To appear in *Physical Review E*

Noise Can Speed Markov Chain Monte Carlo Estimation and Quantum Annealing

Brandon Franzke[1] and Bart Kosko[1, a)]

Center for Quantum Information Science and Technology
Signal and Image Processing Institute
Department of Electrical and Computer Engineering
University of Southern California, Los Angeles, California 90089

Carefully injected noise can speed the average convergence of Markov chain Monte Carlo (MCMC) estimates and simulated annealing optimization. This includes quantum annealing and the MCMC special case of the Metropolis-Hastings algorithm. MCMC seeks the solution to a computational problem as the equilibrium probability density of a reversible Markov chain. The algorithm must cycle through a long burn-in phase until it reaches equilibrium because the Markov samples are statistically correlated. The special injected noise reduces this burn-in period in MCMC. A related theorem shows that it reduces the cooling time in simulated annealing. Simulations showed that optimal noise gave a 76% speed-up in finding the global minimum in the Schwefel optimization benchmark. The noise-boosted simulations found the global minimum in 99.8% of trials compared with only 95.4% of trials in noiseless simulated annealing. Simulations also showed that the noise boost is robust to accelerated cooling schedules and that noise decreased convergence times by more than 32% under aggressive geometric cooling. Molecular dynamics simulations showed that optimal noise gave a 42% speed-up in finding the minimum potential energy configuration of an 8-argon-atom gas system with a Lennard-Jones 12-6 potential. The annealing speed-up also extends to quantum Monte Carlo implementations of quantum annealing. Noise improved ground-state energy estimates in a 1024-spin simulated quantum annealing simulation by 25.6%. The quantum noise flips spins along a Trotter ring. The Noisy MCMC algorithm brings each Markov step closer on average to equilibrium if an inequality holds between two expectations. Gaussian or Cauchy jump probabilities reduce the noise-benefit inequality to a simple quadratic inequality. Simulations show that noise-boosted simulated annealing is more likely than noiseless annealing to sample high probability regions of the search space and to accept solutions that increase the search breadth.

Keywords: Markov chain Monte Carlo (MCMC) simulation, Metropolis-Hastings simulated annealing, quantum Monte Carlo (QMC), quantum annealing, noise benefits, Bayesian statistics

I. NOISE-BOOSTING MCMC ESTIMATION

We show that carefully injected noise can speed the convergence of Markov Chain Monte Carlo (MCMC) estimates. The noise randomly perturbs the signal and widens the breadth of search. The perturbation can be additive or multiplicative or some other measurable function. The two main theorems below use additive noise only for simplicity. They hold for arbitrary combinations of noise and signal. One corollary does use multiplicative noise for a Gaussian jump density.

The injected noise must satisfy an inequality that incorporates the detailed-balance condition of a reversible Markov chain. So the process is not simply blind independent noise injection as in stochastic resonance[8,15,28,31,32,34,40–42,54]. The specially chosen noise perturbs the current state so as to make the state more probable within the constraints of reversibility. This constrained probabilistic noise differs from the search probability even if they are both Gaussian because the system injects only that subset of Gaussian noise that satisfies the inequality.

The noise boost shortens the distance between the current sampled probability density and the desired equilibrium density. It reduces on average the Kullback-Liebler pseudo-distance between these two densities. This leads to a shorter "burn in" time before the user can safely estimate integrals or other statistics based on sample averages as in regular (uncorrelated) Monte Carlo simulation.

The MCMC noise boost extends to simulated annealing with different cooling schedules. It also extends to quantum-annealing search that burrows through a cost surface rather than thermally bounces over it as in classical annealing. The quantum-annealing noise propagates along an Ising lattice. It conditionally flips the corresponding sites on coupled Trotter slices.

MCMC is a powerful statistical optimization technique that exploits the convergence properties of Markov chains[9,17]. These properties include Markov-chain versions of the laws of large numbers and the central limit theorem[21]. It often works well on high-dimensional problems of statistical physics, chemical kinetics, genomics, decision theory, machine learning, quantum computing, financial engineering, and Bayesian inference[7]. Special cases of MCMC include the Metropolis-Hastings algorithm and Gibbs sampling in Bayesian statistical inference[16,19,50].

MCMC solves an inverse problem: How can the system reach a given solution from any starting point of the Markov chain?

MCMC draws random samples from a reversible Markov chain. It then computes sample averages to estimate population statistics. The designer picks the Markov chain so that its equilibrium probability density function corresponds to the solution of a given computational problem. The correlated

[a)] Electronic mail: kosko@usc.edu

samples can require cycling through a long "burn in" period before the Markov chain equilibrates. We show that careful (non-blind) noise injection can speed up this lengthy burn-in period.

MCMC simulation itself arose in the early 1950s when physicists modeled the intense energies and high particle dimensions involved in the design of thermonuclear bombs. These simulations ran on the MANIAC and other early computers[29]. Many refer to this algorithm as the Metropolis algorithm or the Metropolis-Hastings algorithm after Hastings' modification to it in 1970[19]. The original 1953 paper[29] computed thermal averages for 224 hard spheres that collided in the plane. Its high-dimensional state space was $\mathbb{R}^{448}$. So even standard random-sample Monte Carlo techniques were not feasible. The name "simulated annealing" has also become common since Kirkpatrick's work on spin glasses and VLSI optimization in 1983 for MCMC that uses a cooling schedule[22].

The Noisy MCMC (N-MCMC) algorithm below resembles but differs from our earlier "stochastic resonance" work on using noise to speed up stochastic convergence. We showed earlier how adding noise to a Markov chain's state density can speed convergence to the chain's equilibrium probability density π if we know π in advance[14]. The noise did not add to the state. Nor was it part of the MCMC framework that solves the inverse problem of starting with π and finding a Markov chain that leads to it.

The Noisy Expectation-Maximization (NEM) algorithm did show on average how to boost each iteration of the EM algorithm[10,12] as the estimator climbs to the top of the nearest hill on a likelihood surface[38,39]. This noise result also showed how to speed up the popular backpropagation algorithm in neural networks because we also showed that the backpropagation gradient-descent algorithm is a special case of the generalized EM algorithm[3,5]. The same NEM algorithm boosts the popular Baum-Welch method for training hidden-Markov models in speech recognition and elsewhere[4]. It boosts the k-means-clustering algorithm found in pattern recognition and big data[37]. It also boosts recurrent backpropgation in machine learning[1].

The N-MCMC algorithm and theorem below stem from a simple intuition: Find a noise sample n that makes the next choice of location $x+n$ more probable. Define the usual state transition function $Q(y|x)$ as the probability that the system moves or jumps to state y if it is in state x. The sample x is a realization of the location random variable X. The sample n is a realization of the noise random variable N. The Metropolis algorithm requires a symmetric state transition function: $Q(y|x) = Q(x|y)$. This helps explain the common choice of a Gaussian jump function. Neither the Metropolis-Hastings algorithm nor the N-MCMC results require symmetry. But all MCMC algorithms do require that the chain is reversible. Physicists call this *detailed balance*:

$$Q(y|x)\pi(x) = Q(x|y)\pi(y) \tag{1}$$

for all x and y.

Now consider a noise sample n that makes the jump more probable: $Q(y|x+n) \geq Q(y|x)$. This is equivalent to $\ln \frac{Q(y|x+n)}{Q(y|x)} \geq 0$. Replace the denominator jump term with its symmetric dual $Q(x|y)$. Then eliminate this term with detailed balance and rearrange to get the key inequality for a noise boost:

$$\ln \frac{Q(y|x+n)}{Q(y|x)} \geq \ln \frac{\pi(x)}{\pi(y)}. \tag{2}$$

Taking expectations over the noise random variable N and over X gives a simple symmetric version of the sufficient condition in the Noisy MCMC Theorem for a speed-up:

$$E_{N,X}\left[\ln \frac{Q(y|x+N)}{Q(y|x)}\right] \geq E_X\left[\ln \frac{\pi(x)}{\pi(y)}\right]. \tag{3}$$

The inequality (3) has the form $A \geq B$. So it generalizes the structurally similar sufficient condition $A \geq 0$ that governs the NEM algorithm[39]. This is natural since the EM algorithm deals with only the likelihood term $P(E|H)$ on the right side of Bayes Theorem: $P(H|E) = \frac{P(H)P(E|H)}{P(E)}$ for hypothesis H and evidence E. MCMC deals with the converse posterior probability $P(H|E)$ on the left side. The posterior requires the extra prior $P(H)$. This accounts for the right-hand side of (3).

