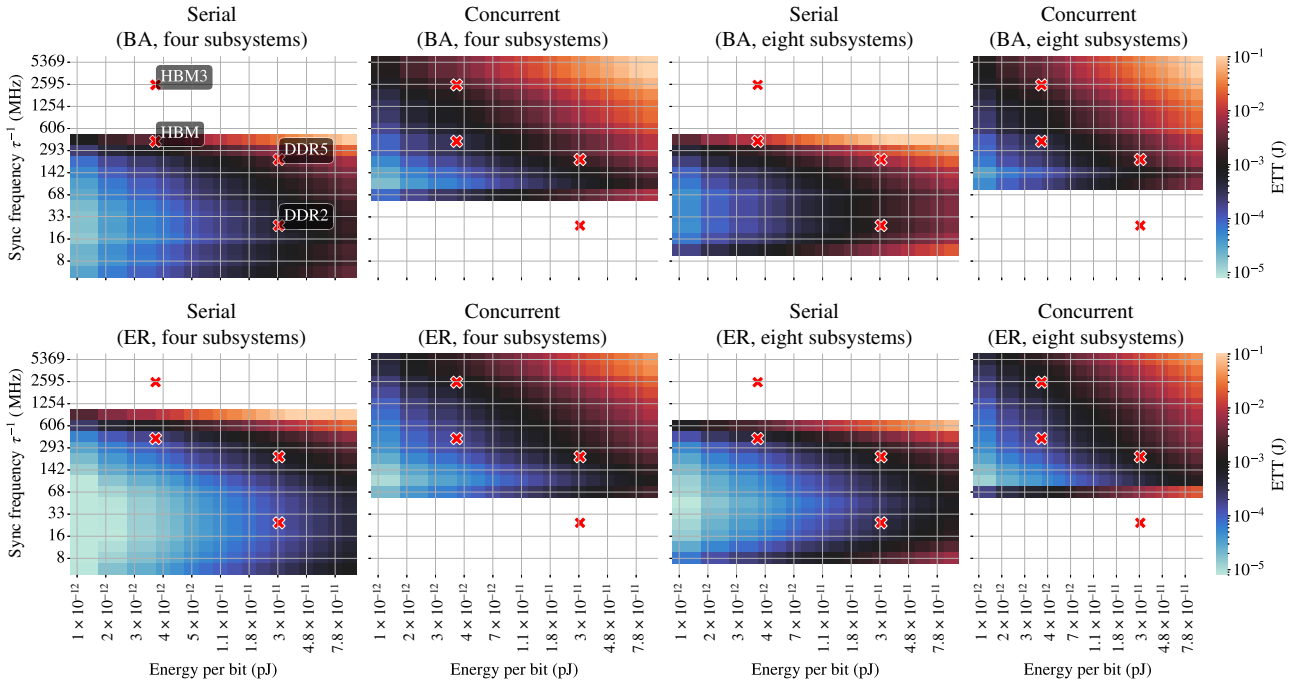


FIG. 10. Mean estimated energy to target (ETT) versus synchronization frequency τ^{-1} and the energy cost per bit for serial and concurrent models on $N = 2000$ GSet Barabási-Albert (top) and Erdős-Rényi (bottom) graphs. The maximum achievable frequencies with modern memory interfaces are shown as red markers. Energy consumption is assumed to be approximately 34 and 4 pJ/bit for DRAM [63] and HBM [64] respectively. HBM and HBM3 bandwidths are assumed to be 1 and 5.6 Tbps [65], and DDR2 and DDR5 bandwidths are 62 and 546 Gbps [66,67].

dependent on the target application and the power requirements of the application. Serial operation is observed to attain minimal ETT across a wider range of available bandwidths, allowing for more flexibility in parameter selection.

As our numerical and theoretical findings demonstrate, the gap between serial- and concurrent-mode operation is highly dependent on the synchronization epoch τ and the spectrum of J. Moreover, a comparison of Figs. 9 and 10 suggests that comparisons between concurrent and serial are sensitive to the choice of metric (time vs energy). Using the results gathered so far, we propose a simple decision tree to determine the operating mode most suitable to a given system or problem, and this is shown in Fig. 11. As Fig. 10 shows, concurrent and serial operation achieve roughly the same ETT performance across energy-per-bit values, with energy efficiency peaking in the 30–70-MHz range. Given the additional theoretical and practical complications incurred by concurrent operation, Fig. 11 shows that serial is preferable to concurrent operation in all scenarios where the ETT is a higher priority than the TTT.

