



Quantum Annealing via Path-Integral Monte Carlo With Data Augmentation


Jianchang Hu & Yazhen Wang


To cite this article: Jianchang Hu & Yazhen Wang (2021) Quantum Annealing via Path-Integral Monte Carlo With Data Augmentation, Journal of Computational and Graphical Statistics, 30:2, 284-296, DOI: [10.1080/10618600.2020.1814787](https://doi.org/10.1080/10618600.2020.1814787)


To link to this article: <https://doi.org/10.1080/10618600.2020.1814787>


 View supplementary material [↗](#)

 Published online: 09 Oct 2020.

 Submit your article to this journal [↗](#)

 Article views: 106

 View related articles [↗](#)

 View Crossmark data [↗](#)



Quantum Annealing via Path-Integral Monte Carlo With Data Augmentation

Jianchang Hu^a and Yazhen Wang^b

^aDepartment of Biostatistics, Yale University, New Haven, CT; ^bDepartment of Statistics, University of Wisconsin-Madison, Madison, WI

ABSTRACT

This article considers quantum annealing in the Ising framework for solving combinatorial optimization problems. The path-integral Monte Carlo simulation approach is often used to approximate quantum annealing and implement the approximation by classical computers, which refers to simulated quantum annealing (SQA). In this article, we introduce a data augmentation scheme into SQA and develop a new algorithm for its implementation. The proposed algorithm reveals new insights on the sampling behaviors in SQA. Theoretical analyses are established to justify the algorithm, and numerical studies are conducted to check its performance and to confirm the theoretical findings. Supplementary materials for this article are available online.

ARTICLE HISTORY

Received August 2018
Revised August 2020

KEYWORDS

Combinatorial optimization; Hamming distance; Parallel Ising model; Simulated quantum annealing; Strong ergodicity

1. Introduction

Combinatorial optimization plays an important role in many scientific studies. Examples include travel salesman's problem, task scheduling and system design, image analysis, machine learning, and portfolio selection. See Kirkpatrick, Gelatt, and Vecchi (1983), Geman and Geman (1987), Winkler (2012), and Wang, Wu, and Zou (2016) for more details about combinatorial optimization and its applications. For a typical combinatorial optimization problem, its search space often exponentially increases in its size or scale, and thus the problem can be NP-hard. As a result, deterministic approaches to solve the problem require computing resources with exponential growth in the problem size or scale, and hence, it is prohibitive to attack the combinatorial optimization problem by any deterministic exhaustive search algorithm in general. As a popular feasible alternative, stochastic search algorithms like annealing methods are widely employed to solve combinatorial optimization problems. One such well-known method is simulated annealing (SA) introduced by Kirkpatrick, Gelatt, and Vecchi (1983). In the annealing framework, the objective function of a given optimization problem is cast as the energy of a physical system, and Markov chain Monte Carlo (MCMC) simulations such as the Metropolis–Hastings algorithm are used to explore the large search space probabilistically. An artificial temperature parameter is introduced into the MCMC simulations so that we may probabilistically drive the system to its lowest possible energy state by decreasing the temperature slowly in the MCMC simulations, and the corresponding state of the system renders a solution to the original combinatorial optimization problem. The lowest energy states are called ground states.

Quantum annealing (QA) is the quantum analog of classical annealing (Kadowaki and Nishimori 1998). While thermodynamics is the physical driven force behind SA, QA replaces it by

quantum dynamics and uses quantum fluctuations to drive a quantum physical system to its lowest possible energy states. Specifically, the procedure starts with an easy initial quantum system prepared in its lowest energy state (or a ground state), and then gradually moves the system toward the target system whose energy matches the objective function of the given optimization problem. According to the quantum adiabatic theorem (Farhi et al. 2000, 2001), as the quantum system slowly evolves, it tends to stay in a ground state. At the end of the QA procedure, if the quantum system stays in a ground state, the energy and the state of the system provide a solution to the given optimization problem. That is, with certain probability QA can solve combinatorial optimization problems. Similar to SA, in practice, we may repeatedly run the QA procedure many times to find solutions of a given optimization problem.

QA is considered as adiabatic quantum computing, where special purpose quantum computers such as D-Wave analog quantum computing devices are built to physically implement QA for solving some combinatorial optimization problems. Like SA in the classical case, simulated quantum annealing (SQA) has been proposed in the quantum scenario. It combines path-integral formulation and MCMC simulations to develop algorithms for approximate implementations of QA on classical computers. SQA has been employed to facilitate the study and understanding of QA and its implementation such as the analysis of quantum performance of D-Wave machines. See Boixo et al. (2014) and Wang, Wu, and Zou (2016) for more details.

SQA algorithms suffer some drawbacks such as extremely time consuming and the lack of good understanding and clear interpretation. This article considers QA for the Ising model. We introduce a data augmentation scheme into SQA to develop a new SQA algorithm. To the best of our knowledge, this is

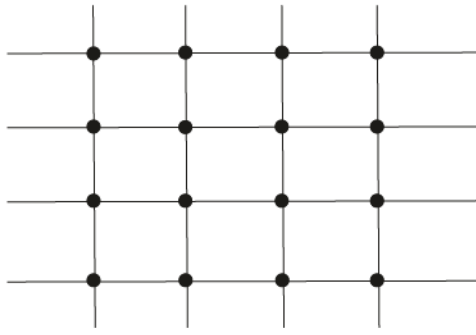


Figure 1. Illustration of a lattice structure as a simple graph.

the first time that SQA is viewed and investigated from the perspective of data augmentation. Our study based on the data augmentation angle reveals new insights that the Ising system involved in SQA essentially behaves like parallel classical SA systems with controlled Hamming distance between each two neighboring SA systems. We establish the strong ergodic theory for the algorithm under an appropriate condition on annealing schedules. The analysis may shed new light on the performance of SQA under various annealing schedules.

The rest of the article is organized as follows. Section 2 provides a brief review on classical and quantum annealing for the Ising model. Section 3 introduces the proposed new SQA algorithm based on data augmentation. Section 4 presents theoretical results for the algorithm. Numerical studies are conducted in Section 5 to evaluate the performances of SQA algorithms and validate our theoretical analysis. Conclusion and discussion are featured in Section 6. All proofs are relegated in the appendix.

2. Classical and Quantum Annealing With the Ising Model

2.1. The Ising Model and Simulated Annealing (SA)

The Ising model can be characterized by a graph \mathcal{G} with \mathcal{V} and \mathcal{E} being the sets of sites and edges, respectively. Each site represents a random variable taking values in $\{+1, -1\}$, and each edge indicates the interaction (or coupling) between the variables on the two sites linked by the edge. For a lattice with d sites, a configuration or state $s = (s_1, s_2, \dots, s_d)$ is a d -dimensional vector with each site variable being $s_i = \pm 1$. Figure 1 shows an example of a lattice structure as a simple graph for the Ising model. The Hamiltonian of the classical Ising model is given by

$$H_I^c(s) = - \sum_{(i,j) \in \mathcal{E}} J_{ij} s_i s_j - \sum_{i \in \mathcal{V}} h_i s_i, \quad (1)$$

where J_{ij} gives the strength of the interaction between sites i and j associated with edge (i, j) in graph \mathcal{G} , and h_i is the strength of the external local fields imposed on site i . A set of fixed values $\{J_{ij}, h_i\}$ is referred to as one instance of the Ising model. For simplicity, we consider no local fields and set $h_i = 0$ for the rest of this article. For a given configuration s , the energy of the Ising system is equal to $H_I^c(s)$. According to Boltzmann's law, the probability of a given configuration s can be described by the Boltzmann (or

Gibbs) distribution

$$P_\beta(s) = \frac{e^{-\beta H_I^c(s)}}{Z_\beta}, \quad Z_\beta = \sum_s e^{-\beta H_I^c(s)}, \quad (2)$$

where $\beta = \frac{1}{k_B T}$, and k_B is a generic physical constant called the Boltzmann constant, T is the absolute temperature of the system, and the normalization constant Z_β is called the partition function.

When a combinatorial optimization problem is represented by the Ising model, a set of $\{J_{ij}\}$ is specified, and the goal is to find a configuration s^* which minimizes the Hamiltonian $H_I^c(s)$ over all s . The configuration s^* is often referred to as a ground state of the Ising model. Finding a ground state is a hard computational problem, because the search space has 2^d configurations, an exponential increase in the system size d . One usual approach is to consider SA with a decreasing temperature $T = T(t)$ as a function of evolution time t . The initial temperature is set high to induce thermal fluctuations for exploring the large search space, the SA process samples configurations using MCMC simulations like the Metropolis algorithm, and the MCMC simulations lead the system to concentrate more and more frequently at the thermal equilibrium. As the temperature decreases slowly with typical schedule $T(t) \propto 1/\log t$, SA tends to drive the system to a ground state at the end of the annealing process, which gives a solution to the optimization problem. See Geman and Geman (1987), Hajek (1988), and Winkler (2012) for more detailed discussions on SA.