The next sections review MCMC and then extend it to the noise-boosted case. Theorem 1 proves that at each step the noise-boosted chain is closer on average to the equilibrium density than is the noiseless chain. Theorem 2 proves that noisy simulated annealing increases the sample acceptance rate to exploit the noise-boosted chain. The first corollary uses an exponential term to weaken the sufficient condition. The next two corollaries state a simple quadratic condition for the noise boost when the jump probability is a Gaussian bell curve or a Cauchy bell curve. A Cauchy bell curve has slightly thicker tails and gives occasional longer jumps.

The next section presents the Noisy Markov Chain Monte Carlo Algorithm and the Noisy Simulated Annealing Algorithm. Three simulations show the predicted MCMC noise benefit. The first shows that noise decreases convergence time in Metropolis Hastings optimization of the highly nonlinear Schwefel function (Figure 1) by 75%. Figure 2 shows two sample paths and describes the origin of the convergence noise benefit. Then we show noise benefits in an 8-argon-atom molecular dynamics simulation that uses a Lennard-Jones 12-6 interatomic potential and a Gaussian-jump model. Figure 8 shows that the optimal noise gives a 42% speed-up. It took 173 steps to reach equilibrium with N-MCMC compared with 300 steps in the noiseless case. The third simulation shows that a noise-boosted path-integral Monte Carlo quantum annealing improves the estimated ground state of a 1024-spin Ising spin glass system by 25.6%. We were not able to quantify the decrease in convergence time because the non-noisy quantum annealing algorithm did not converge to a ground state this low in any trial.

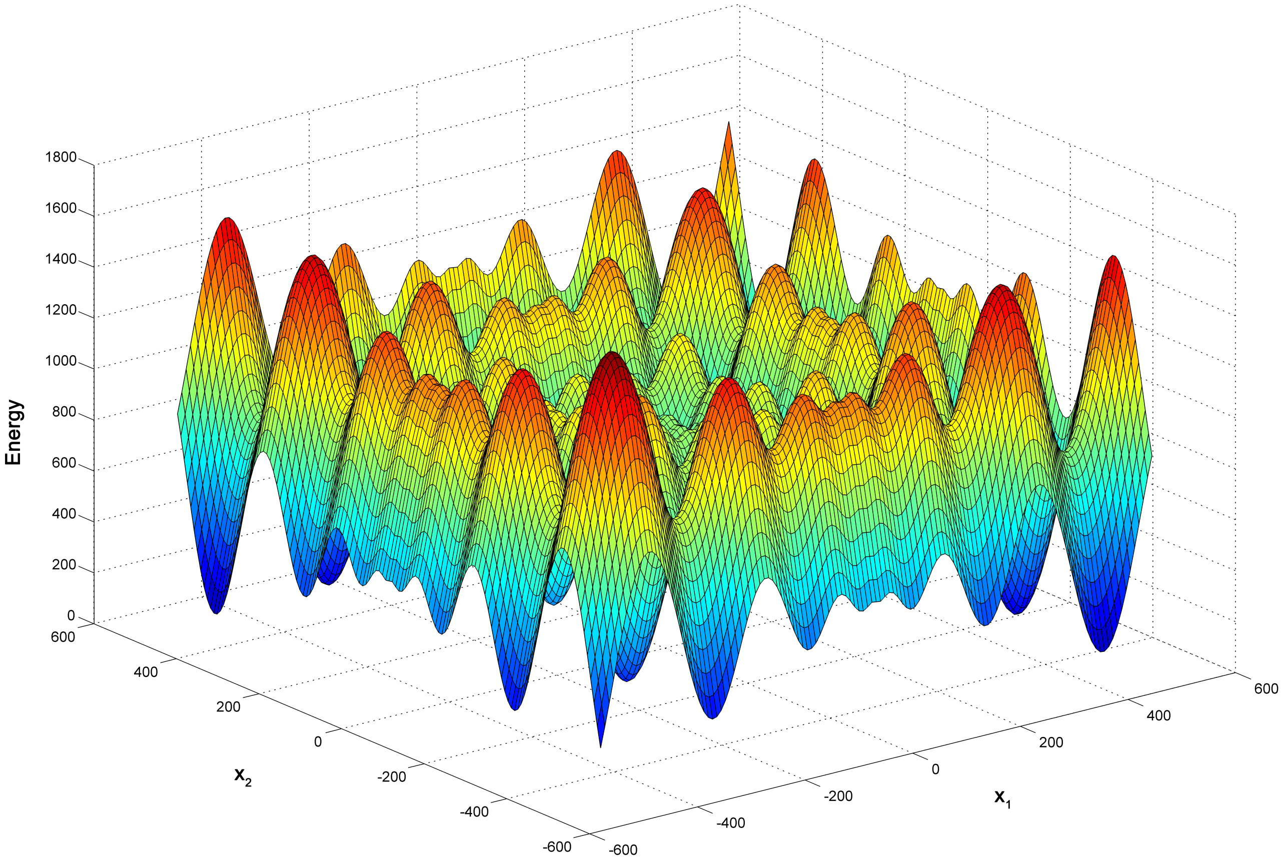


FIG. 1. Schwefel function $f(x) = 419.9829d - \sum_{i=1}^{d} x_i \sin\left(\sqrt{|x|}\right)$ is a d-dimensional optimization benchmark on the hypercube $-512 \leq x_i \leq 512$[11,33,48,53]. It has a single global minimum $x_{min}(=)0$ at $x_{min} = (420.9687, \ldots, 420.9687)$. Energy peaks separate irregular troughs on the surface. This leads to estimate capture in search algorithms that emphasize local search.

II. MARKOV CHAIN MONTE CARLO

We first review the Markov chains that underlie the MCMC algorithm[44]. This includes the important MCMC special case called the Metropolis-Hastings algorithm.

A Markov chain is a random process whose future depends only on the present. It has no memory of the past. So its transitions from one state to another obey the Markov property

$$P(X_{t+1} = x \mid X_1 = x_1, \ldots, X_t = x_t) = P(X_{t+1} = x \mid X_t = x_t). \tag{4}$$

P is the single-step transition probability matrix where

$$P_{i,j} = P(X_{t+1} = j \mid X_t = i) \tag{5}$$

is the probability that the chain in state i at time t moves to state j at time $t+1$.

State j is *accessible* from state i if there is some non-zero probability of transitioning from state i to state j ($i \rightarrow j$) in any finite number of steps

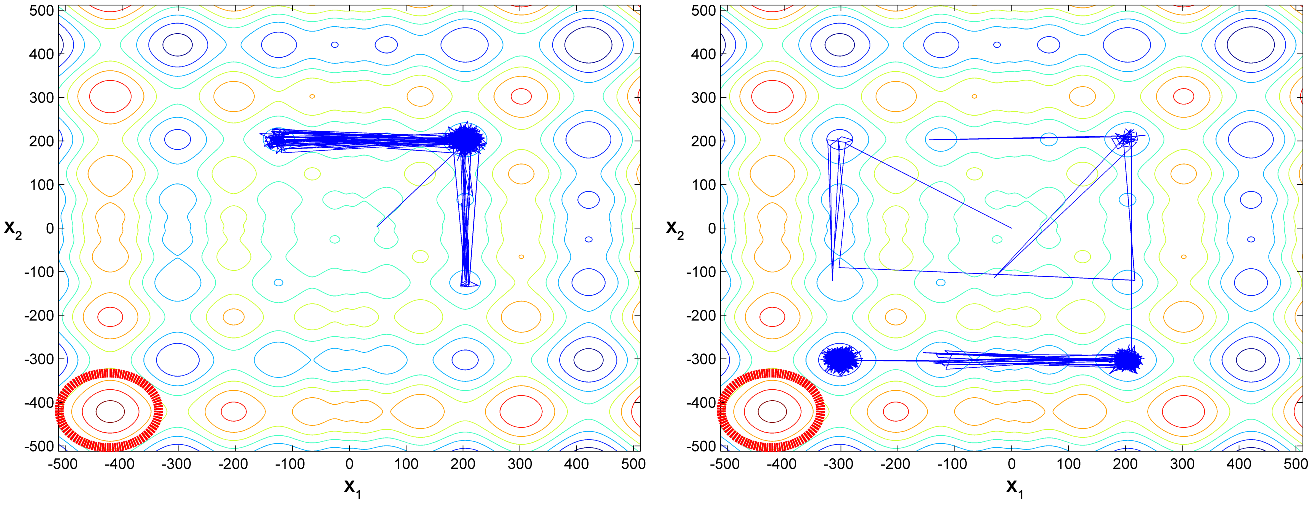
$$P_{i,j}^{(n)} > 0 \tag{6}$$

for some $n > 0$. A Markov chain is *irreducible* if each state is accessible from all other states[30,44]. Irreducibility implies that for all states i and j there exists $m > 0$ such that $P(X_{n+m} = j \mid X_n = i) = P_{i,j}^{(m)} > 0$. This holds if and only if P is a *regular stochastic matrix*.