If a low TTT is preferred, then the recommendation depends on the system parameters and the problem structure. Equation (48) provides an empirically conservative heuristic for a maximal admissible synchronization period τ_{Flip} for effective concurrent operation based on the spectral radius $\rho(J)$ (or a more liberal substitute such as the average eigenvalue magnitude). In our system model, $\rho(J)$ depends on both the problem structure (the problem couplings $\tilde{J}$) and on the system $R \cdot C$ time constant, related by $J = \frac{1}{R \cdot C}\tilde{J}$; therefore, $\rho(J) = \frac{1}{R \cdot C}\tilde{J}$.

If the proposed system cannot natively achieve $1/\tau_{\text{Flip}}$ synchronization frequencies, then the user can attempt to reduce $\rho(J)$, either by increasing the system's RC constant or by altering the problem formulation. For instance, lowering the penalty parameters in constrained Hamiltonian encodings lowers the spectral radius of the coupling matrix, though this must be considered in tandem with maintaining high feasibility rates. If the user is unable to alter system time constants or sufficiently change the problem formulation, then serial mode is preferable.

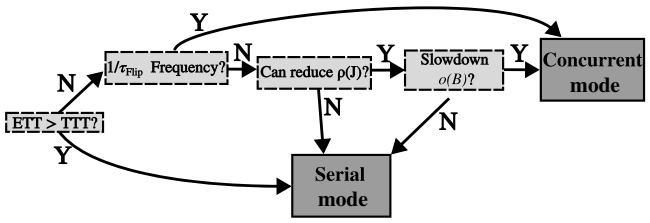


FIG. 11. Parallel Ising-machine operating-mode decision tree. Note that most cases prefer serial operation to concurrent; however, latency-critical applications with sufficiently high available bandwidth can derive advantage from concurrent mode.

Even if the system's time constant can be changed from RC to $R'C' > RC$, doing so will likely incur a net slowdown in system performance. If that slowdown $R'C'/RC = O(B)$ (where B is again the number of subsystems), then the performance edge of concurrent mode over serial mode will likely be maintained; however, if $R'C'/RC > B$, it is highly probable that the advantage of concurrent mode over serial mode has been lost, making serial again preferable by virtue of system simplicity and theoretical grounding.

To summarize, if the user prioritizes latency over energy efficiency *and all other system parameters are amenable*, then concurrent mode should be used. *In all other cases*, serial operation is preferable.

VI. CONCLUSION

In this work, we provide theoretical and numerical analyses of parallel analog Ising-machine architectures. Noting the lack of existing results for concurrent execution models, we provide lower bounds on asymptotic KL divergence as well as proving sufficient conditions for per-iteration contraction in 1-Wasserstein distance.

Numerical experiments on the linear IM model support our theoretical results, demonstrating that the problem smoothness and synchronization frequency are primary determinants of parallel-solver performance. We further characterize the energy and time trade-offs between serial and concurrent execution models, noting that concurrent architectures require optimized synchronization tuning to optimize energy efficiency, while serial execution allows for energy-efficient operation across a range of update frequencies. Finally, we propose practical heuristics for users to decide between serial and concurrent execution, dependent on user preferences, problem structure, system dynamics, and bandwidth constraints.

ACKNOWLEDGMENTS

This work was supported in part by NSF under Awards No. 2231036 and No. 2233378, and by DARPA under contract No. FA8650-23-C-7312.

DATA AVAILABILITY

The data that support the findings of this article are openly available [69].

APPENDIX A: PROOF OF PROPOSITION 1

Proof. Recall that for two probability measures μ and ν over $\mathcal{D} \subseteq \mathbb{R}^d$ with μ absolutely continuous with respect to ν, the Radon-Nikodym derivative $d\mu/d\nu$ is an integrable function that satisfies

$$\mu(A) = \int_A \nu(x) \frac{d\mu}{d\nu}(x) dx \tag{A1}$$

for any set $A \subseteq \mathcal{D}$. Let μ_t and ν_t be defined as in the previous section. Since we are concerned with the behavior of continuous, finite potentials over a bounded domain, we can assume that the system gradients ∇U, ∇U_{Int}, and ∇U_{Ext} are Lipschitz with some finite constants. Therefore, we can express the Radon-Nikodym derivative using the multidimensional Girsanov formula [see Eq. 7.138 of Ref. [70]]