2.2. The Quantum Ising Model

A quantum system is described by its quantum state and the dynamic evolution of the state. The quantum state is often characterized by a unit vector in a complex vector space, and the dynamic evolution of the state is governed by a Hermitian matrix via the so-called Schrödinger equation. To introduce the quantum Ising model, we use the same graphical structure as in the classical Ising model given by (1) to define a quantum Ising Hamiltonian. Suppose that the graph \mathcal{G} has d sites. Each site now represents a quantum spin with possible states $|\uparrow\rangle$ and $|\downarrow\rangle$ for symbolizing spin up and spin down, respectively, where we use the customary Dirac notation $|\cdot\rangle$ in the quantum literature to denote the state. Mathematically the states $|\uparrow\rangle$ and $|\downarrow\rangle$ can be represented by the following unit vectors,

$$|\uparrow\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad \text{and} \quad |\downarrow\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}.$$

To specify a quantum Ising model we need to define its quantum Hamiltonian. As the quantum Ising model is based on matrices and vectors with dimensionality equal to 2^d , we define

$$I_j = \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}, \quad \sigma_j^x = \begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \\ \sigma_j^z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix}, \quad j = 1, \dots, d, \quad (3)$$

where σ_j^x and σ_j^z are called Pauli matrices in x and z axes, respectively. Substituting σ_j^z for s_j in the classical Ising Hamiltonian H_I^c defined by (1) (with $h_i = 0$), we obtain the following quantum

Hamiltonian of the quantum Ising model associated with the graph \mathcal{G}

$$\mathbf{H}_I^q = - \sum_{(i,j) \in \mathcal{E}} J_{ij} \sigma_i^z \sigma_j^z, \quad (4)$$

where J_{ij} is the interaction between sites i and j associated with edge (i, j) in the graph \mathcal{G} . Here, we use the convention in the quantum literature that $\sigma_i^z \sigma_j^z$ denotes the tensor product of σ_i^z and σ_j^z along with identity matrices in such a way that

$$\sigma_i^z \sigma_j^z \equiv I_1 \otimes \cdots \otimes I_{i-1} \otimes \sigma_i^z \otimes I_{i+1} \otimes \cdots \otimes I_{j-1} \otimes \sigma_j^z \otimes I_{j+1} \otimes \cdots \otimes I_d.$$

Consequently, each term in (4) is a diagonal matrix of size 2^d , and \mathbf{H}_I^q is also a diagonal matrix, with all diagonal elements equal to the 2^d values of the classical Hamiltonian \mathbf{H}_I^c in (1) (with $h_i = 0$). The energies of the quantum system are equal to the eigenvalues of the quantum Hamiltonian, with quantum states given by the corresponding eigenvectors. We refer the 2^d eigenvectors to quantum configurations, and the lowest energy states to its ground states. Thus, minimizing the classical Hamiltonian \mathbf{H}_I^c over the 2^d configurations is equivalent to finding the lowest energy of the quantum Ising model. See Nielsen and Chuang (2010), Wang (2012), Wang, Wu, and Zou (2016), and Wang and Song (2020) for more details.

2.3. Quantum Annealing (QA)

To describe QA for solving the combinatorial optimization problem, we need to introduce a magnetic field orthogonal to the Ising axis and obtain the following Hamiltonian of the quantum Ising model in the transverse field (Kadowaki and Nishimori 1998),

$$\mathbf{H}_\Gamma = - \sum_{(i,j)} J_{ij} \sigma_i^z \sigma_j^z - \Gamma \sum_i \sigma_i^x, \quad (5)$$

where σ_i^x stands for the following tensor product of d matrices of size 2,

$$\sigma_i^x \equiv I_1 \otimes \cdots \otimes I_{i-1} \otimes \sigma_i^x \otimes I_{i+1} \otimes \cdots \otimes I_d,$$

and Pauli matrix σ_i^x in the tensor product is defined in (3).

The two terms of \mathbf{H}_Γ in (5) are noncommutable matrices of size 2^d , and represent the potential and kinetic energies. The second term $-\sum_{i=1} \sigma_i^x$ is a Hamiltonian (or Hermitian matrix) with explicit expressions for its smallest eigenvalue and the corresponding eigenvector, and the quantum system governed by the Hamiltonian $-\sum_{i=1} \sigma_i^x$ can be easily prepared in its ground state. The nonnegative scalar Γ controls the strength of the transverse field. The QA procedure is described as follows. By decreasing Γ from a high level to zero gradually, we engineer the quantum system to slowly evolve from \mathbf{H}_Γ toward \mathbf{H}_I^q . According to the adiabatic quantum theorem, as the quantum system is initially started in its ground state, during the Hamiltonian evolution the system tends to stay in the ground states of the instantaneous Hamiltonian via quantum tunneling (Farhi et al. 2000, 2001). Therefore, at the end of the QA evolution, if the system stays in its ground state, we measure the system energy to render a solution to the optimization problem. That is, like SA, each run of QA can yield a solution to the optimization problem with certain probability, and running QA many times enables us to solve the optimization problem. See Wang, Wu, and Zou (2016) for more details.

2.4. Simulated Quantum Annealing (SQA)

Different approaches have been developed to approximately implement QA by various MCMC based simulations on classical computers (see Morita and Nishimori 2008; Wang, Wu, and Zou 2016 and the reference therein for more information). One popular SQA algorithm is the so-called SQA-PI algorithm (also known as the PIQA algorithm) introduced by Martoňák, Santoro, and Tosatti (2002). The main idea behind the SQA-PI algorithm is the path-integral formulation with the Trotter formula. Specifically, to derive a path-integral representation for the transverse field quantum Ising model (5), we introduce the following notations,

$$\mathbf{H}_\Gamma = \mathbf{H}_I^q + \mathbf{K}, \quad \mathbf{H}_I^q = - \sum_{(i,j)} J_{ij} \sigma_i^z \sigma_j^z, \quad \mathbf{K} = -\Gamma \sum_i \sigma_i^x,$$

where terms \mathbf{H}_I^q and \mathbf{K} represent the noncommutable potential and kinetic energies. The Boltzmann law of the transverse field quantum Ising model is given by $e^{-\beta \mathbf{H}_\Gamma} / Z_\beta$, where Z_β is the partition function defined as follows,

$$\begin{aligned} Z_\beta &= \text{tr}(e^{-\beta \mathbf{H}_\Gamma}) = \text{tr}(e^{-\beta(\mathbf{H}_I^q + \mathbf{K})}) = \lim_{\tau \rightarrow \infty} \text{tr} \left[(e^{-\frac{\beta}{\tau} \mathbf{H}_I^q} e^{-\frac{\beta}{\tau} \mathbf{K}})^\tau \right] \\ &= \lim_{\tau \rightarrow \infty} \sum_s \langle s | (e^{-\frac{\beta}{\tau} \mathbf{H}_I^q} e^{-\frac{\beta}{\tau} \mathbf{K}})^\tau | s \rangle, \end{aligned} \quad (6)$$

and the third equality follows from the Trotter breakup formula, which may be stated as that for Hermitian matrices A_1 and A_2 ,

$$\exp(A_1 + A_2) = \lim_{\tau \rightarrow \infty} [\exp(A_1/\tau) \exp(A_2/\tau)]^\tau.$$

On the right-hand side of (6), $s = \{s_i, i = 1, \dots, d\}$, $s_i = \pm 1$, and the summation runs over all 2^d possible s . With some algebraic manipulations on the right-hand side of (6) based on quantum mechanics, we obtain an approximation Z_τ to Z_β whose error is proportional to the square of the Trotter breakup time $\Delta t = \beta/\tau$, where Z_τ has the following expression,

$$Z_\beta \approx Z_\tau = C \sum_{s^1} \cdots \sum_{s^\tau} e^{-\mathbf{H}_{d+1}/\tau T}, \quad (7)$$

for some constant C ,

$$\mathbf{H}_{d+1} = - \sum_{k=1}^{\tau} \left(\sum_{(i,j)} J_{ij} s_i^k s_j^k + J^\perp \sum_i s_i^k s_i^{k+1} \right), \quad (8)$$

and

$$J^\perp = \frac{\tau T}{2} \ln \coth \frac{\Gamma}{\tau T} > 0.$$

It turns out that Z_τ is the partition function of a classical $(d+1)$ -dimensional *anisotropic* Ising system at temperature τT , with couplings J_{ij} along the original d -dimensional slices (the same for all Trotter slices and independent of k), and J^\perp along the extra dimension (positive and same for all sites i). The system has a finite length τ , and also periodic boundary conditions have to be assumed along the extra dimension ($s_i^{\tau+1} = s_i^1$, $i = 1, \dots, d$). We call $s^k = \{s_i^k, i = 1, \dots, d\}$, $k = 1, \dots, \tau$, Trotter slices, and τ the number of Trotter slices. Figure 2 illustrates a classical anisotropic Ising system with 3 Trotter slices.

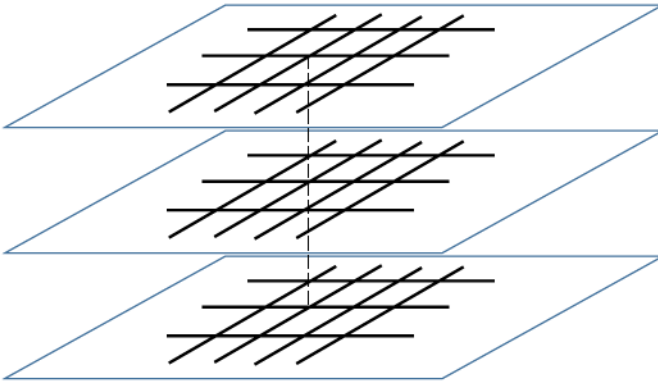


Figure 2. Structure of a classical anisotropic Ising system with three Trotter slices.

With the path-integral representation (7) and (8), we can approximately implement QA by conducting MCMC simulations from the Boltzmann distribution

$$p_{\text{SQA}}(s_1, \dots, s_\tau) = \frac{1}{Z_\tau} e^{-H_{d+1}/\tau T},$$

with some annealing schedule $\Gamma = \Gamma(t)$, which changes with time t as it evolves.

One common approach to carrying out this sampling process is to adopt a standard Metropolis algorithm with both local and global moves. To be specific, in each sweep, the local move is first performed where it attempts individual flips site by site in all Trotter slices with the Metropolis acceptance rule. After the local moves, the algorithm implements the global move where it attempts to flip simultaneously all the replicas of the same site in all Trotter slices; see Martoňák, Santoro, and Tosatti (2002) and Wang, Wu, and Zou (2016) for more details on this implementation. From the sampling perspective, the inclusion of the local move is natural. However, the theoretical rationale of the global move needs further elaboration and explanation. SQA and its implementations on classical computers allow us to gain insights on QA related quantum behaviors, and the insights may help us to better understand and study QA and its implementation like the quantum performance of D-Wave machines.