The period d_i of state i is $d_i = \gcd\left\{n \geq 1 : P_{i,i}^{(n)} > 0\right\}$ or $d_i = \infty$ if $P_{i,i}^{(n)} = 0$ for all $n \geq 1$ if gcd denotes the greatest common divisor. State i is aperiodic if $d_i = 1$. A Markov chain with transition matrix P is *aperiodic* if $d_i = 1$ for all states i.

A sufficient condition for a Markov chain to have a unique stationary distribution π is that the state transitions satisfy *detailed balance*: $P_{j,k} \cdot x_j^\infty = P_{k,j} \cdot x_k^\infty$ for all states j and k. We can also write this as $Q(k|j)\pi(j) = Q(j|k)\pi(k)$. This is called the reversibility condition. A Markov chain is *reversible* if it obeys detailed balance.

Markov Chain Monte Carlo algorithms exploit the Markov convergence guarantee construct Markov chains with samples drawn from complex probability densities. But MCMC methods suffer from problem-specific parameters that govern sample acceptance and convergence assessment[18,51]. Strong dependence on initial conditions also biases the MCMC sampling unless the simulation allows a lengthy period of "burn-



(a) Without noise

(b) With noise

FIG. 2. Noise increased the breadth of search for simulated-annealing sample sequences from a 5-dimensional (projected to 2-D) Schwefel surface with log cooling schedule. Noisy simulated annealing visited more local minima and quickly moved out of those minima that trapped non-noisy SA. Both figures show sample sequences with initial condition $x_0 = (0,0)$ and $N = 10^6$. The red circle (lower left) locates the global minimum at $x_{min} = (-420.9687, -420.9687)$. (a) The noiseless algorithm found the $(205,205)$ local minimum within the first 100 time steps. Thermal noise did not induce the noiseless algorithm to search the space beyond three local minima. (b) The noisy simulation followed the noiseless simulation at the simulation start. It also sampled the same regions but in accord with Algorithm 2. The noise injected in accord with Theorem 2 both enhanced the thermal jumps and increased the breadth of the simulation. The noise-injected simulation visited the same three minima as in (a) but it performed a local optimization for only a few hundred steps before it jumped to the next minimum. The estimate settled at $(-310, -310)$ just one hop away from the global minimum x_{min}.

in” to allow the driving Markov chain to mix adequately.

A. The Metropolis-Hastings Algorithm

We next present Hastings’[19] generalization of the MCMC Metropolis algorithm now called Metropolis-Hastings. This starts with the classical Metropolis algorithm[29].

Suppose we want to sample $x_1, \ldots, x_n$ from a random variable X with probability density function (pdf) $p(x)$. Suppose $p(x) = \frac{f(x)}{K}$ for some function $f(x)$ and normalizing constant K. We may not know the normalizing constant K or it may be hard to compute. The Metropolis algorithm constructs a Markov chain with the target equilibrium density π. The algorithm generates a sequence of samples from $p(x)$.

1. Choose an initial x_0 with $f(x_0) > 0$.

2. Generate a candidate x^*_{t+1} by sampling from the *jump distribution* $Q(x_{t+1}|x_t)$. The jump pdf must be symmetric: $Q(x_{t+1}|x_t) = Q(x_t|x_{t+1})$.

3. Calculate the density ratio for x^*_{t+1}: $\alpha = \frac{p\left(x^*_{t+1}\right)}{p(x_t)} = \frac{f\left(x^*_{t+1}\right)}{f(x_t)}$. Note that the normalizing constant K cancels.

4. Accept the candidate point ($x_{t+1} = x^*_{t+1}$) if the jump increases the probability ($\alpha > 1$). Also accept the candidate point with probability α if the jump *decreases* the probability. Else reject the jump ($x_{t+1} = x_t$) and return to step 2.

Hastings’[19] replaced the symmetry constraint on the jump distribution Q with $\alpha = \min\left(\frac{f\left(x^*_{t+1}\right)Q\left(x_t|x^*_{t+1}\right)}{f(x_t)Q\left(x^*_{t+1}|x_t\right)}, 1\right)$. Then detailed balance still holds[44]. Gibbs sampling is a special case of the Metropolis-Hastings algorithm when $\alpha = 1$ always holds for each conditional pdf[7,44].

B. Simulated Annealing

We next present a time-varying version of the Metropolis-Hastings algorithm for global optimization of a high-dimensional surface with many extrema. Kirkpatrick[22] called this process *simulated annealing* because it resembles the metallurgical annealing process that slowly cools a heated substance until it reaches a low-energy crystalline state.

The simulated version uses a temperature-like parameter T. T is so high at first that search is essentially random. T lowers until the search is greedy or locally optimal. Then the system state tends to get trapped in a large minimum or even in the global minimum. Kirkpatrick applied this thermodynamically inspired algorithm to finding optimal layouts for VLSI circuits.

Suppose we want to find the global minimum of a cost function $C(x)$. Simulated annealing maps the cost function to a

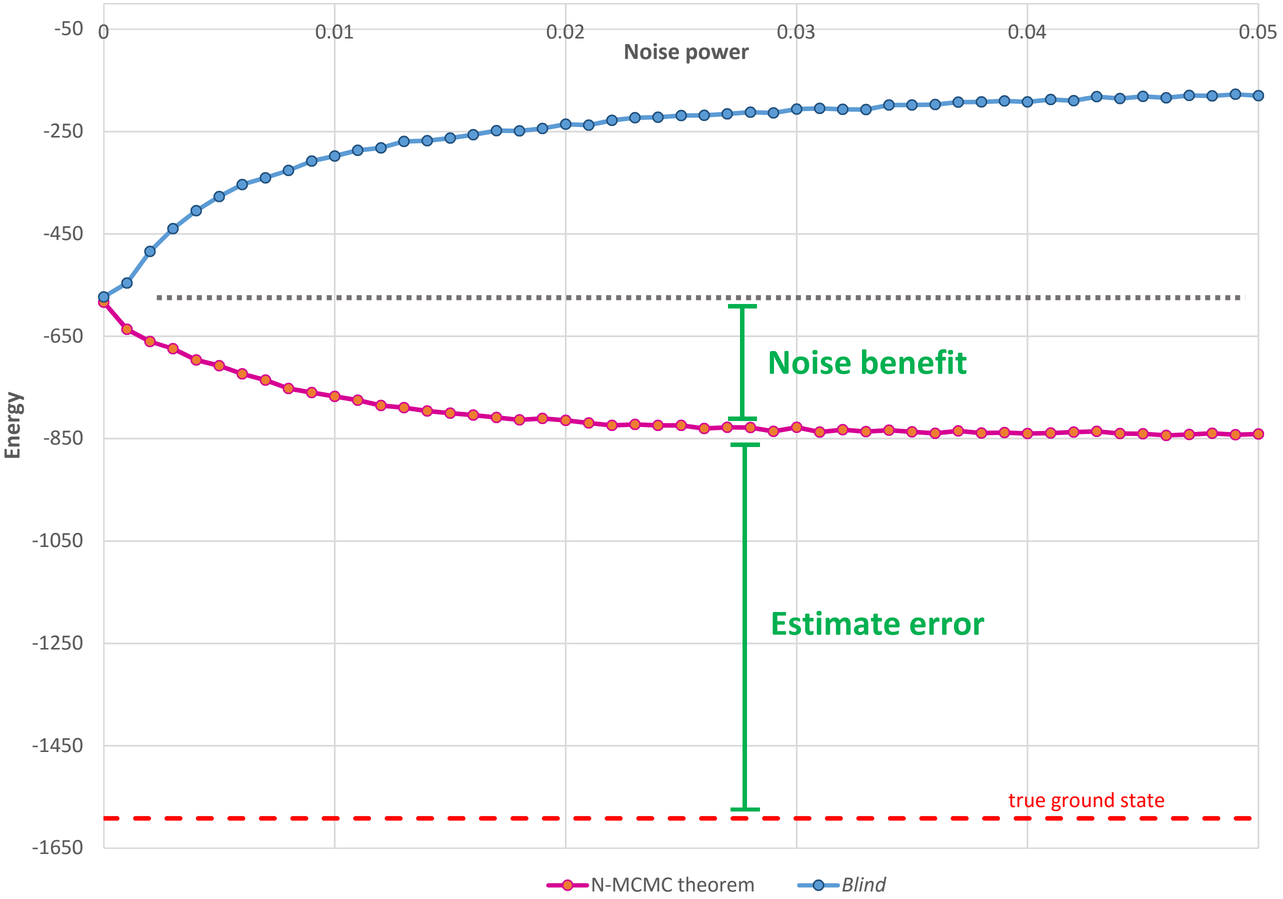


FIG. 3. Simulated quantum annealing noise benefit in a 1024 Ising spin simulation. The pink line shows that noise improved the estimated ground-state energy of a 32x32 spin lattice by 25.6%. The plot shows the ground state energy after 100 path-integral Monte Carlo steps. The true ground state energy (red) was $E_0 = -1591.92$. Each plotted point shows the average calculated ground state from 100 simulations at each noise power. The blue line shows that *blind* (independent and identically distributed sampling) noise did not benefit the simulation. Blind noise only made the estimates worse. So the N-MCMC noise-benefit condition is central to the S-QA noise benefit.