$$\begin{aligned}\frac{d\mu_t}{d\nu_t}(x) = \exp\Bigg[&\frac{\beta}{2}\int_0^t \Big\langle \nabla U_{\text{Int}}(X_s) + \nabla U_{\text{Ext}}(X_0) - \nabla U(X_s), \nabla U(X_s) ds + \sqrt{2\beta^{-1}} dW_s \Big\rangle \\ &- \frac{\beta}{4}\int_0^t \langle \nabla U(X_s) - \nabla U_{\text{Int}}(X_s) - \nabla U_{\text{Ext}}(X_0), \nabla U(X_s) + \nabla U_{\text{Int}}(X_s) + \nabla U_{\text{Ext}}(X_0) \rangle \, ds \Bigg].\end{aligned} \tag{A2}$$

Disregarding the dW_s term (since it will be disregarded in later expectation calculations), we have

$$\begin{aligned}\frac{d\mu}{d\nu}(x) &= \exp\Bigg[\frac{\beta}{2}\int_0^t \langle \nabla U_{\text{Int}}(X_s) + \nabla U_{\text{Ext}}(X_0) - \nabla U(X_s), \nabla U(X_s) \rangle \, ds \\ &\qquad - \frac{\beta}{4}\int_0^t \langle \nabla U_{\text{Ext}}(X_s) - \nabla U_{\text{Ext}}(X_0), 2\nabla U(X_s) - \nabla U_{\text{Ext}}(X_s) + \nabla U_{\text{Ext}}(X_0) \rangle \, ds \Bigg] \\ &= \exp\Bigg[-\frac{\beta}{4}\int_0^t \langle \nabla U_{\text{Ext}}(X_s) - \nabla U_{\text{Ext}}(X_0), -\nabla U_{\text{Ext}}(X_s) + \nabla U_{\text{Ext}}(X_0) \rangle \, ds \Bigg] \\ &= \exp\Bigg[\frac{\beta}{4}\int_0^t \|\nabla U_{\text{Ext}}(X_s) - \nabla U_{\text{Ext}}(X_0)\|^2 \, ds \Bigg] = \exp\Bigg[\frac{t\beta}{4}\langle \|\delta \nabla U\|^2 \rangle_t \Bigg].\end{aligned} \tag{A3}$$

The KL divergence is then given by

$$\mathrm{D}_{\mathrm{KL}}(\mu_t\|\nu_t) = \int \mu_t(X)\frac{t\beta}{4}\langle\|\delta\nabla U\|^2\rangle_t dX = \mathbb{E}_{\mu_t}\left[\frac{t\beta}{4}\langle\|\delta\nabla U\|^2\rangle_t\right]. \tag{A4}$$

■

APPENDIX B: PROOF OF PROPOSITIONS 2 AND 3

1. Proof of Proposition 2

Proof. For simplicity, we let $X_0 = X^m$ be the initial state and consider the evolution between synchronization steps from time $t \in [0, \tau]$. The 1-Wasserstein distance between μ_t and ν_t is

$$\begin{aligned} W_1(\mu_t, \nu_t) &= \inf_{\gamma\in\mathcal{C}(\mu_t,\nu_t)} \mathbb{E}_\gamma\left[\|X_t - Y_t\|_1\right] \\ &= \inf_{\gamma\in\mathcal{C}(\mu_t,\nu_t)} \mathbb{E}_\gamma\left[\left\|\int_0^t \nabla U_{\mathrm{Int}}(X_s) + \nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Int}}(Y_s) - \nabla U_{\mathrm{Ext}}(Y_s)ds\right\|_1\right], \end{aligned} \tag{B1}$$

which we can rearrange to form

$$W_1(\mu_t, \nu_t) = \inf_{\gamma\in\mathcal{C}(\mu_t,\nu_t)} \mathbb{E}_\gamma\left[\left\|\int_0^t \nabla U(X_s) - \nabla U(Y_s) + \nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_s)ds\right\|_1\right]. \tag{B2}$$

By the optimality of γ, we can upper-bound $W_1(\mu_t, \nu_t)$ by the transport distance using the product measure $\nu \times \mu$:

$$W_1(\mu_t, \nu_t) \le \mathbb{E}_{\nu\times\mu}\left[\left\|\int_0^t \nabla U(X_s) - \nabla U(Y_s) + \nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_s)ds\right\|_1\right]. \tag{B3}$$