3. Data Augmentation Algorithm for SQA

This section presents a data augmentation algorithm for SQA from the viewpoint of theoretical understanding, which is called the SQA-DA algorithm. The algorithm demonstrates that data augmentation may provide better understanding of SQA and QA.

3.1. The SQA-DA Algorithm

Now we describe the data augmentation algorithm as follows. First from (7), we have

$$p_{\text{SQA}}(s^1, \dots, s^\tau) \propto e^{-H_{d+1}/\tau T} \\ = \prod_{k=1}^{\tau} e^{\frac{1}{\tau T} (\sum_{(i,j)} J_{ij} s_i^k s_j^k + J^\perp \sum_i s_i^k s_i^{k+1})}$$

$$= \left(\prod_{k=1}^{\tau} e^{\frac{1}{\tau T} \sum_{(i,j)} J_{ij} s_i^k s_j^k} \right) e^{\frac{J^\perp}{\tau T} \sum_{k=1}^{\tau} \sum_i s_i^k s_i^{k+1}} \\ \propto \left(\prod_{k=1}^{\tau} p_{\text{SA}}(s^k; \frac{1}{\tau T}) \right) e^{-\frac{2J^\perp}{\tau T} \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1})} \\ \propto \left(\prod_{k=1}^{\tau} p_{\text{SA}}(s^k; \frac{1}{\tau T}) \right) \mathbf{P} \\ \left(\text{Exp} \left(\frac{2J^\perp}{\tau T} \right) > \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1}) \right), \quad (9)$$

where $p_{\text{SA}}(\cdot; \frac{1}{\tau T})$ stands for the Boltzmann distribution for SA with temperature T , $\Lambda(a, b)$ is the Hamming distance between vectors a and b , and $\text{Exp}(\lambda)$ denotes an exponential random variable with mean $1/\lambda$. Expressions in (9) lead to a data augmentation algorithm for SQA, which we refer to as SQA-DA. The algorithm generates samples from the desired joint distribution $p_{\text{SQA}}(s^1, \dots, s^\tau)$ by the following procedure,

Step 1: $y|s^1, \dots, s^\tau \sim \text{Exp} \left(\frac{2J^\perp}{\tau T} \right) \mathbf{I}_{(\Lambda(s), \infty)}(y)$,

Step 2: $s^1, \dots, s^\tau | y \sim \prod_{k=1}^{\tau} p_{\text{SA}}(s^k; \frac{1}{\tau T}) \mathbf{I}_{[0, y]}(\Lambda(s))$,

where $\Lambda(s) = \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1})$. More specifically, in each sweep, in the first step we calculate $\Lambda(s)$, which provides the total neighboring Hamming distance of the current configuration matrix $\mathbf{s} = (s^1, \dots, s^\tau)$, and then generate control variable y from the truncated exponential distribution with parameter $\frac{2J^\perp}{\tau T}$ and truncation at $\Lambda(s)$. In the second step, we conduct τ parallel sampling from the classical Ising model with temperature τT , and under the constraint that the updated configuration matrix \mathbf{s} must satisfy $\Lambda(s) < y$, where y is generated from the first step. There are a couple of ways to carry out the procedure. One straightforward approach is to propose and accept a new state by generating samples from τ independent Ising models and then performing an update if they satisfy the constraint. Another way is to incorporate the constraint into the proposal and then update according to the usual Metropolis rule.

3.2. Implication of the SQA-DA Algorithm

We provide some implication of the SQA-DA algorithm for the step-wise update and limiting cases.

3.2.1. The Step-Wise Update Case

First we would like to trace the effect of Γ on the anisotropic Ising system. Recall that in the Ising system defined in (8), Γ appears only in the definition of J^\perp . In the SQA-DA algorithm, J^\perp shows up only in $\lambda = \frac{2J^\perp}{\tau T}$, the rate of exponential distribution, and the magnitude of parameter λ would directly affect the augmented variable y which controls the total Hamming distance $\Lambda(s) = \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1})$ between neighboring slices. Hence, the strength of the transverse field Γ has an important impact on the Ising system solely through the control over the total Hamming distance between neighboring slices.

The control of the total neighboring Hamming distance can be considered as a regulation on the search range allowed for generating configuration slices from the Ising system in each

step. With a loose control, different slices can behave freely, which leads to searching for a ground state in a wide range of the state space. On the other hand, if the control is tight, different slices cannot vary from each other too much, thus each time the generated sample path from the system can explore only a small range of the state space. Combining it with the impact of the transverse field on the system, we may conclude that at each step the transverse field affects the Ising system by managing the searching range of the parallel Ising slices.

Moreover, in comparison with the SQA-PI algorithm, the proposed SQA-DA algorithm has more explicit and strict control on the total neighboring Hamming distance. The SQA-PI algorithm governs such distance indirectly through probability, and thus it is neither explicit nor strict. The difference, to certain extent, is similar to that between ridge regression and lasso where both regularization methods shrink the coefficients toward zero, but the ridge may hardly produce exact zero coefficients while lasso can yield exactly zero coefficients. It will be further confirmed by the established theory as well as the conducted numerical study later.

3.2.2. The Limiting Case

We have the following observations when transverse field parameter Γ approaches either 0 or $+\infty$.

1. When $\Gamma \rightarrow +\infty$, then $\lambda \rightarrow 0$. Therefore, the Hamming distance constraint plays very little role, and in this case all slices s^k for $k = 1, 2, \dots, \tau$ behave almost independently. This is close to what is happening at the initial stage of SQA with $\Gamma \in (0, +\infty)$ where at the initial stage we sample all slices independently from the Ising model to explore the state space as much as possible.
2. When $\Gamma \rightarrow 0$, then $\lambda \rightarrow +\infty$. In this case, the constraint on the Hamming distance in the Step 2 is extremely strong, which leads to all τ slices converging to the same configuration. This resembles the scenario at the end stage of SQA with $\Gamma \in (0, +\infty)$ where all slices essentially merge to the same configuration and thus effectively reduce to one slice.

4. Convergence Theory

This section presents theoretical convergence results for the proposed SQA-DA algorithm. We start with the constant transverse field case where Γ is assumed to be a fixed positive number. Because our ultimate goal is to study the asymptotic behavior of the inhomogeneous Markov chain associated with the annealing procedure, we then consider the case where Γ varies over time. Before showing any theoretical results we point out that our analyses are always conducted under the fixed d and τ situation, and hence, the state space of the Markov chain is finite.

4.1. Constant Transverse Field

We have the following result for the Markov chain associated with the SQA-DA algorithm.

Theorem 1. For fixed $T > 0$ and $\Gamma > 0$ (then $J^\perp < \infty$), the Markov chain associated with the configuration matrix s in the SQA-DA algorithm is (1) Harris recurrent; (2) geometrically

ergodic, meaning that there exists a function $M : S \rightarrow [0, \infty)$, where S is the state space, and a constant $\rho \in [0, 1)$ such that, for all $s \in S$ and all $n = 1, 2, \dots$,

$$\|P^n(s, \cdot) - p_{\text{SQA}}(\cdot)\| \leq M(s)\rho^n,$$

where $P^n(a, b)$ denotes the n -step transition probability of the Markov chain from state a to b .

Remark 1. One implication of the geometric ergodicity under the constant Γ is that during the annealing process the system does not need to stay at each Γ level for too long because under each fixed Γ , the Markov chain has a fast mixing rate. We often resort to some asymptotic convergence to justify MCMC and simulated annealing algorithms, and their asymptotic justifications typically rely on Markov chain limit theorems such as ergodic theorems (Hajek 1988; Morita and Nishimori 2008; Winkler 2012). Theorem 1 is in line with the standard approach to provide theoretical justifications for the SQA-DA algorithm as follows. It shows that the sequence generated from the SQA-DA algorithm quickly converges in distribution to the target equilibrium distribution, and consequently, after we run the algorithm long enough, the generated sequence should approximately follow the equilibrium distribution.

4.2. Time-Dependent Transverse Field

So far we have considered the asymptotic behavior of the Markov chain associated with the SQA-DA algorithm with constant Γ (and J^\perp). The annealing procedure requires to employ an annealing schedule changing with time, and thus we need to consider the case that Γ varies with time t . Strong ergodicity is introduced to study the asymptotic behaviors of the associated inhomogeneous Markov chain.

Definition 1. An inhomogeneous Markov chain is said to be strongly ergodic if the probability distribution of the Markov chain converges to a unique distribution irrespective of the initial distribution, namely,

$$\exists \mu, \forall t_0 \geq 0, \limsup_{t \rightarrow \infty} \|p(t_0, t) - \mu\| = 0, \quad (10)$$

where μ is a fixed distribution, $p(t_0, t)$ is the probability distribution of the chain at time t under the initial distribution p_0 at t_0 , and for two distributions ν_1 and ν_2 , $\|\nu_1 - \nu_2\|$ denotes the total variation of $\nu_1 - \nu_2$.

We need to impose some technical conditions for the theoretical analysis.

(C1) Assume that the transition probability of the Markov chain is of the following form,

$$G(s, \tilde{s}; t) = \begin{cases} N(s, \tilde{s})A(q(\tilde{s})/q(s)) & (s \neq \tilde{s}), \\ e^{-\frac{2J^\perp(t)}{\tau}((\Lambda(\tilde{s}) - \Lambda(s)) \vee 0)} & (s \neq \tilde{s}), \\ 1 - \sum_{\tilde{s} \neq s} G(s, \tilde{s}; t) & (s = \tilde{s}), \end{cases} \quad (11)$$

where $N(s, \tilde{s})$ is the generation probability, $A(u) = \min\{1, u\}$ is the usual acceptance function for the

Metropolis method, $a \vee b = \max\{a, b\}$,

$$q(\mathbf{s}) = \prod_{k=1}^{\tau} p_{SA}(s^k; \frac{1}{\tau T}), \text{ and } J^{\perp}(t) = \frac{\tau T}{2} \ln \coth \frac{\Gamma(t)}{\tau T}, \tag{12}$$

and $\Lambda(\mathbf{s}) = \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1})$ is the total neighboring Hamming distances of the configuration matrix \mathbf{s} .