potential energy surface through the Boltzmann factor

$$\widetilde{p}(x_t) \propto \exp\left[-\frac{C(x_t)}{kT}\right] \qquad (7)$$

for some scaling constant $k > 0$. It then performs the Metropolis-Hastings algorithm with the $\widetilde{p}(x_t)$ in place of the probability density $p(x)$. This operation preserves the Metropolis-Hastings framework because $\widetilde{p}(x_t)$ is an unnormalized probability density.

Simulated annealing introduces a temperature parameter to tune the Metropolis-Hastings acceptance probability α. The algorithm slowly cools the system according to a cooling schedule $T(t)$. This reduces the probability of accepting candidate points with higher energy. The algorithm provably attains a global minimum in the limit but this requires an extremely slow $\log(t+1)$ cooling. Accelerated cooling schedules such as geometric or exponential often yield satisfactory approximations in practice. The procedure below describes the algorithm. The algorithm attains the global minimum as $t \to \infty$.

1. Choose an initial x_0 with $C(x_0) > 0$ and initial temperature T_0.

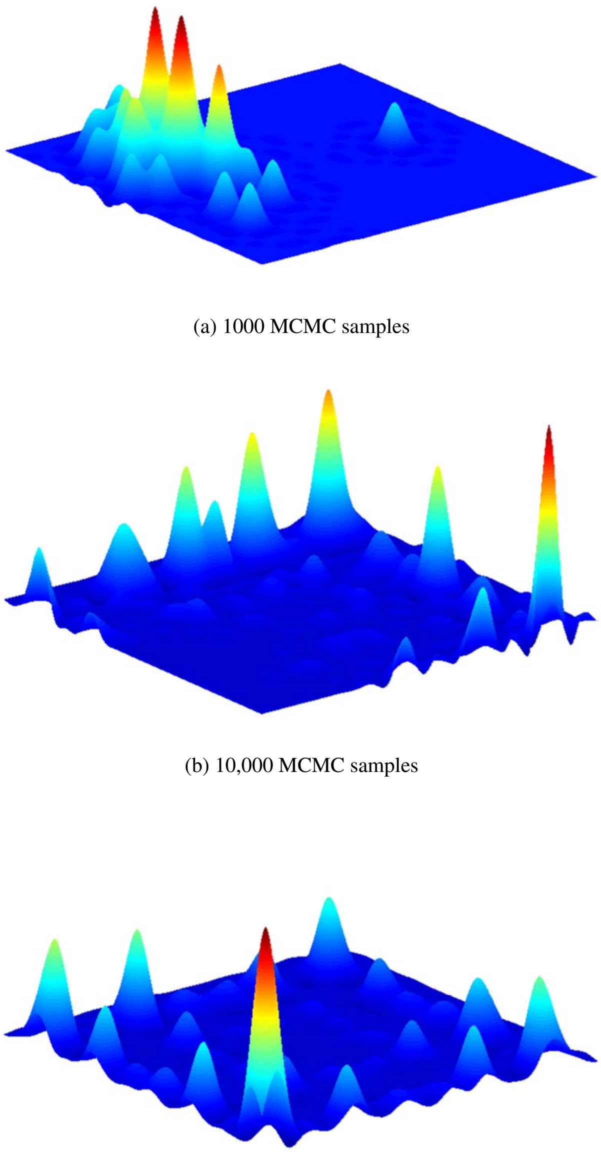
2. Generate a candidate x^*_{t+1} by sampling from the *jump distribution* $Q(x_{t+1}|x_t)$.

3. Compute the Boltzmann factor $\alpha = \exp\left(-\frac{C\left(x^*_{t+1}\right)-C(x_t)}{kT}\right)$.

4. Accept the candidate point ($x_{t+1} = x^*_{t+1}$) if the jump decreases the energy. Also accept the candidate point with probability α if the jump *increases* the energy. Else reject the jump ($x_{t+1} = x_t$).

5. Update the temperature $T_t = T(t)$. $T(t)$ is usually a monotonic decreasing function.

6. Return to step 2.



(a) 1000 MCMC samples

(b) 10,000 MCMC samples

(c) 100,000 MCMC samples

FIG. 4. Time evolution of a 2-dimensional histogram of MCMC samples from the 2-D Schwefel function in Figure 1. (a) The simulation has explored only a small region of the space after 1000 samples. The simulation has not sufficiently *burned in*. The samples remain close to the initial state because the MCMC random walk proposed new samples near the current state. This early histogram did not match the Schwefel density. (b) The 10,000 sample histogram better matched the target density but there were still large unexplored regions. (c) The 100,000 sample histogram shows that the simulation explored most of the search space. The tallest (red) peak shows that the simulation found the global minimum. The histogram peaks corresponded to energy minima on the Schwefel surface.

III. NOISY MARKOV CHAIN MONTE CARLO

We now show how carefully injected noise can speed the average convergence of MCMC simulations in terms of reducing the relative-entropy (Kullback-Liebler divergence) pseudo-distance. This basic theorem leads to many variants.

Theorem 1 states the Noisy MCMC (N-MCMC) Theorem and gives a simple inequality as a sufficient condition for the speed-up. The Appendix gives the proof. We also include algorithm statements of the main results. We note that reversing inequalities in the N-MCMC Theorem leads to noise that on average slows convergence.

Corollary 1 weakens the sufficient condition of Theorem 1 through the use of a new exponential term. Corollary 2 allows noise injection with any measurable combination of noise and state. Corollary 3 shows that a Gaussian jump function reduces the sufficient condition to a simple quadratic inequality. Figure 7 shows simulation instances of Corollary 2 for a Lennard-Jones model of the interatomic potential of a gas of 8 argon atoms. The graph shows the optimal Gaussian variance for the quickest convergence to the global minimum of the potential energy. Corollary 5 states a similar quadratic inequality when the jump function is the thicker-tailed Cauchy probability bell curve. Earlier simulations showed that a Cauchy jump function can lead to “fast” simulated annealing because sampling from its thicker tails can lead to more frequent long jumps out of shallow local minima[52].

Theorem 1 is the main contribution of this paper. It shows that injecting noise into a jump density that satisfies detailed balance can only bring the jump density closer to the equilibrium density if the noise-injected jump density satisfies an average inequality at each iteration.

Theorem 1 (Noisy Markov Chain Monte Carlo Theorem (N-MCMC))**.** *Suppose that $Q(x|x_t)$ is a Metropolis-Hastings jump pdf for time t and that it satisfies the detailed balance condition $\pi(x_t)\,Q(x|x_t) = \pi(x)\,Q(x_t|x)$ for the target equilibrium pdf $\pi(x)$. Then the MCMC noise benefit $d_t(N) \le d_t$ holds on average at time t if*

$$E_{N,X}\left[\ln\frac{Q(x_t+N\mid x)}{Q(x_t\mid x)}\right] \ge E_N\left[\ln\frac{\pi(x_t+N)}{\pi(x_t)}\right] \tag{8}$$

where $d_t = D\Big(\pi(x)\,\Big\|\,Q(x\mid x_t)\Big)$, $d_t(N) = D\Big(\pi(x)\,\Big\|\,Q(x\mid x_t+N)\Big)$, $N \sim f_{N|x_t}(n|x_t)$ is noise that may depend on x_t, and $D\Big(\cdot\,\Big\|\,\cdot\Big)$ is the relative-entropy pseudo-distance: $D\Big(P\,\Big\|\,Q\Big) = \int_X p(x)\ln\Big(\frac{p(x)}{q(x)}\Big)\,dx$.