Using the triangle inequality and integrating each term, noting that $X_0 = Y_0$, we obtain

$$\begin{aligned} W_1(\mu_t, \nu_t) &\le \mathbb{E}_{\nu\times\mu}\left\|\int_0^t \nabla U(X_s) - \nabla U(Y_s)dt\right\|_1 + \mathbb{E}_{\nu\times\mu}\left\|\int_0^t \nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_s)dt\right\|_1 \\ &= \mathbb{E}_{\nu\times\mu}\|\nabla U(Y_t) - \nabla U(X_t)\|_1 + \mathbb{E}_\mu\|\nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_t)\|_1. \end{aligned} \tag{B4}$$

We then have an upper bound on $W_1(\mu, \nu)$ in terms of the iterates X and Y. Applying the triangle inequality, we obtain

$$\begin{aligned} W_1(\mu_t, \pi) &\le W_1(\nu_t, \pi) + W_1(\nu_t, \mu_t) \\ &\le \underbrace{C(t)W_1(\nu_0, \pi) + \mathbb{E}_{\nu\times\mu}\|\nabla U(Y_t) - \nabla U(X_t)\|_1 + \mathbb{E}_\mu\|\nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_t)\|_1}_{\text{(a)}}. \end{aligned} \tag{B5}$$

If we have (a) $< W_1(\mu_0, \pi)$, then we have guaranteed contraction. With a bit of algebra, we obtain

$$\frac{\mathbb{E}_{\nu\times\mu}\|\nabla U(Y_t) - \nabla U(X_t)\|_1 + \mathbb{E}_\mu\|\nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_t)\|_1}{1 - C(t)} < W_1(\nu_0, \pi) \tag{B6}$$

as a sufficient condition for contraction. ■

2. Proof of Proposition 3

We use Eq. (B6) in numerical experiments, where we can empirically estimate ν_t and μ_t; however, the behavior of the ideal process Y_t is unknown in practice, making (B6) interesting, but not particularly useful.

Assuming that the target potential is L-smooth, we provide a stronger condition that depends only on device characteristics, problem structure, and the concurrent process X_t. If we assume that the expected distance $\mathbb{E}_\gamma\|X_s - Y_s\|_1$ is increasing with time and that $tL < 1$, we have

$$\frac{\mathbb{E}_{\mu_t}\|\nabla U_{\mathrm{Ext}}(X_0) - \nabla U_{\mathrm{Ext}}(X_t)\|_1}{1 - tL} < W_1(\mu_0, \nu_t). \tag{B7}$$

Proof. Applying the triangle inequality, we obtain

$$W_1(\mu_t, \nu_t) \leq \inf_{\gamma \in \mathcal{C}(\mu_t, \nu_t)} \mathbb{E}_\gamma \left[\left\| \int_0^t \nabla U(X_s) - \nabla U(Y_s) \right\|_1 ds + \left\| \int_0^t \nabla U_{\text{Ext}}(X_0) - \nabla U_{\text{Ext}}(X_s) ds \right\|_1 \right]. \tag{B8}$$

Using Jensen's inequality and the smoothness of U, we obtain

$$W_1(\mu_t, \nu_t) \leq \inf_{\gamma \in \mathcal{C}(\mu_t, \nu_t)} \mathbb{E}_\gamma \left[\int_0^t L \left\| X_s - Y_s \right\|_1 ds + \left\| \int_0^t \nabla U_{\text{Ext}}(X_0) - \nabla U_{\text{Ext}}(X_s) ds \right\|_1 \right]. \tag{B9}$$

Assuming that $\mathbb{E}_\gamma \left\| X_s - Y_s \right\|_1$ is nondecreasing (i.e., $\sup_{s \in [0,t]} \mathbb{E}_\gamma \left\| X_s - Y_s \right\|_1 = \mathbb{E}_\gamma \left\| X_t - Y_t \right\|_1$), we can apply the upper bound

$$\begin{aligned} W_1(\mu_t, \nu_t) &\leq \inf_{\gamma \in \mathcal{C}(\mu_t, \nu_t)} t \mathbb{E}_\gamma \left[L \left\| X_t - Y_t \right\|_1 + \left\| \int_0^t \nabla U_{\text{Ext}}(X_0) - \nabla U_{\text{Ext}}(X_s) ds \right\|_1 \right] \\ &= tLW_1(\mu_t, \nu_t) + \mathbb{E}_{\mu_t} \left\| \nabla U_{\text{Ext}}(X_0) - \nabla U_{\text{Ext}}(X_t) \right\|_1 . \end{aligned} \tag{B10}$$