(C2) The generation probability N in Condition (C1) is a positive and irreducible symmetric transition probability.

Let

$$\mathcal{N}_s = \{\tilde{\mathbf{s}} : N(\mathbf{s}, \tilde{\mathbf{s}}) > 0\},$$

$$\mathcal{S}_m = \{\mathbf{s} : \forall \tilde{\mathbf{s}} \in \mathcal{N}_s, \Lambda(\tilde{\mathbf{s}}) \leq \Lambda(\mathbf{s})\}, \tag{13}$$

and denote by R the maximum number of minimum steps needed to reach an arbitrary state $\mathbf{s} \in \mathcal{S}_m$ from any other state. The following theorem addresses the asymptotic behavior for the inhomogeneous Markov chain associated with the SQA-DA algorithm.

Theorem 2. Under Conditions (C1) and (C2), for fixed $T > 0$, the inhomogeneous Markov chain associated with \mathbf{s} in the SQA-DA algorithm is strongly ergodic and converges to the equilibrium state corresponding to the distribution $q(\mathbf{s})\mathbf{I}(\Lambda(\mathbf{s}) = 0)$, where $q(\mathbf{s})$ is defined in (12), if

$$\Gamma(t) \geq \tau T \tanh^{-1} \frac{1}{(t+2)^{1/d\tau R}}, \tag{14}$$

where d is the dimension of each slice, τ is the total number of Trotter slices, and R is defined right after (13).

Remark 2. To establish the strong ergodicity, one implicit requirement for the annealing schedule is the positivity, that is, $\Gamma(t)$ decreases to 0 from above, which is required to maintain the irreducibility of the Markov chain. As we discussed in Remark 1, the ergodicity established in Theorem 2 provides theoretical justifications for the SQA-DA algorithm with time varying $\Gamma(t)$.

From the theorem we immediately obtain a corollary regarding the role of the global move in the SQA-PI algorithm.

Corollary 1. Let $N_{\text{global}}(\mathbf{s}, \tilde{\mathbf{s}})$ be the proposal distribution of the global move in the SQA-PI algorithm. Then we have that

- (i) when the state space consists of all possible configuration matrices with τ slices, then $N_{\text{global}}(\mathbf{s}, \tilde{\mathbf{s}})$ is not irreducible, and thus for $\Gamma(t) > 0$, the corresponding transition kernel is not irreducible at any given time point t ;
- (ii) when the state space consists of all configuration matrices with τ identical slices, $N_{\text{global}}(\mathbf{s}, \tilde{\mathbf{s}})$ is irreducible and sampling according to the global move rule within such state space is equivalent to sampling from the corresponding classical Ising model.

Remark 3. Corollary 1 indicates that for $\Gamma(t) > 0$, the global move is not essential. On the other hand, for $\Gamma(t) = 0$, when SQA-DA reaches to the set of configuration matrices \mathbf{s} with $\Lambda(\mathbf{s}) = 0$, the global move may be used but it is effectively the same as sampling with a single slice.

5. Numerical Studies

We conducted numerical studies to check the performance of the proposed SQA-DA algorithm and compared it with the SQA-PI algorithm. In the following studies, we fixed temperature $T = 0.1$, took the number of slices in the configuration matrix to be $\tau = 30$, and utilized 15,000 sweeps for each annealing procedure. The sampling distribution used was

$$p(s^1, \dots, s^{\tau}; A(t), B(t)) \propto \prod_{k=1}^{\tau} e^{\frac{1}{\tau T} (B(t) \sum_{(i,j)} J_{ij} s_i^k s_j^k + J^{\perp}(t) \sum_i s_i^k s_i^{k+1})},$$

with

$$J^{\perp}(t) = \frac{\tau T}{2} \ln \coth \frac{A(t)}{\tau T} > 0.$$

We selected annealing schedules

$$B(t) = 5.2t^2 + 0.2t + 0.1, t \in [0, 1],$$

and three choices for $A(t), t \in [0, 1]$,

1. $A_1(t) = (8t^2 - 9.6t + 2.88)I\{0 \leq t \leq 0.6\}$,
2. $A_2(t) = (-3.6t + 2.88)I\{0 \leq t \leq 0.8\}$,
3. $A_3(t) = -2.88t + 2.88$,

where $I\{A\}$ denotes the indicator function of event A . The annealing schedules are plotted in Figure 3 along with the theoretical lower bound derived from Theorem 2 as a reference.

5.1. Success Probability

We considered the 1000 instances used in the study of quantum performance of D-Wave machine in Boixo et al. (2014) and Wang, Wu, and Zou (2016). The graph size is $d = 108$, with J_{ij} 's being randomly assigned values ± 1 . For each instance, the ground state success probability was estimated by the frequency of finding a ground state among the 1000 runs under each annealing schedule.

Figure 4 illustrates the histograms of the ground state success probabilities generated by the D-Wave machine and by SA for the classical Ising model (1) with no local fields ($h_i = 0$). The annealing schedule used for D-Wave is $A_1(t)$ along with $B(t)$, with annealing schedule $B(t)$ for SA (which needs only schedule $B(t)$). More details on the study of D-Wave, SQA-PI, and SA can be found in Boixo et al. (2014) and Wang, Wu, and Zou (2016).

Figures 5–7 display the histograms of the ground state success probability data generated from the SQA-PI and proposed SQA-DA algorithms. Here, the SQA-PI algorithm involves both local and global moves, while SQA-DA employs only the local move.

The results lead to the following observations. The similar bimodal shapes shown in Figures 5–7 for the SQA-PI algorithm bears some resemblance to that for the D-Wave data in Figure 4, as demonstrated in Boixo et al. (2014) and Wang, Wu, and Zou (2016). Recall that the effect of the transverse field on the Ising system is through its control over the total neighboring Hamming distance, and as we discussed in Section 3.2.2, when $A(t) \rightarrow 0$, all slices tend to merge into one slice. We argue that such merging leads to the bi-modal shapes displayed in the figures. First the SA result indicates that these 1000 instances

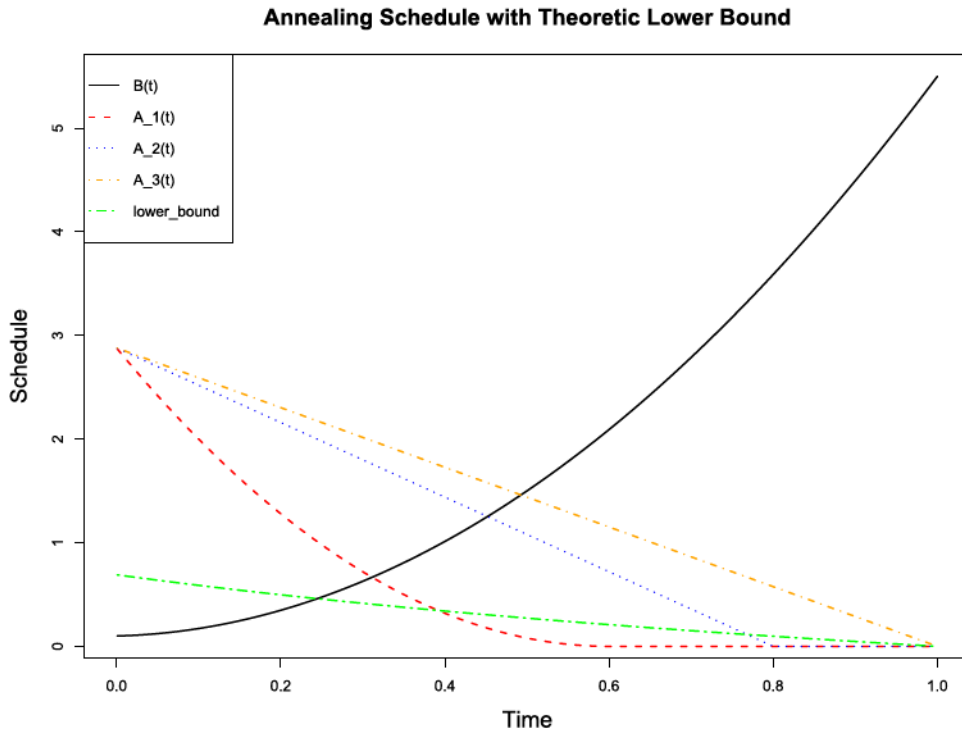


Figure 3. Annealing schedules used in the numerical studies along with the theoretical lower bound derived in Theorem 2.

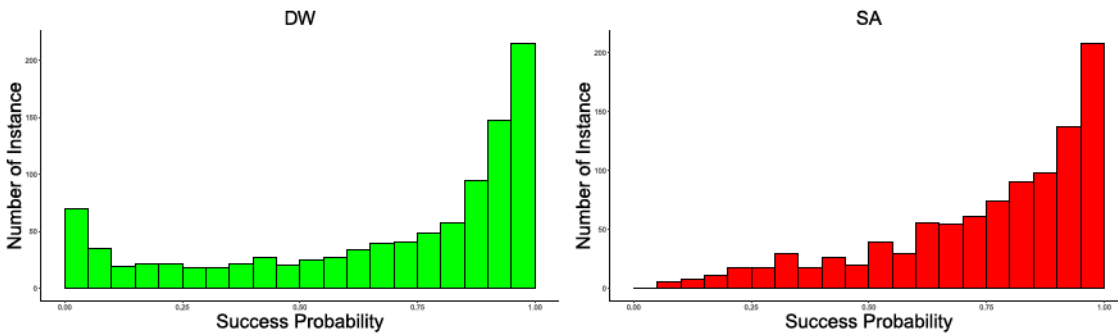


Figure 4. Histogram plots of ground state success probability data from D-Wave machine with annealing schedules $A_1(t)$ and $B(t)$, and from SA with schedule $B(t)$.