We next present five corollaries to Theorem 1. Corollary 1 shows that an expectation-based exponential term e^A can weaken the N-MCMC inequality (8) and thereby broaden the theorem’s range of application. The Appendix gives the proof.

Corollary 1. *The N-MCMC noise benefit condition holds if*

$$Q(x_t+n|x) \ge e^A\;Q(x_t|x) \tag{9}$$

for almost all x and n where

$$A = E_N\left[\ln\frac{\pi(x_t+N)}{\pi(x_t)}\right]. \tag{10}$$

Corollary 2 shows that *any* measurable combination $g(x,n)$ of the noise n and state x applies in Corollary 1. So it applies in the N-MCMC Theorem as well. An important case is multiplicative noise injection: $g(x,n) = xn$. We omit the proof because it just replaces $x+n$ with $g(x,n)$ in the proof of Corollary 1.

Corollary 2. *The N-MCMC noise benefit condition holds if*

$$Q(g(x_t,n)|x) \geq e^A\ Q(x_t|x) \tag{11}$$

for almost all x and n where

$$A = E_N\left[\ln\frac{\pi(g(x_t,N))}{\pi(x_t)}\right]. \tag{12}$$

Corollary 3 is a practical result. It shows that the special case of a Gaussian jump density reduces the N-MCMC inequality to a simple quadratic constraint on the noise n. The quadratic condition depends on the Gaussian densities mean and variance σ^2. The similar quadratic condition for the noise-boosted EM algorithm depends only on the mean[38].

Corollary 3. *Suppose* $Q(x_t|x)\ \sim \mathscr{N}\left(x,\sigma^2\right)$. *Then the sufficient noise benefit condition (9) holds if*

$$n(n-2(x_t-x)) \leq -2\sigma^2 A \tag{13}$$

for A in (12).

Corollary 3 shows how noise induces the N-MCMC benefit condition under the Gaussian jump density. The corollary reduces to $n^2+2x \leq -2A = -2E_N\left[\ln\frac{\pi(x_t+N)}{\pi(x_t)}\right] = 2E_N\left[\ln\frac{\pi(x_t)}{\pi(x_t+N)}\right]$ in the standardized case where $\sigma^2 = 1$ and $x_t = 0$. Then

$$n^2+2E_N\left[\ln\pi(x_t+N)\right] \leq b-2x \tag{14}$$

with constant $b = 2\ln\pi(x_t)$. This shows that noise is most effective with small jump samples x. It also suggests a noise upper-bound to satisfy the Corollary 1 sufficient condition.

This condition hinges only on *samples* x and n given an estimate $\widetilde{E_N\left[\ln\pi(x_t+N)\right]}$. This points to a simple naïve implementation of the N-MCMC algorithm that assumes a fixed noise distribution $f_n(n)$ and tunes a multiplicative scaling constant $\alpha\cdot n$. The naïve approach would choose α to meet some predefined acceptance threshold for the noise condition. Corollaries 4 and 5 lead to similar implementations if we substitute the appropriate constraint.

Corollary 4 shows that a Gaussian jump density still gives a simple quadratic noise constraint in the case of *multiplicative* noise injection.

Corollary 4. *Suppose* $Q(x_t|x)\ \sim \mathscr{N}\left(x,\sigma^2\right)$ *and* $g(x_t,n) = nx_t$. *Then the sufficient noise benefit condition (11) holds if*

$$nx_t(nx_t-2x)-x_t(x_t-2x) \leq -2\sigma^2 A \tag{15}$$

for A in (12).

Corollary 5 shows that the jump density need not have finite variance. It shows that infinite-variance Cauchy noise also produces a quadratic constraint on the noise. Cauchy bell curves resemble Gaussian bell curves but have thicker tails. The quadratic constraint depends on the Cauchy noise dispersion (unlike the noisy-EM Cauchy quadratic constraint that depends only on the Cauchy density's location parameter[38]).

Corollary 5. *Suppose* $Q(x_t|x)\ \sim Cauchy(x,d)$. *Then the sufficient condition (9) holds if*

$$n^2+2n(x_t-x) \leq \left(e^{-A}-1\right)\left(d^2+(x_t-x)^2\right) \tag{16}$$

for A in (12).

IV. NOISY SIMULATED ANNEALING

We now show how carefully injected noise can speed convergence of simulated annealing. We will later extend a version of this result to quantum annealing.

Theorem 2 states the Noisy Simulated Annealing (N-SA) Theorem for an annealing temperature schedule $T(t)$ and exponential occupancy probabilities $\pi(x;T) \propto \exp\left(-\frac{C(x)}{T}\right)$. It also gives a simple inequality as a sufficient condition for the speed-up. Its proof uses Jensen's inequality for concave functions and appears in the Appendix. Algorithm 2 in the next section states an annealing algorithms based on the N-SA Theorem. Two corollaries further extend the theorem.

Theorem 2 (Noisy Simulated Annealing Theorem (N-SA))**.** *Suppose* $C(x)$ *is an energy surface with occupancy probabilities* $\pi(x;T) \propto \exp\left(-\frac{C(x)}{T}\right)$. *Then the simulated-annealing noise benefit*

$$E_N\left[\alpha_N(T)\right] \geq \alpha(T) \tag{17}$$

holds on average if

$$E_N\left[\ln\frac{\pi(x_t+N;T)}{\pi(x_t;T)}\right] \geq 0 \tag{18}$$

where $\alpha(T)$ *is the simulated annealing acceptance probability from state* x_t *to the candidate* x^*_{t+1} *that depends on a temperature T (with cooling schedule* $T(t)$*):*

$$\alpha(T) = \min\left\{1, \exp\left(-\frac{\Delta E}{T}\right)\right\} \tag{19}$$

and $\alpha_N(T)$ *is the noisy simulated annealing acceptance probability from state* x_t *to the candidate* $x^*_{t+1}+N$*:*

$$\alpha_N(T) = \min\left\{1, \exp\left(-\frac{\Delta E_N}{T}\right)\right\} \tag{20}$$

where $\Delta E = E^*_{t+1}-E_t = C\left(x^*_{t+1}\right)-C(x_t)$ *is the energy difference of states* x^*_{t+1} *and* x_t *and* $\Delta E_N = E^*_{N,t+1}-E_t = C\left(x^*_{t+1}+N\right)-C(x_t)$ *is the energy difference of states* $x^*_{t+1}+N$ *and* x_t .

Two important annealing corollaries follow from Theorem 2. The first corollary allows the acceptance probability $\beta(T)$ to depend on any increasing convex function m of the occupancy probability ratio. The proof also relies on Jensen's inequality and appears in the Appendix.

Corollary 6. *Suppose m is a convex increasing function. Then the N-SA Theorem noise benefit*

$$E_N\left[\beta_N(T)\right] \geq \beta(T) \tag{21}$$

holds on average if

$$E_N\left[\ln\frac{\pi(x_t+N;T)}{\pi(x_t;T)}\right] \geq 0 \tag{22}$$

where β is the acceptance probability from state x_t to the candidate x^*_{t+1}:

$$\beta(T) = \min\left\{1, m\left(\frac{\pi\left(x^*_{t+1};T\right)}{\pi(x_t;T)}\right)\right\}. \tag{23}$$

and β_N is the noisy acceptance probability from state x_t to the candidate $x^*_{t+1}+N$:

$$\beta_N(T) = \min\left\{1, m\left(\frac{\pi\left(x^*_{t+1}+N;T\right)}{\pi(x_t;T)}\right)\right\}. \tag{24}$$

Corollary 7 gives a simple inequality condition for the noise benefit in the N-SA Theorem when the occupancy probability $\pi(x)$ has a softmax or Gibbs form of an exponential normalized with a partition function or integral of exponentials. Its proof also appears in the Appendix.

Corollary 7. *Suppose $\pi(x) = Ce^{g(x)}$ if C is the normalizing constant $C = \frac{1}{\int_X e^{g(x)}\,dx}$. Then there is an N-SA noise benefit if*

$$E_N\left[g(x_t+N)\right] \geq g(x_t). \tag{25}$$

V. NOISY MCMC ALGORITHMS AND RESULTS

We now present algorithms for noisy MCMC and noisy simulated annealing. We follow each with applications of the algorithms and results that show improvements over existing noiseless algorithms.