For $tL < 1$, we obtain

$$W_1(\mu_t, \nu_t) \leq \frac{\mathbb{E}_{\mu_t} \left\| \nabla U_{\text{Ext}}(X_0) - \nabla U_{\text{Ext}}(X_t) \right\|_1}{1 - tL}. \tag{B11}$$

■

APPENDIX C: W_1 EXPERIMENTAL DETAILS

Here, we provide experimental details for the numerical results in Sec. IV. The simulation code was written in JULIA and run on an Intel i9-13900k CPU with 64 GB of RAM.

For a target problem, we randomly initialize $N = 3000$ problem states from a uniform random distribution over $\{-1, 1\}$. Each state is a different initialization of X_0, allowing for statistical averaging.

We then evolve two synchronously coupled processes, X_t and Y_t, for $t \in [0, \tau]$ using an Euler-Maruyama discretization of (15) with a step size of $dt = 10^{-12}$ s. We use $R = 310$ kΩ and $C = 50$ fF for the device parameters, leading to an effective time constant of 15.5 ns. After τ seconds, we quantize the states to the nearest hypercube vertex using the sign function. At $\beta = 10$, the continuous processes are strongly concentrated around the hypercube vertices, making the quantized state a reasonable approximation of the continuous state.

The quantized states are used to compute an empirical distribution over the $2^{12} = 4096$ vertices of the hypercube, forming a distribution $\mu \in \mathbb{R}^{4096}$. We also compute the true distribution π by fully enumerating the spin-glass states. Using the package PYTHONOT.JL (the JULIA interface for PYTHONOT [71]), we compute the W_1 distance between the distributions with the emd2 function. The distance matrix was computed using pairwise Hamming distances between states.

Our target graph was a 12-spin ferromagnetic system, arranged in a two-dimensional, periodic 4×3 lattice. Despite our broader focus on nonconvex problems, we selected a ferromagnetic instance to avoid the high estimation variance resulting from solver samples being stuck in local minima.

[1] T. Albash, and D. A. Lidar, Adiabatic quantum computation, Rev. Mod. Phys. **90**, 015002 (2018), publisher: American Physical Society.

[2] P. Hauke, H. G. Katzgraber, W. Lechner, H. Nishimori, and W. D. Oliver, Perspectives of quantum annealing: Methods and implementations, Rep. Prog. Phys. **83**, 054401 (2020), publisher: IOP Publishing.

[3] R. Afoakwa, Y. Zhang, U. K. R. Vengalam, Z. Ignjatovic, and M. Huang, in *2021 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (2021), p. 749, ISSN: 2378-203X.

[4] T. Wang, and J. Roychowdhury, OIM: Oscillator-based Ising machines for solving combinatorial optimisation problems, ArXiv:1903.07163.

[5] T. Inagaki, Y. Haribara, K. Igarashi, T. Sonobe, S. Tamate, T. Honjo, A. Marandi, P. L. McMahon, T. Umeki,

K. Enbutsu, O. Tadanaga, H. Takenouchi, K. Aihara, K.-i. Kawarabayashi, K. Inoue, S. Utsunomiya, and H. Takesue, A coherent Ising machine for 2000-node optimization problems, Science **354**, 603 (2016), publisher: American Association for the Advancement of Science.

[6] N. Mohseni, P. L. McMahon, and T. Byrnes, and Ising machines as hardware solvers of combinatorial optimization problems, Nat. Rev. Phys. **4**, 363 (2022).

[7] T. Zhang, Q. Tao, B. Liu, A. Grimaldi, E. Raimondo, M. Jiménez, M. J. Avedillo, J. Nuñez, B. Linares-Barranco, T. Serrano-Gotarredona, G. Finocchio, and J. Han, A review of Ising machines implemented in conventional and emerging technologies, IEEE Trans. Nanotechnol. **23**, 704 (2024), conference Name: IEEE Transactions on Nanotechnology.