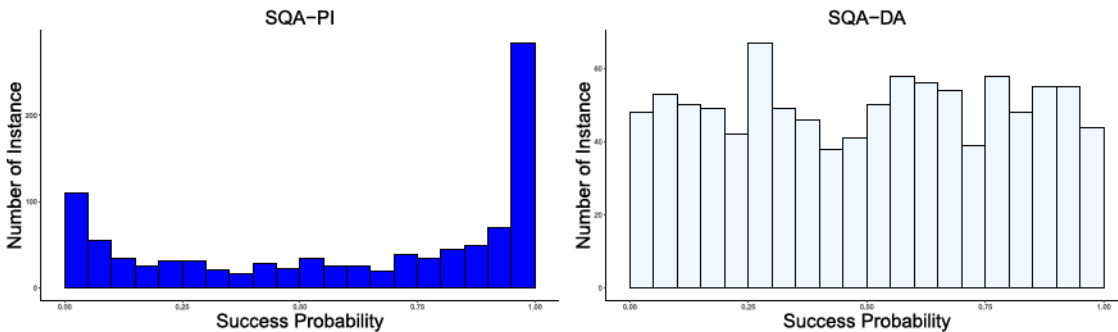


Figure 5. Histogram plots of ground state success probability data for the SQA-PI and SQA-DA algorithms with annealing schedules $A_1(t)$ and $B(t)$.

consist of both easy and hard problems. For easy problems, a majority of the parallel Ising slices have the tendency to be around a ground state, so when being pushed to merge together, they likely converge to a nearby ground state, which results in their success probabilities to cluster around 1. On the other hand, for hard problems, most slices in the system may not settle near a ground state. Reasons may include that ground states are sparsely distributed in the space, and thus it is very hard to

search for such ground states; or there is a high energy barrier around a ground state such that reaching to its neighborhood is difficult. Even a small portion of slices are close to a ground state, due to the stringent control over the total neighboring Hamming distance toward the end of the annealing process, they may be pulled away from the ground state to be merged together with other slices. All of these may yield the other mode around 0 in the histogram of the success probability data.

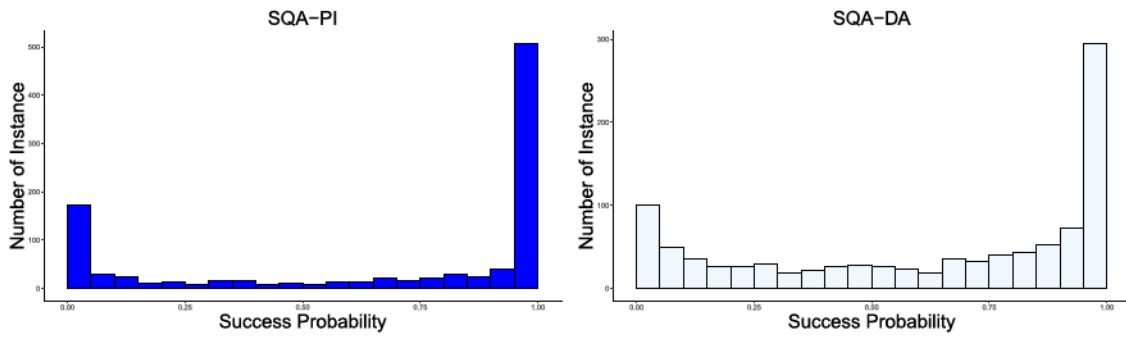


Figure 6. Histogram plots of ground state success probability data for the SQA-PI and SQA-DA algorithms with annealing schedules $A_2(t)$ and $B(t)$.

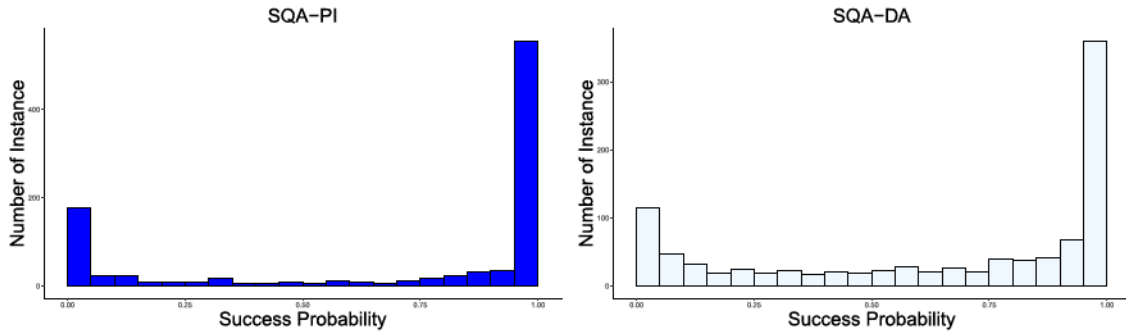


Figure 7. Histogram plots of ground state success probability data for the SQA-PI and SQA-DA algorithms with annealing schedules $A_3(t)$ and $B(t)$.

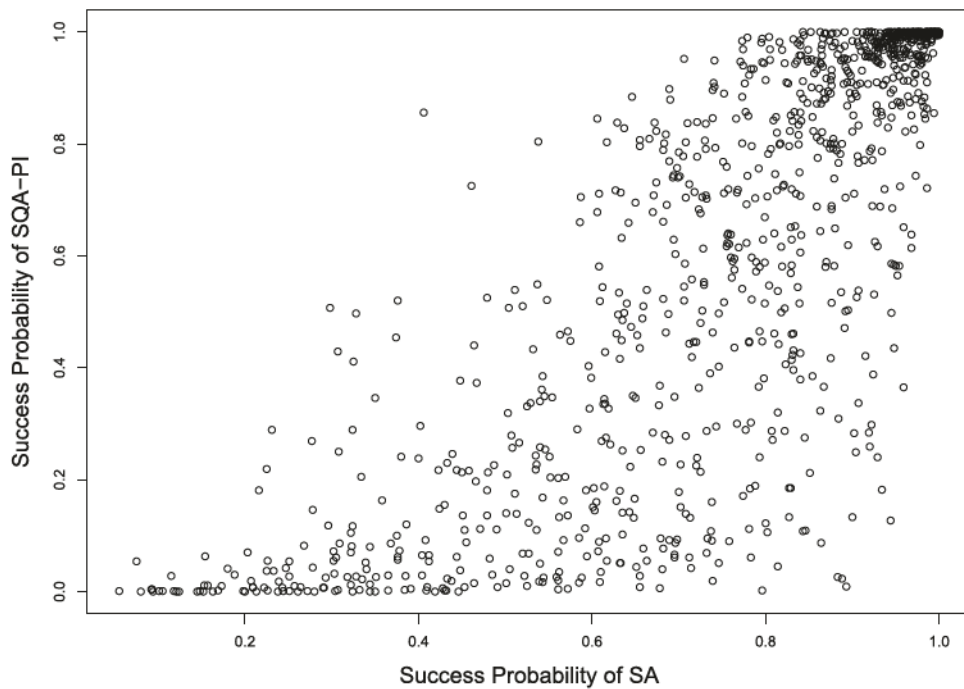


Figure 8. Instance-by-instance success probability comparison between SA and SQA-PI.

Figure 8 features the success probability scatterplot of SQA-PI against SA for the 1000 instances. The plot shows a concurrence that both SQA-PI and SA exhibit near-zero or near-one success probability for a large number of instances. It confirms that the performance of the SQA algorithms is decided by the majority of the slices in the system, especially when the problems are extremely easy or hard.

Both SQA-PI and SQA-DA algorithms exhibit bimodal shapes shown in Figures 5–7. However, the bimodal shapes

are more concentrated around endpoints 0 and 1 in Figures 6 and 7 than in Figure 5. This is especially true for the SQA-PI algorithm. The phenomenon may be explained as follows. As $A_2(t)$ and $A_3(t)$ decrease to 0 much slower than $A_1(t)$, the constraint on the total neighboring Hamming distance is weaker for the case of $A_2(t)$ and $A_3(t)$ than for the case of $A_1(t)$, and thus slices may have a higher chance to visit or escape from a ground state for the case of schedules $A_2(t)$ and $A_3(t)$ than for the case of schedule $A_1(t)$. Moreover, $A_2(t)$ and $A_3(t)$ stay

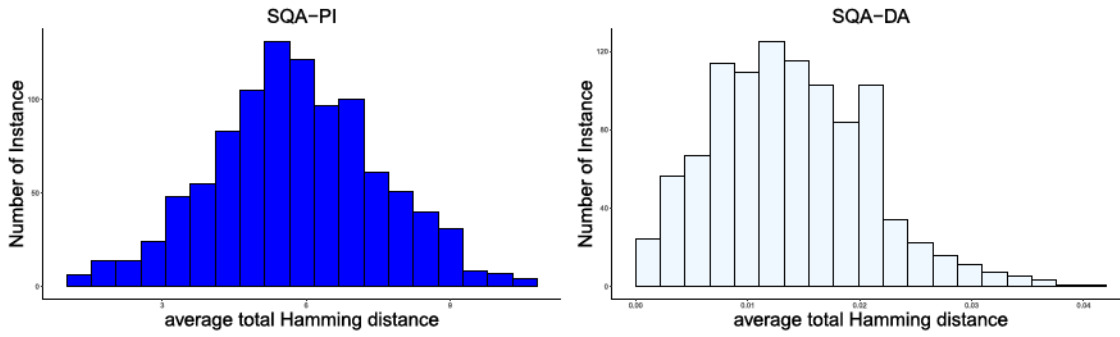


Figure 9. Histogram plots of average total neighboring Hamming distance for the SQA-PI and SQA-DA algorithms under annealing schedule $A_1(t)$.