A. The Noisy MCMC Algorithms

This section presents two noisy variants of MCMC algorithms. Algorithm 1 extends Metropolis-Hastings MCMC for sampling. Algorithm 2 describes how to use noise to benefit stochastic optimization with simulated annealing.

Algorithm 1 The Noisy Metropolis Hastings Algorithm

```
procedure NoisyMetroloplisHastings($X$)
    $x_0 \leftarrow$ Initial$(X)$
    for $t \leftarrow 0, N$ do
        $x_{t+1} \leftarrow$ Sample$(x_t)$
procedure Sample($x_t$)
    $x^*_{t+1} \leftarrow x_t +$ JumpQ$(x_t) +$ Noise$(x_t)$
    $\alpha \leftarrow \frac{\pi(x^*_{t+1})}{\pi(x_t)}$
    if $\alpha > 1$ then
        return $x^*_{t+1}$
    else if Uniform$[0,1] < \alpha$ then
        return $x^*_{t+1}$
    else
        return $x_t$
procedure JumpQ($x_t$)
    return $y \sim Q(y|x_t)$
procedure Noise($x_t$)
    return $y \sim f(y|x_t)$
```

Algorithm 2 The Noisy Simulated Annealing Algorithm

```
procedure NoisySimulatedAnnealing($X$, $T_0$)
    $x_0 \leftarrow$ Initial$(X)$
    for $t \leftarrow 0, N$ do
        $T \leftarrow Temp(t)$
        $x_{t+1} \leftarrow$ Sample$(x_t, T)$
procedure Sample($x_t$, $T$)
    $x^*_{t+1} \leftarrow x_t +$ JumpQ$(x_t) +$ Noise$(x_t)$
    $\alpha \leftarrow \pi\left(x^*_{t+1}\right) - \pi(x_t)$
    if $\alpha \leq 0$ then
        return $x^*_{t+1}$
    else if Uniform$[0,1] < \exp(-\alpha/T)$ then
        return $x^*_{t+1}$
    else
        return $x_t$
procedure JumpQ($x_t$)
    return $y \sim Q(y|x_t)$
procedure Noise($x_t$)
    return $y \sim f(y|x_t)$
```

B. Noise improves complex optimization

The first simulation produced a noise benefit in simulated annealing on a complex cost function. The Schwefel function[48] is a standard optimization benchmark because it has many local minima and has the unique global minimum

$$f(x) = 419.9829d - \sum_{i=1}^{d} x_i \sin\left(\sqrt{|x_i|}\right) \tag{26}$$

where d is the dimension over the hypercube $-500 \leq x_i \leq 500$ for $i = 1, \ldots, d$. The Schwefel function has a single global minimum $f(x_{min}) = 0$ at $x_{min} = (420.9687, \ldots, 420.9687)$. Figure 1 shows a representation of the surface for $d = 2$.

The simulation used a zero-mean Gaussian jump density with $\sigma_{jump} = 5$ and thus with variance $\sigma^2_{jump} = 25$. It also used a zero-mean Gaussian noise density with $0 < \sigma_{noise} \leq 5$. Figure 5.(a) shows that noisy simulated annealing in accord with Theorem 2 converged 76% faster than did standard noiseless simulated annealing when using log-cooling. Figure 5.(b) shows that the estimated global minimum from noisy simulated annealing was almost 2 orders of magnitude better than non-noisy simulations on average (0.05 vs 4.6). The simulation annealed a 5-dimensional Schwefel surface. It estimated the minimum energy configuration and averaged the result over 1000 trials. We defined the convergence time as the number of steps that the simulation required to reach the global minimum energy within 10^{-3}:

$$|f(x_t) - f(x_{min})| \leq 10^{-3}. \tag{27}$$

Figure 2 shows projections of trajectories from a simulation without noise (a) and a simulation with noise (b). We initialized each simulation with the same x_0. The figure shows the global minimum circled in red (lower left). It shows that noisy simulated annealing boosted the sequences through more local minima while the no-noise simulation could not escape cycling between three local minima.

Figure 5.(c) shows that the noise decreased the failure rate of the simulation. We defined a failed simulation as a simulation that did not converge before $t < 10^7$. Noiseless simulations produced the failure rate 4.5%. Even moderate noise reduced the failure rate to less than 1 in 200 (< 0.5%).

Figure 6 shows that appropriate noise also boosted simulated annealing with accelerated cooling schedules. Noise reduced convergence time by 40.5% under exponential cooling and 32.8% under geometric cooling. The simulations attained comparable solution error and failure rate (0.05%) across all noise levels. So we have omitted the corresponding figures.

C. Noise speeds Lennard-Jones 12-6 simulations

The second simulation shows a noise benefit in an MCMC molecular dynamics model. This model used the noisy Metropolis-Hastings algorithm (Algorithm 1) to search a 24-dimensional energy landscape. It used the Lennard-Jones 12-6 potential well to model the pairwise interactions between an 8 argon atom gas.

The Lennard Jones (12-6) potential well approximates the interaction energy between two neutral atoms[23,24,45]

$$V_{LJ} = \epsilon\left[\left(\frac{r_m}{r}\right)^{12} - 2\left(\frac{r_m}{r}\right)^{6}\right] \tag{28}$$

$$= 4\epsilon\left[\left(\frac{\sigma}{r}\right)^{12} - \left(\frac{\sigma}{r}\right)^{6}\right] \tag{29}$$

where ϵ is the depth of the potential well, r is the distance between the two atoms, r_m is the interatomic distance corresponding to the minimum energy, and σ is the zero potential interatomic distance. Figure 7 shows how the two terms interact to form the energy surface: (1) the *12*-term dominates at short distances since overlapping electron orbitals cause strong Pauli repulsion to push the atoms apart and (2) the *6*-term dominates at longer distances because van der Waals and dispersion forces pull the atoms toward a finite equilibrium distance r_m. Table I shows the value of the Lennard-Jones parameters for argon.

TABLE I. Argon Lennard-Jones 12-6 parameters

ϵ	1.654×10^{-21} J
σ	3.405×10^{-10} m
r_m	3.821 Å

The simulation estimated the minimum energy coordinates for 8 argon atoms in 3 dimensions. We performed 200 trials at each noise level. We summarized each trial as the average number of steps to estimate the minimum energy within 10^{-2}.

Figure 8 shows that noise produced a 42% reduction in convergence time over the non-noisy simulation in the 8-argon-atom system. The simulation found the global noise optimum at a noise variance of $\sigma^2 = 0.56$. We found this optimal noise value through repeated trial and error. The N-MCMC theorems guarantee only that noise will improve system performance on average if the noise obeys the N-MCMC inequality. The results do not directly show how to find the optimal noise value.

VI. QUANTUM SIMULATED ANNEALING

Quantum annealing (QA) uses quantum perturbations to evolve the system state in accord with the quantum Hamiltonian[2,6,47]. Classical simulated annealing instead evolves the system with thermodynamic excitations.