[8] E. Elmitwalli, Z. Ignjatovic, and S. Kose, Utilizing multibody interactions in a CMOS-based Ising machine for LDPC decoding, IEEE Trans. Circuits Syst. I: Regul. Papers **71**, 40 (2024), conference Name: IEEE Transactions on Circuits and Systems I: Regular Pap..

[9] A. Sharma, M. Burns, A. Hahn, and M. Huang, Augmenting an electronic Ising machine to effectively solve Boolean satisfiability, Sci. Rep. **13**, 22858 (2023), publisher: Nature Publishing Group.

[10] S. Tan, M. Yu, A. Python, Y. Shang, T. Li, L. Lu, and J. Yin, in *2023 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, Montreal, QC Canada, 2023), p. 731, ISSN: 2378-203X.

[11] U. K. R. Vengalam, Y. Liu, T. Geng, H. Wu, and M. Huang, in *Proceedings of the 56th Annual IEEE/ACM International Symposium on Microarchitecture*, MICRO '23 (Association for Computing Machinery, New York, NY, USA, 2023), p. 465.

[12] N. Matsumoto, Y. Hamakawa, K. Tatsumura, and K. Kudo, Distance-based clustering using QUBO formulations, Sci. Rep. **12**, 2669 (2022).

[13] When solving spin-glass Hamiltonians (and the combinatorial optimization problems they define) using a naive divide-and-conquer approach, the need for digital pre- and postprocessing for each subproblem quickly dominates the workload, limiting the potential speedup and incurring high energy costs from the additional communication and processing required.

[14] C. McGeoch, P. Farre, and W. Bernoudy, D-Wave Hybrid Solver Service + Advantage: Technology Update, 2020, https://www.dwavequantum.com/media/m2xbmlhs/14-1048a-a_d-wave_hybrid_solver_service_plus_advantage_technology_update.pdf

[15] Y. Du, H. Wang, R. Hennig, A. Hulandageri, G. Kochenberger, and F. Glover, New advances for quantum-inspired optimization, Int. Trans. Oper. Res. **32**, 6 (2025).

[16] P. R. Kalehbasti, H. Ushijima-Mwesigwa, A. Mandal, and I. Ghosh, in *Advances in Intelligent Data Analysis XIX*, edited by P. H. Abreu, P. P. Rodrigues, A. Fernández, and J. Gama (Springer International Publishing, Cham, 2021), p. 350.

[17] A. Sharma, R. Afoakwa, Z. Ignjatovic, and M. Huang, in *Proceedings of the 49th Annual International Symposium on Computer Architecture*, ISCA '22 (Association for Computing Machinery, New York, NY, USA, 2022), p. 508.

[18] M. Booth, S. P. Reinhardt, and A. Roy, Partitioning Optimization Problems for Hybrid Classical/Quantum Execution, 2017, https://www.dwavequantum.com/media/jhlpvult/partitioning_qubos_for_quantum_acceleration-2.pdf

[19] Y. Ye, Gset benchmark, https://web.stanford.edu/~yyye/yyye/Gset/

[20] Reference Workflows – Ocean Documentation 8.2.0 documentation, https://docs.ocean.dwavesys.com/en/latest/docs_hybrid/reference/reference.html

[21] D. J. Earl, and M. W. Deem, Parallel tempering: Theory, applications, and new perspectives, Phys. Chem. Chem. Phys. **7**, 3910 (2005), publisher: Royal Society of Chemistry.

[22] The coupling matrix J equivalently describes a weighted graph with spins as nodes.

[23] F. Barahona, On the computational complexity of Ising spin glass models, J. Phys. A: Math. Gen. **15**, 3241 (1982).

[24] S. Istrail, in *Proceedings of the Thirty-Second Annual ACM Symposium on Theory of Computing*, STOC '00 (Association for Computing Machinery, New York, NY, USA, 2000), p. 87.

[25] A. Lucas, Ising formulations of many NP problems, Front. Phys. **2**, 5 (2014).

[26] A route such that each city is visited exactly once.

[27] M. Ayodele, in *Evolutionary Computation in Combinatorial Optimization*, Lecture Notes in Computer Science, edited by L. Pérez Cáceres and S. Verel (Springer International Publishing, Cham, 2022), p. 159.

[28] A probability distribution over the joint space $X \times Y$ satisfying $\int_X \gamma\, dx = \nu$ and $\int_Y \gamma\, dy = \mu$.