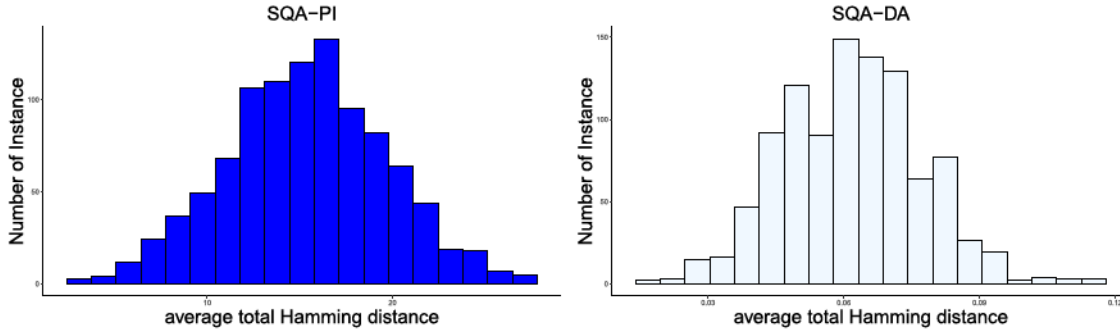


Figure 10. Histogram plots of average total neighboring Hamming distance for the SQA-PI and SQA-DA algorithms under annealing schedule $A_2(t)$.

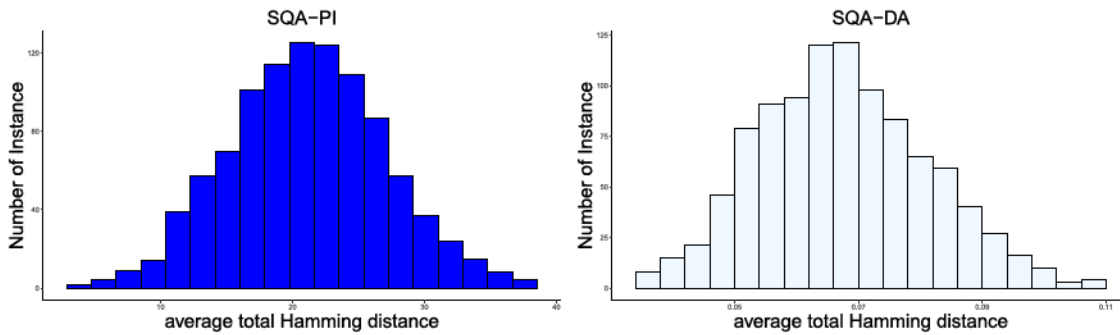


Figure 11. Histogram plots of average total neighboring Hamming distance for the SQA-PI and SQA-DA algorithms under annealing schedule $A_3(t)$.

at 0 in a much shorter time period than $A_1(t)$, so the time for allowing such a visit to or an escape from a ground state is longer for $A_2(t)$ and $A_3(t)$ than for $A_1(t)$. These may cause the major difference in the shapes between Figures 6 and 7 and Figure 5.

Furthermore, the shape difference in the success probability histograms for the SQA-DA algorithm under different schedules may be further explained by the implicit requirement pointed out in Remark 2. Schedule $A_1(t)$ decreases to zero very fast with 40% of time being 0, it does not meet the irreducibility requirement of the chain. Once $A_1(t) = 0$, the set of configuration matrices with all τ slices being the same becomes the absorbing set, and thus the Markov chain associated with such a configuration matrix is neither irreducible nor ergodic. From the algorithm perspective, if schedules quickly decrease to 0 in a relatively short period of time, all slices are driven to the same state without fully exploring the state space. Whenever $A(t) = 0$, no increase in total neighboring Hamming distance will be allowed, and therefore the algorithm stops prematurely.

5.2. Total Neighboring Hamming Distance

For each instance and each of the 1000 runs, we recorded the total neighboring Hamming distance $\Lambda(s)$ of the final configuration matrix s , and then computed the average total Hamming distance based on the 1000 runs for each instance. The histograms of all 1000 instances under each annealing schedule are displayed in Figures 9–11.

Figures 9–11 indicate that the distributions for the total neighboring Hamming distance of the final configuration matrix are approximately normal except for the SQA-DA algorithm under annealing schedule $A_1(t)$. The normality result may be due to the ergodicity of the Markov chains associated with the algorithms. On the other hand, as we have pointed out that the chain associated with the SQA-DA algorithm under schedule $A_1(t)$ is not ergodic, it is natural to expect the exception for the case of the SQA-DA algorithm under annealing schedule $A_1(t)$. Furthermore, we may observe from Figures 9–11 that the average total Hamming distance for the SQA-DA algorithm has negligible means in comparison with very large means of the average total Hamming distance for the SQA-PI algorithm. It

Table 1. Mean of the average total Hamming distance of final configuration matrix with standard deviation in parentheses.

	$A_1(t)$	$A_2(t)$	$A_3(t)$
SQA-PI	5.79(1.74)	15.36(4.33)	21.12(5.88)
SQA-DA	0.014(0.47)	0.062(0.78)	0.069(0.94)

may be attributed to the fact that as schedules $A_i(t)$ decay toward 0, the Hamming distance constraint introduced via the augmented variable tends to drive all slices to the same state. This is clearly explained by the SQA-DA algorithm description in Section 3 and the limiting behavior of the algorithm in Theorem 2 with equilibrium distribution on $I(\Lambda(s) = 0)$. In contrast, without such an explicit and strict Hamming distance constraint, the SQA-PI algorithm produces final configuration matrices whose slices differ at a substantial number of sites. This is further confirmed by the numerical evidence reported in Table 1.

Moreover, Table 1 suggests that the average total neighboring Hamming distance has an increasing mean for both algorithms as we change annealing schedule from $A_1(t)$ to $A_2(t)$ and then to $A_3(t)$. Again the findings may be explained as follows. As we move from $A_1(t)$ to $A_2(t)$ and then to $A_3(t)$, the schedules decrease to zero more slowly and then stay at 0 for a shorter time, thus the slices in the configuration matrix tend to slowly converge toward a single slice state, and the time allowed to do so is shortened. Consequently these may cause the increase in the total neighboring Hamming distance for the final configuration matrix.

5.3. Effect of Global Move

As we discussed early, the SQA-DA algorithm does not explicitly involve global moves, and thus its implementation has no

explicit steps to enforce the global moves for the studies conducted so far. We added extra steps in the analyses shown in this section to explicitly require the global moves in the SQA-DA algorithm and check the ground state success probability outputs produced by the SQA-DA algorithm under the cases of with and without global move. Figures 12 and 13 display the histograms of the output results for the SQA-DA algorithm with and without global moves.

From Figures 12 and 13, we can observe that the global moves do not have a large impact on the proposed SQA-DA algorithm, especially for annealing schedules with a substantial amount of time staying away from 0. However, for annealing schedule with a long time period of being at 0 such as $A_1(t)$, which is equal to zero for 40% of the total annealing time, there is an increasing trend toward the right endpoint in the histogram of Figure 12 corresponding to the case of with global move. Since a clear increasing pattern appears in Figure 4 for the SA case, the phenomenon may be explained as follows. Corollary 1 implies that once $A_1(t) = 0$ the Markov chain lands in the set of configuration matrices with all τ identical slices, and after then the global moves essentially make updates similar to the SA procedure within this set. Therefore, the SQA-DA algorithm exhibits some increasing pattern in its success probability histogram.

6. Conclusion and Discussion

We have considered solving combinatorial optimization problems by QA in the framework of the Ising model and investigated its implementation by SQA algorithms on classical computers. We introduced data augmentation to SQA and proposed a new SQA algorithm to approximately implement

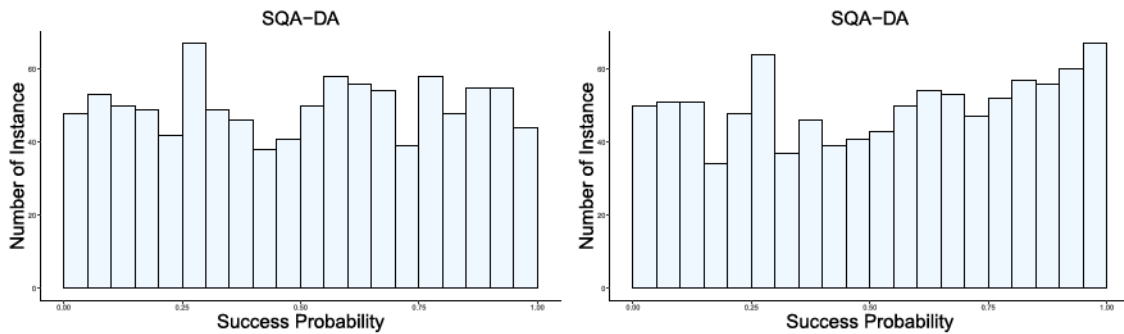


Figure 12. Histogram plots of ground state success probability data for the SQA-DA algorithm with and without global move under annealing schedule $A_1(t)$. The left and right panels correspond to histograms without and with global move, respectively.