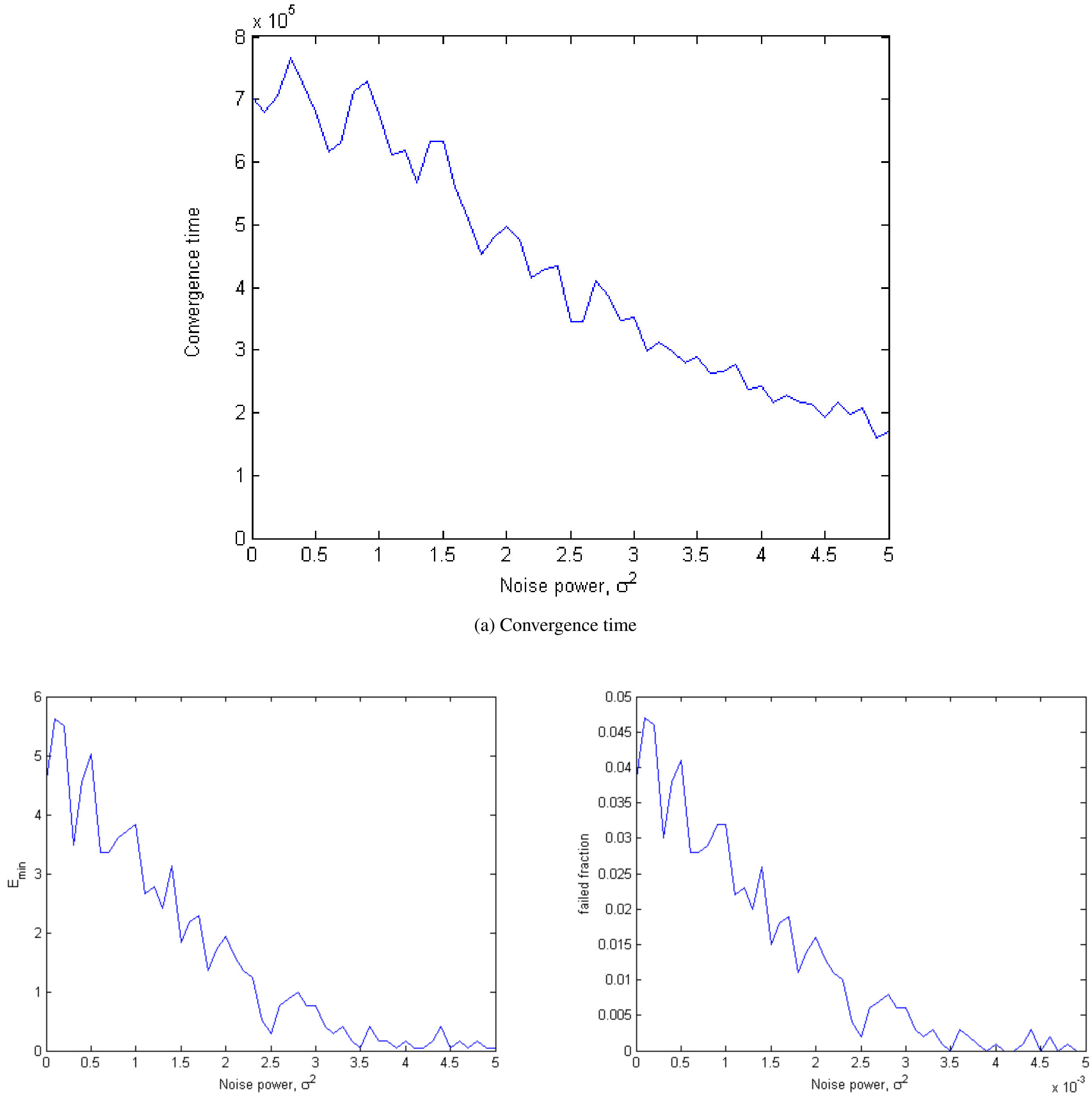
Simulated QA uses an MCMC framework to simulate draws from the square magnitude of the wave function $\Psi(r,t)$ instead of solving the time-dependent Schrödinger equation:

$$i\hbar\frac{\partial}{\partial t}\Psi(r,t) = \left[\frac{-\hbar^2}{2\mu}\nabla^2 + V(r,t)\right]\Psi(r,t) \tag{30}$$

where μ is the particle's *reduced mass*, V is the potential energy, and ∇^2 is the Laplacian operator of appropriately summed second partial derivatives of the spatial variables.

The acceptance probability is proportional to the ratio of a function of the energy of the old and new states in classical simulated annealing . This discourages beneficial hops if there are energy peaks between minima. QA uses probabilistic tunneling to allow occasional jumps *through* high energy peaks.

Ray and Chakrabarti[43] recast Kirkpatrick's thermodynamic simulated annealing using quantum fluctuations to drive the state transitions. The resulting QA algorithm uses a transverse magnetic field Γ in place of the temperature T in classical simulated annealing. Then the strength of the magnetic field governs the transition probability between system states. The adiabatic theorem[20] ensures that the system remains near the ground state during slow changes of the field strength.



(a) Convergence time

(b) Minimum energy

(c) Failure rate

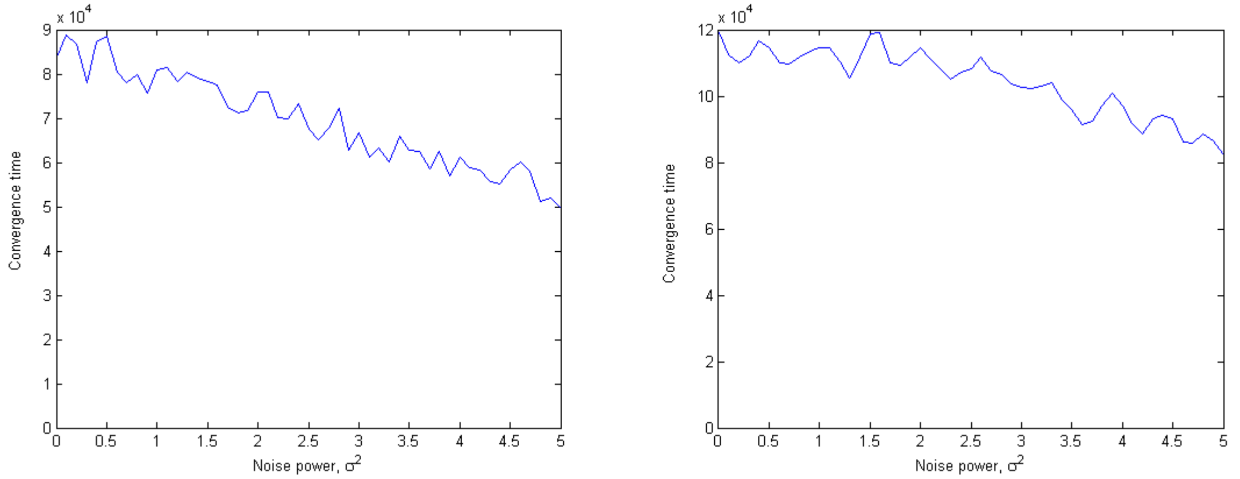
FIG. 5. Simulated-annealing noise benefits with a 5-dimensional Schwefel energy surface and a log cooling schedule. Noise improved three distinct performance metrics when using Algorithm 2. (a) Noise reduced convergence time by 76%. We defined convergence time as the number of steps the simulation took to estimate the global minimum energy with error $< 10^{-3}$. Simulations with faster convergence will in general find better estimates given the same computational time. (b) Noise improved the estimated minimum system energy by 2 orders of magnitude in simulations with a fixed run time ($t_{max} = 10^6$). Figure 2 shows how the estimated minimum corresponds to samples. Noise increased the breadth of the search and pushed the simulation to make *good* jumps toward new minima. (c) Noise decreased the likelihood of failure in a given trial by almost 100%. We defined a simulation *failure* if it did not converge by $t = 10^7$. This was about 20 times longer than the average convergence time. 4.5% of noiseless simulations failed. The simulation produced no sign of failure except an increased estimated variance between trials. Noisy simulated annealing produced only 2 failures in 1000 trials (0.2%).

The adiabatic Hamiltonian evolves smoothly from the transverse magnetic dominance to the Edwards-Anderson Hamiltonian:

$$H(t) = \left(1 - \frac{t}{T}\right) H_0 + \frac{t}{T} H_P. \qquad (31)$$

This evolution leads to the minimum energy configuration of the underlying potential energy surface as time t approaches a fixed large value T.

QA outperforms classical simulated annealing in cases where the potential energy landscape contains many high but



(a) Exponential cooling schedule

(b) Geometric cooling schedule

FIG. 6. Noise benefits decreased convergence time under accelerated cooling schedules. Simulated annealing algorithms often use accelerated cooling schedules such as exponential cooling $T_{exp}(t) = T_0 \cdot A^t$ or geometric cooling $T_{geom}(t) = T_0 \cdot \exp\left(-At^{1/d}\right)$ where $A < 1$ and T_0 are user parameters and d is the sample dimension. Accelerated cooling schedules do not have convergence guarantees as do log cooling $T_{log}(t) = \log(t+1)$ but often provide better estimates given a fixed run time. Noise enhanced simulated annealing reduced convergence time under an (a) exponential cooling schedule by 40.5% and under a (b) geometric cooling schedule by 32.8%.