[29] A. Durmus, É. Moulines, and M. Pereyra, Efficient Bayesian computation by proximal Markov chain Monte Carlo: When Langevin meets Moreau, SIAM J. Imaging Sci. **11**, 473 (2018).

[30] A. S. Dalalyan, and A. Karagulyan, User-friendly guarantees for the Langevin Monte Carlo with inaccurate gradient, Stoch. Process. Their Appl. **129**, 5278 (2019).

[31] L. Li, and Y. Wang, A sharp uniform-in-time error estimate for stochastic gradient Langevin dynamics, ArXiv:2207.09304.

[32] N. Shiraishi, K. Funo, and K. Saito, Speed limit for classical stochastic processes, Phys. Rev. Lett. **121**, 070601 (2018), publisher: American Physical Society.

[33] N. Shiraishi, Wasserstein distance in speed limit inequalities for Markov jump processes, J. Stat. Mech.: Theory Exp. **2024**, 074003 (2024), publisher: IOP Publishing.

[34] A. Dechant, Minimum entropy production, detailed balance and Wasserstein distance for continuous-time Markov processes, J. Phys. A: Math. Theor. **55**, 094001 (2022), publisher: IOP Publishing.

[35] M. Nakazato, and S. Ito, Geometrical aspects of entropy production in stochastic thermodynamics based on Wasserstein distance, Phys. Rev. Res. **3**, 043093 (2021), publisher: American Physical Society.

[36] A. Dechant, and Y. Sakurai, Thermodynamic interpretation of Wasserstein distance, ArXiv:1912.08405.

[37] S. Boyd, N. Parikh, E. Chu, B. Peleato, and J. Eckstein, Distributed optimization and statistical learning via the alternating direction method of multipliers, Foundations and TrendsâŞĞin Machine Learning **3**, 1 (2011), publisher: Now Publishers, Inc..