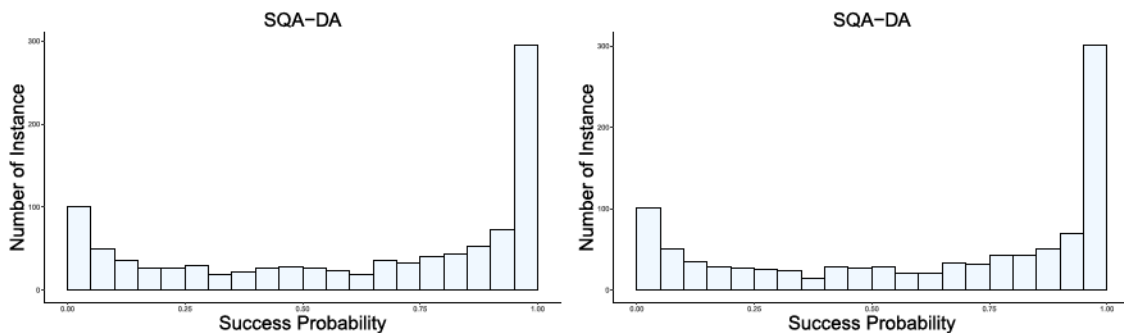


Figure 13. Histogram plots of ground state success probability data for the SQA-DA algorithm with and without global move under annealing schedule $A_2(t)$. The left and right panels correspond to histograms without and with global move, respectively.

QA on classical computers. Note that QA is considered as special-purpose quantum computing, and it is intractable to directly simulate quantum systems (or quantum computers) or implement QA on classical computers. SQA is often employed to gain insight about QA and investigate the performance of QA devices. Such studies play an important role in quantum information science particularly in the certification of quantum communication and computation devices like quantum computers.

Through the proposed SQA-DA algorithm, we have shown that sampling from the approximate target distribution for QA is essentially sampling from parallel classical Ising models with total neighboring Hamming distance appropriately controlled. The strong ergodicity for the proposed algorithm has been established under suitable conditions on annealing schedules. Numerical studies have been conducted to confirm theoretical results and check the performance of the proposed algorithm. In particular, our findings can provide new insights on the understanding of different types of moves (especially global move) under different annealing schedules in the SQA-PI algorithm as well as the proposed SQA-DA algorithm.

We would also like to point out that there is a possibility to incorporate cluster updating into SQA. For instance, a Swendsen–Wang type algorithm (Swendsen and Wang 1987) can be considered where besides the usual “bond” variables between sites within each slice, additional “link” variables at the same sites between neighboring slices can also be introduced. This new rule can update a cluster of sites in the configuration matrix at the same time, which may lead to speed-up for the sampling procedure compared with the usual site-by-site and slice-by-slice update in the existing SQA algorithms.

We leave some issues and problems for the future study. For example, our work offers a better understanding of different types of moves, and gives a first-step explanation on shape patterns of histograms for success probability data. There are many statistical issues in the study of QA and SQA. For example, a further study is needed to better understand QA and SQA in particular how their performances are related to the finite-time annealing, certain type of annealing schedules, or some other more fundamental unknown factors.

Appendix A. Proof of Theorem 1

Proof. For two configuration matrix \mathbf{s} and $\bar{\mathbf{s}}$, denote by $K(\mathbf{s}, \bar{\mathbf{s}}) > 0$ the irreducible transition probability associated with the equilibrium distribution $q(\mathbf{s}) = \prod_{k=1}^{\tau} p_{SA}(s^k; \frac{1}{\tau T})$. Then we have the transition kernel of the Markov chain associated with \mathbf{s} in the algorithm as

$$\begin{aligned} k(\bar{\mathbf{s}}|\mathbf{s}) &= \int f(\bar{\mathbf{s}}|y, \mathbf{s})f(y|\mathbf{s})dy \\ &= K(\mathbf{s}, \bar{\mathbf{s}}) \int \phi(y) \mathbf{I}_{(\sum_{k=1}^{\tau} \Lambda(\bar{s}^k, \bar{s}^{k+1}) - \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1}), \infty)}(y) dy \\ &= K(\mathbf{s}, \bar{\mathbf{s}}) e^{-\frac{2J^{\perp}}{\tau T} ((\sum_{k=1}^{\tau} \Lambda(\bar{s}^k, \bar{s}^{k+1}) - \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1})) \vee 0)} > 0, \end{aligned}$$

where $\phi(y)$ is the density for the exponential distribution with parameter $\frac{2J^{\perp}}{\tau T}$. It is easy to check that $p_{\text{SQA}}(\cdot)$ satisfies the detailed balance condition with this transition kernel,

$$p_{\text{SQA}}(\mathbf{s})k(\bar{\mathbf{s}}|\mathbf{s}) = p_{\text{SQA}}(\bar{\mathbf{s}})k(\mathbf{s}|\bar{\mathbf{s}}).$$

Therefore, $p_{\text{SQA}}(\cdot)$ is the equilibrium distribution. Since $p_{\text{SQA}}(\cdot)$ is positive on the finite state space, so the chain is recurrent and hence Harris recurrent.

The result of geometric ergodicity is a direct consequence of Theorem 11.2.1 in Winkler (2012) with a Harris irreducible and reversible transition kernel associated with $q(\mathbf{s})$.

Another way to prove the result is to consider the Geometric Ergodic Theorem (Theorem 15.0.1) in Meyn and Tweedie (2012). Let $\Lambda(\mathbf{s}) = \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1})$ be the drift function for the Markov chain. For each $\beta \in \mathbb{R}$, the sub-level set $\{\mathbf{s} : \Lambda(\mathbf{s}) \leq \beta\}$ is compact since the set only contain isolated points. In addition, we have

$$\begin{aligned} E[\Lambda(\mathbf{s}_{n+1})|\mathbf{s}_n = \mathbf{s}] &\leq E[y_{n+1}|\mathbf{s}_n = \mathbf{s}] \\ &= \Lambda(\mathbf{s})e^{-\lambda\Lambda(\mathbf{s})} + 1/\lambda e^{-\lambda\Lambda(\mathbf{s})} \\ &\leq e^{-\lambda} \Lambda(\mathbf{s}) + 1/\lambda, \end{aligned}$$

where $e^{-\lambda} \in [0, 1)$. By Lemma 15.2.8 of Meyn and Tweedie (2012), we can show that the geometric drift condition holds. Therefore, by geometric ergodic theorem (Theorem 15.0.1) in Meyn and Tweedie (2012), we establish that the chain is geometrically ergodic. \square

Appendix B. Proof of Theorem 2

We need to define weak ergodicity for proving the theorem.

Definition 2. An inhomogeneous Markov chain is weak ergodic if the probability distribution of the chain becomes independent of the initial conditions after a sufficiently long time, namely,

$$\forall t_0 > 0, \limsup_{t \rightarrow \infty} \|p(t_0, t) - p'(t_0, t)\| = 0, \quad (\text{B.1})$$

where $p(t_0, t)$ and $p'(t_0, t)$ are the probability distributions of the Markov chain at t with initial distributions p_0 and p'_0 at time t_0 , respectively, and for two distributions μ and ν , $\|\mu - \nu\|$ denotes the total variation of $\mu - \nu$.

Proof. We adopt proof arguments similar to those for Theorem 4.5.1 in Winkler (2012) and Theorem 5.3 in Morita and Nishimori (2008). To show the convergence to the equilibrium and the strong ergodicity, we need to prove the weak ergodicity of the Markov chain and

$$\sum_n \|\mu_n - \mu_{n+1}\| < \infty, \quad (\text{B.2})$$

where $\mu_n(\mathbf{s}) \propto q(\mathbf{s}) \exp(-\frac{2J^{\perp}(n)}{\tau T} \Lambda(\mathbf{s}))$ is an invariant distribution of the chain at time n .

We first prove Equation (B.2). By Lemma 4.5.2 in Winkler (2012), we only need to show that μ_n eventually decreases as $\Gamma(n) \rightarrow 0$, which is easy to obtain. For \mathbf{s} such that $\Lambda(\mathbf{s}) = 0$, $\mu_n(\mathbf{s}) \propto q(\mathbf{s})$ does not change over time. For other \mathbf{s} , as $\Gamma(n) \rightarrow 0$, $\frac{2J^{\perp}(n)}{\tau T}$ is monotone increasing to $+\infty$, and thus μ_n is eventually decreasing. This also proves that the equilibrium distribution is $q(\mathbf{s})\mathbf{I}(\Lambda(\mathbf{s}) = 0)$.

Next, we show the weak ergodicity of the Markov chain. With Theorem 5.1 in Morita and Nishimori (2008), we only need to show that there exists a strictly increasing sequence of positive number $\{t_i, i = 0, 1, \dots\}$ such that

$$\sum_{k=1}^{\infty} (1 - \alpha(G^{t_i, t_{i+1}})) \rightarrow \infty, \quad (\text{B.3})$$

where $\alpha(G^{t_i, t_{i+1}})$ is the contraction coefficient defined by

$$\alpha(G^{t_i, t_{i+1}}) = 1 - \min_{\mathbf{s}, \hat{\mathbf{s}}} \left\{ \sum_{\hat{\mathbf{s}}} \min\{G^{t_i, t_{i+1}}(\mathbf{s}, \hat{\mathbf{s}}), G^{t_i, t_{i+1}}(\hat{\mathbf{s}}, \hat{\mathbf{s}})\} \right\}, \quad (\text{B.4})$$

with $G^{t_i+1, t_i}(s, \bar{s})$ denotes the transition probability of the chain to move from s at time t_i to \bar{s} at time t_{i+1} .