[38] X. Yi, C. Caramanis, and S. Sanghavi, in *Proceedings of the 31st International Conference on Machine Learning* (PMLR, Beijing, China, 2014), p. 613, ISSN: 1938-7228.
[39] P. Jain, P. Netrapalli, and S. Sanghavi, in *Proceedings of the forty-fifth annual ACM symposium on Theory of Computing*, STOC '13 (Association for Computing Machinery, New York, NY, USA, 2013), p. 665.
[40] A. Sokal, in *Functional Integration: Basics and Applications*, edited by C. DeWitt-Morette, P. Cartier, and A. Folacci (Springer US, Boston, MA, 1997), p. 131.
[41] M. X. Burns, Q. Hou, and M. Huang, in *The Thirteenth International Conference on Learning Representations* (Curran Associates Inc, Singapore, Singapore, 2024).
[42] A. T. Ogielski, Dynamics of three-dimensional Ising spin glasses in thermal equilibrium, Phys. Rev. B **32**, 7384 (1985).
[43] Y. Fang, S. Feng, K.-M. Tam, Z. Yun, J. Moreno, J. Ramanujam, and M. Jarrell, Parallel tempering simulation of the three-dimensional Edwards–Anderson model with compact asynchronous multispin coding on GPU, Comput. Phys. Commun. **185**, 2467 (2014).
[44] M. E. J. Newman, and G. T. Barkema, *Monte Carlo Methods in Statistical Physics* (Oxford University Press, Oxford, New York, 1999).
[45] R. Santos, L. Buffoni, and Y. Omar, Enhancing quantum annealing via entanglement distribution, ArXiv:2212.02465.
[46] K. Tatsumura, M. Yamasaki, and H. Goto, Scaling out Ising machines using a multi-chip architecture for simulated bifurcation, Nat. Electron. **4**, 208 (2021).
[47] S. R. Sundara Raman, L. K. John, and J. P. Kulkarni, in *2024 IEEE International Symposium on High-Performance Computer Architecture (HPCA)* (IEEE, Edinburgh, UK, 2024), p. 719, ISSN: 2378-203X.
[48] J. Raymond, R. Stevanovic, W. Bernoudy, K. Boothby, C. C. McGeoch, A. J. Berkley, P. Farré, J. Pasvolsky, and A. D. King, Hybrid quantum annealing for larger-than-QPU lattice-structured problems, ACM Trans. Quantum Comput. **4**, 1 (2023).
[49] A. Sharma, M. Burns, and M. C. Huang, in *26th International Conference on Theory and Applications of Satisfiability Testing (SAT 2023)* (Schloss Dagstuhl–Leibniz-Zentrum für Informatik, Alghero, Italy, 2023), p. 25.
[50] F. Böhm, T. V. Vaerenbergh, G. Verschaffelt, and G. Van der Sande, Order-of-magnitude differences in computational performance of analog Ising machines induced by the choice of nonlinearity, Commun. Phys. **4**, 1 (2021).
[51] S. Chewi, J. Niles-Weed, and P. Rigollet, *Statistical Optimal Transport: École d'Été de Probabilités de Saint-Flour XLIX – 2019*, Lecture Notes in Mathematics (Springer Nature Switzerland, Cham, 2025), Vol. 2364.
[52] Z. Ding, and Q. Li, Langevin Monte Carlo: Random coordinate descent and variance reduction, J. Mach. Learn. Res. **22**, 1 (2021).
[53] X. Cheng, N. S. Chatterji, Y. Abbasi-Yadkori, P. L. Bartlett, and M. I. Jordan, Sharp convergence rates for Langevin dynamics in the nonconvex setting, ArXiv:1805.01648.
[54] X. S. Liang, Information flow and causality as rigorous notions *ab initio*, Phys. Rev. E **94**, 052201 (2016).
[55] C. Villani, *Topics in Optimal Transportation*, UK ed. (American Mathematical Society, Providence, RI, 2003).
[56] S. Chewi, M. A. Erdogdu, M. Li, R. Shen, and S. Zhang, in *Proceedings of Thirty Fifth Conference on Learning Theory* (PMLR, London, UK, 2022), p. 1, ISSN: 2640-3498.
[57] S. Vempala, and A. Wibisono, in *Advances in Neural Information Processing Systems* (Curran Associates, Inc., Vancouver, Canada, 2019), Vol. 32.
[58] T.-S. Chiang, C.-R. Hwang, and S. J. Sheu, Diffusion for global optimization in $\mathbb{R}^n$, SIAM J. Control Optim. **25**, 737 (1987), publisher: Society for Industrial and Applied Mathematics.
[59] D. Panchenko, The Sherrington-Kirkpatrick model: An overview, J. Stat. Phys. **149**, 362 (2012).
[60] We define the Hamming distance $d_H(x,y) = 0$ if $\text{sign}(x) = \text{sign}(y)$ and 1 otherwise..
[61] P. Erdos, and A. Renyi, On the evolution of random graphs, Publ. Math. Inst. Hungary. Acad. Sci. **5**, 17 (1960).
[62] R. Albert, and A.-L. Barabási, Statistical mechanics of complex networks, Rev. Mod. Phys. **74**, 47 (2002), publisher: American Physical Society.
[63] B. Giridhar, M. Cieslak, D. Duggal, R. Dreslinski, H. M. Chen, R. Patti, B. Hold, C. Chakrabarti, T. Mudge, and D. Blaauw, in *Proceedings of the International Conference on High Performance Computing, Networking, Storage and Analysis* (ACM, Denver Colorado, 2013), p. 1.
[64] M. O'Connor, N. Chatterjee, D. Lee, J. Wilson, A. Agrawal, S. W. Keckler, and W. J. Dally, in *2017 50th Annual IEEE/ACM International Symposium on Microarchitecture (MICRO)* (IEEE, Boston, MA US, 2017), p. 41, ISSN: 2379-3155.
[65] K. Kim, and M.-j. Park, in *2024 IEEE International Memory Workshop (IMW)* (2024), p. 1, ISSN: 2573-7503.
[66] Micron Technology, Inc., RAM Memory Speeds & Compatibility, 2025, https://www.crucial.com/support/memory-speeds-compatability
[67] Micron Technology, Inc., DDR5 RAM: Everything You Need to Know, 2025, https://www.crucial.com/articles/about-memory/everything-about-ddr5-ram
[68] F. Ma, and J.-K. Hao, A multiple search operator heuristic for the max-k-cut problem, Ann. Oper. Res. **248**, 365 (2017).
[69] M. X. Burns, Parallel-Ising-Analysis, 2025, https://github.com/mxburns2022/Parallel-Ising-Analysis
[70] R. S. Liptser, and A. N. Shiryaev, *Statistics of Random Processes* (Springer, Berlin, Heidelberg, 2001),
[71] R. Flamary *et al.*, POT: Python Optimal Transport, J. Mach. Learn. Res. **22**, 1 (2021).