First, we establish the lower bound on the transition probability defined by (C1) and (C2). Define

$$L_0 = \max_{s, \bar{s}} \left| \sum_{k=1}^{\tau} \sum_{(i,j)} J_{ij} s_i^k s_j^k - \sum_{k=1}^{\tau} \sum_{(i,j)} J_{ij} \bar{s}_i^k \bar{s}_j^k \right|,$$

and

$$w = \min_{s, \bar{s}} \{N(s, \bar{s}) : N(s, \bar{s}) > 0\}.$$

Lemma 1. For any $s \neq \bar{s}$ with $N(s, \bar{s}) > 0$, then we have for any $t > 0$,

$$G(s, \bar{s}; t) \geq wA(e^{-\frac{L_0}{\tau T}})e^{-\frac{2J^\perp(t)d}{T}}. \quad (\text{B.5})$$

For any state $s \notin \mathcal{S}_m$, there exists $t_1 > 0$ such that $\forall t > t_1$,

$$G(s, s; t) > wA(e^{-\frac{L_0}{\tau T}})e^{-\frac{2J^\perp(t)d}{T}}. \quad (\text{B.6})$$

Proof. For any $s \neq \bar{s}$ with $N(s, \bar{s}) > 0$, we have for any $t > 0$,

$$G(s, \bar{s}; t) \geq wA(e^{-\frac{L_0}{\tau T}})e^{-\frac{2J^\perp(t)d}{T}} = wA(e^{-\frac{L_0}{\tau T}})e^{-\frac{2J^\perp(t)d}{T}},$$

where we have used the monotonicity of $A(\cdot)$ and

$$\Lambda(\bar{s}) - \Lambda(s) \vee 0 \leq \max_s \Lambda(s) = \max_s \sum_{k=1}^{\tau} \Lambda(s^k, s^{k+1}) = d\tau.$$

For any state $s \notin \mathcal{S}_m$, there exists $\hat{s} \in \mathcal{N}_s$ such that $\hat{s} \neq s$, $\Lambda(\hat{s}) - \Lambda(s) > 0$, so by the discreteness of the Hamming distance, we have $\Lambda(\hat{s}) - \Lambda(s) \geq 1$. Since $J^\perp(t) \rightarrow \infty$ as $t \rightarrow \infty$, for any $0 < \epsilon < 1$, there exists $t_1 > 0$ such that

$$\forall t > t_1, e^{-\frac{2J^\perp(t)}{\tau T}((\Lambda(\hat{s}) - \Lambda(s)) \vee 0)} \leq e^{-\frac{2J^\perp(t)}{\tau T}} < \epsilon.$$

Therefore, we conclude

$$\begin{aligned} \sum_{\bar{s} \neq s} G(s, \bar{s}; t) &= G(s, \hat{s}; t) + \sum_{\bar{s} \neq s, \bar{s} \neq \hat{s}} G(s, \bar{s}; t) \\ &\leq N(s, \hat{s})\epsilon + \sum_{\bar{s} \neq s, \bar{s} \neq \hat{s}} N(s, \bar{s}) \\ &= N(s, \bar{s})\epsilon + 1 - N(s, \bar{s}) = 1 - (1 - \epsilon)N(s, \bar{s}), \end{aligned}$$

and

$$G(s, s; t) = 1 - \sum_{\bar{s} \neq s} G(s, \bar{s}; t) \geq (1 - \epsilon)N(s, \bar{s}) > 0.$$

Finally, we can easily prove (B.6) by noting that its right-hand side can be arbitrarily small for sufficiently large t . \square

The generation probability N is positive and irreducible, and we define R as the maximum number of minimum steps needed to reach an arbitrary state $s \in \mathcal{S}_m$ from any other state. Then as long as $\Gamma(t) > 0$, this number remains the same for the inhomogeneous Markov chain. Now consider state $s^* \in \mathcal{S}_m$ such that the number of maximum steps needed to reach it from any other state is at most R . Then there exists a path such that

$$s = s_0 \neq s_1 \neq \dots \neq s_k = s_{k+1} = \dots = s_R = s^*.$$

The above lemma yields that, for sufficiently large t , the transition probability at each time step has the following lower bound,

$$G(s_i, s_{i+1}; t - R + i) \geq wA(e^{-\frac{L_0}{\tau T}})e^{-\frac{2J^\perp(t-R+i)d}{T}}.$$

Combining all of them together, we obtain

$$\begin{aligned} G^{t-R, t}(s, s^*) &\geq G(s, s_1; t - R) \dots \\ &G(s_{R-2}, s_{R-1}; t - 2)G(s_{R-1}, s^*; t - 1) \\ &\geq \prod_{i=0}^{R-1} wA(e^{-\frac{L_0}{\tau T}})e^{-\frac{2J^\perp(t-R+i)d}{T}} \\ &\geq w^R A(e^{-\frac{L_0}{\tau T}})^R e^{-\frac{2J^\perp(t-1)dR}{T}}, \end{aligned} \quad (\text{B.7})$$

where we have used the monotonicity of $J^\perp(t)$. Hence, there exists an integer $k_0 \geq 0$ such that for all $k \geq k_0$, the contraction coefficient satisfies

$$\begin{aligned} 1 - \alpha(G^{kR-R, kR}) &= \min_{s, \bar{s}} \left\{ \sum_{\hat{s}} \min\{G^{kR-R, kR}(s, \hat{s}), G^{kR-R, kR}(\bar{s}, \hat{s})\} \right\} \\ &\geq \min_{s, \bar{s}} \left\{ \min\{G^{kR-R, kR}(s, s^*), G^{kR-R, kR}(\bar{s}, s^*)\} \right\} \\ &\geq w^R A(e^{-\frac{L_0}{\tau T}})^R e^{-\frac{2J^\perp(kR-1)dR}{T}}. \end{aligned} \quad (\text{B.8})$$

Finally with annealing schedule (14), we easily establish the weak ergodicity by

$$\sum_{k=1}^{\infty} (1 - \alpha(G^{kR, kR-R})) \geq w^R A(e^{-\frac{L_0}{\tau T}})^R \sum_{k=1}^{\infty} \frac{1}{kR+1} \rightarrow \infty.$$

This concludes the proof of the theorem. \square

Appendix C. Proof of Corollary 1

Proof. With global move, the total neighboring Hamming distance $\Lambda(s)$ of the configuration matrix is intact, which indicates that the proposed distribution cannot be irreducible. Thus, when $\Gamma(t) > 0$, the irreducibility of the transition kernels for the SQA-PI algorithm as well as the proposed SQA-DA algorithm is the same as that for $N_{\text{global}}(s, \bar{s})$. This immediately leads to (i).

For the proof of (ii), since all slices are the same, and we only move within the set of configuration matrices with all same slices, the state space is equivalent to that of an Ising model (1) with $h_i = 0$. Then the global move is essentially doing a site-by-site update, that is, the usual Metropolis sampler for the SA case. \square

Supplementary Materials

Code and data: An R package which consists of datasets and programs for all methods used in the numerical studies, along with an example code file necessary to reproduce the results in this article. (zip file).

Acknowledgments

The authors thank Editor Tyler McCormick, an associate editor, and two anonymous referees for helpful comments and suggestions which led to significant improvements of the article.

Funding

The research of Yazhen Wang was supported in part by NSF grants DMS-15-28375 and DMS-17-07605.

References

- Boixo, S., Rønnow, T. F., Isakov, S. V., Wang, Z., Wecker, D., Lidar, D. A., Martinis, J. M., and Troyer, M. (2014), “Evidence for Quantum Annealing With More Than One Hundred Qubits,” *Nature Physics*, 10, 218–224. [284,289]
- Farhi, E., Goldstone, J., Gutmann, S., Lapan, J., Lundgren, A., and Preda, D. (2001), “A Quantum Adiabatic Evolution Algorithm Applied to Random Instances of an NP-Complete Problem,” *Science*, 292, 472–475. [284,286]
- Farhi, E., Goldstone, J., Gutmann, S., and Sipser, M. (2000), “Quantum Computation by Adiabatic Evolution,” arXiv no. quant-ph/0001106. [284,286]
- Geman, S., and Geman, D. (1987), “Stochastic Relaxation, Gibbs Distributions, and the Bayesian Restoration of Images,” in *Readings in Computer Vision*, eds. M. A. Fischler and O. Firschein, Amsterdam: Elsevier, pp. 564–584. [284,285]
- Hajek, B. (1988), “Cooling Schedules for Optimal Annealing,” *Mathematics of Operations Research*, 13, 311–329. [285,288]
- Kadowaki, T., and Nishimori, H. (1998), “Quantum Annealing in the Transverse Ising Model,” *Physical Review E*, 58, 5355–5363. [284,286]
- Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983), “Optimization by Simulated Annealing,” *Science*, 220, 671–680. [284]
- Martoňák, R., Santoro, G. E., and Tosatti, E. (2002), “Quantum Annealing by the Path-Integral Monte Carlo Method: The Two-Dimensional Random Ising Model,” *Physical Review B*, 66, 094203. [286,287]
- Meyn, S. P., and Tweedie, R. L. (2012), *Markov Chains and Stochastic Stability*, London: Springer-Verlag. [294]
- Morita, S., and Nishimori, H. (2008), “Mathematical Foundation of Quantum Annealing,” *Journal of Mathematical Physics*, 49, 125210. [286,288,294]
- Nielsen, M. A., and Chuang, I. L. (2010), *Quantum Computation and Quantum Information*, Cambridge Series on Information and the Natural Sciences, New York: Cambridge University Press. [286]
- Swendsen, R. H., and Wang, J.-S. (1987), “Nonuniversal Critical Dynamics in Monte Carlo Simulations,” *Physical Review Letters*, 58, 86. [294]
- Wang, Y. (2012), “Quantum Computation and Quantum Information,” *Statistical Science*, 27, 373–394. [286]
- Wang, Y., and Song, X. (2020), “Quantum Science and Quantum Technology,” *Statistical Science*, 35, 51–74. [286]
- Wang, Y., Wu, S., and Zou, J. (2016), “Quantum Annealing With Markov Chain Monte Carlo Simulations and D-Wave Quantum Computers,” *Statistical Science*, 31, 362–398. [284,286,287,289]
- Winkler, G. (2012), *Image Analysis, Random Fields and Markov Chain Monte Carlo Methods: A Mathematical Introduction* (Vol. 27), Berlin, Heidelberg: Springer. [284,285,288,294]