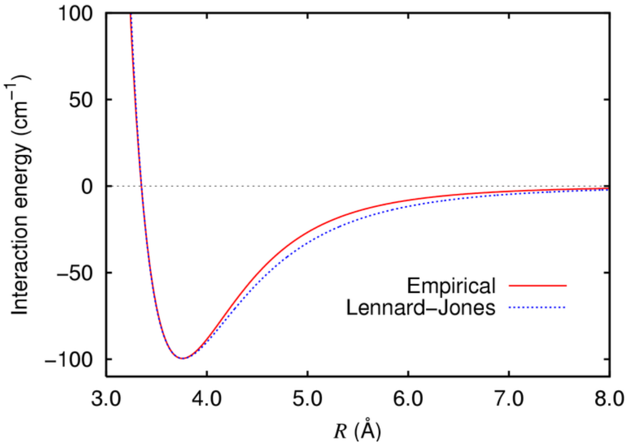


FIG. 7. The Lennard-Jones 12-6 potential well approximated pairwise interactions between two neutral atoms. The figure shows the energy of a two-atom system as a function of the interatomic distance. The well resulted from two competing atomic effects: (1) overlapping electron orbitals caused strong Pauli repulsion to push the atoms apart at short distances and (2) van der Waals and dispersion attractions pulled the atoms together at longer distances. Three parameters characterized the potential: (1) ϵ was the depth of the potential well, (2) r_m was the interatomic distance corresponding to the minimum energy, and (3) σ was the zero-potential interatomic distance. Table I lists parameter values for argon.

thin energy barriers between shallow local minima[43]. QA is well suited to problems in discrete search spaces that have vast numbers of local minima. A good example is finding the ground state of an Ising spin glass. Lucas recently found Ising formulations for Karp's 21 NP-complete problems[25]. The NP-complete problems include such optimization benchmarks as graph-partitioning, calculating an exact cover, integer weight knapsack packing, graph coloring, and the traveling salesman problem. NP-complete problems are a special class of decision problem that have time complexity super-polynomial (NP-hard) to the input size but only polynomial time to verify the solution (NP). Advances by D-Wave Systems have brought quantum annealers to market and shown how adiabatic quantum computers can have real-world applications[49].

Spin glasses are systems with localized magnetic moments. *Quenched disorder* characterizes the steady-state interactions between atomic moments. Thermal fluctuations drive changes within the system. Ising spin-glass models use a two-dimensional or three-dimensional lattice of discrete variables to represent the coupled dipole moments of atomic spins. The discrete variables take one of two values: +1 (*up*) or -1 (*down*). The two-dimensional square-lattice Ising model is one of the simplest statistical models that shows a phase transition.

Simulated QA for an Ising spin glass usually applies the Edwards-Anderson model Hamiltonian with a transverse magnetic field $J^{\perp}$

$$\mathbf{H} = \mathbf{U} + \mathbf{K} = -\sum_{\langle ij \rangle} J_{ij} s_i s_j - J^{\perp} \sum_i s_i. \qquad (32)$$

The transverse field $J^{\perp}$ and classical Hamiltonian J_{ij} have a nonzero commutator in general:

$$\left[\mathbf{J}^{\perp}, \mathbf{J_{ij}}\right] \neq \mathbf{0} \qquad (33)$$

for the commutator operator $[\mathbf{A}, \mathbf{B}] = \mathbf{AB} - \mathbf{BA}$.

The path-integral Monte Carlo method is a standard QA method[27] that uses the Trotter ("break-up") approximation for

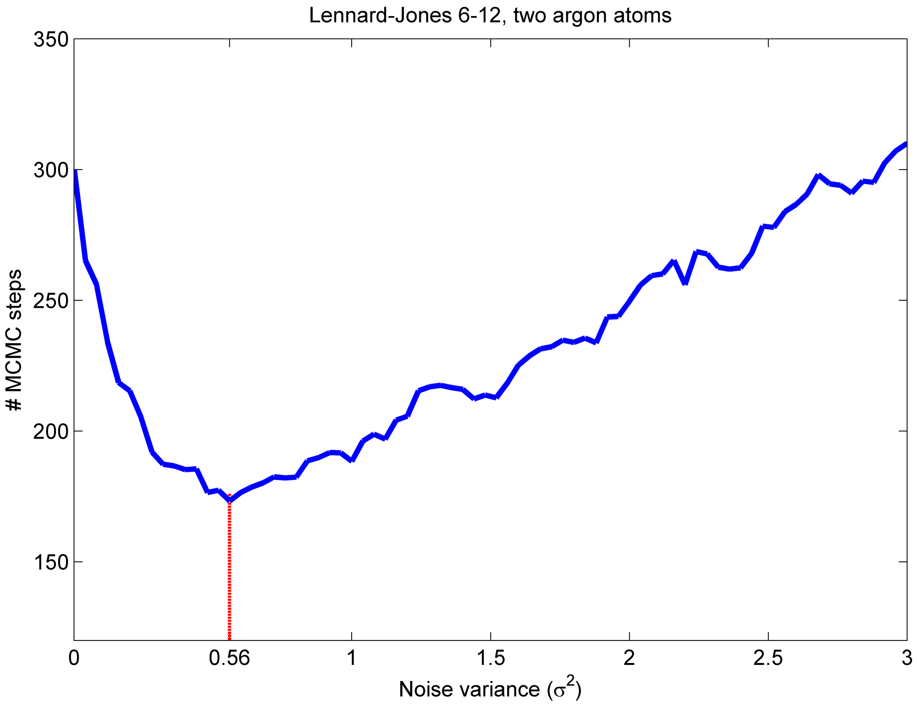


FIG. 8. MCMC noise benefit for an MCMC molecular dynamics simulation of 8 argon atoms. Noise decreases the convergence time for an MCMC simulation to find the energy minimum by 42%. The plot shows the number of steps that an MCMC simulation needed to converge to the minimum energy in an eight-argon-atom gas system. The optimal noise had a standard deviation of 0.56. The plot shows 100 noise levels with variance noise powers that range between 0 (no noise) and $\sigma^2 = 3$. Each point averaged 200 simulations and shows the average number of MCMC steps required to estimate the minimum to within 0.01. The Lennard-Jones 12-6 model described the interaction between two argon atoms with $\epsilon = 1.654 \times 10^{-21} J$ and $\sigma = 3.405 \times 10^{-10} m = 3.405$Å[45].

quantum operators that do not commute:

$$e^{-\beta(\mathbf{K}+\mathbf{U})} \approx e^{-\beta\mathbf{K}} e^{-\beta\mathbf{U}} \tag{34}$$

where $[\mathbf{K}, \mathbf{U}] \neq \mathbf{0}$ and $\beta = \frac{1}{k_B T}$. Then the Trotter approximation estimates the partition function Z:

$$\begin{aligned} Z &= \mathrm{Tr}\left(e^{-\beta\mathbf{H}}\right) && (35)\\ &= \mathrm{Tr}\left(\exp\left[-\frac{\beta(\mathbf{K}+\mathbf{U})}{P}\right]\right)^P && (36)\\ &= \sum_{s^1}\cdots\sum_{s^P}\left\langle s^1|e^{-\beta(\mathbf{K}+\mathbf{U})/P}|s^2\right\rangle \\ &\qquad \times\left\langle s^2|e^{-\beta(\mathbf{K}+\mathbf{U})/P}|s^3\right\rangle\times\cdots\times\left\langle s^P|e^{-\beta(\mathbf{K}+\mathbf{U})/P}|s^1\right\rangle && (37)\\ &\approx C^{NP}\sum_{s^1}\cdots\sum_{s^P} e^{-\frac{\mathbf{H_{d+1}}}{PT}} && (38)\\ &= Z_P && (39) \end{aligned}$$

where N is the number of lattice sites in the d-dimensional Ising lattice, P is the *Trotter number* of imaginary-time slices,

$$C = \sqrt{\frac{1}{2}\sinh\left(\frac{2\Gamma}{PT}\right)} \tag{40}$$

and

$$\mathbf{H_{d+1}} = -\sum_{k=1}^{P}\left(\sum_{\langle ij\rangle} J_{ij} s_i^k s_j^k + J^{\perp}\sum_{i} s_i^k s_i^{k+1}\right) \tag{41}$$

where Γ is the transverse field strength and $s^{P+1} = s^1$ to satisfy periodic bounding conditions. The temperature T in the exponent of (38) absorbs the β coefficient because in Planck units the Boltzmann coefficient k_B is $k_B = 1$. So $T = \frac{1}{k_B\beta} = \frac{1}{\beta}$.

A Trotter slice subdivides the system's evolution into short-time intervals. Then the system Hamiltonian is approximately time-independent and includes an error term. The product PT in (38) determines the spin replica couplings both between neighboring Trotter slices and between the spins within slices.

Shorter simulations did not show a strong dependence on the number of Trotter slices P. This is likely because shorter simulations spend relatively less time under the lower transverse magnetic field to induce strong coupling between the slices. So the Trotter slices tend to behave more independently than if they evolved under the increased coupling from longer simulations.

High Trotter numbers ($N = 40$) show substantial improvements for very long simulations. Martoňák[27] compared high Trotter simulations to classical annealing. The computations showed that path-integral QA gave a relative speed-up of four orders of magnitude over classical annealing: "one can calculate using path-integral quantum annealing in one day what

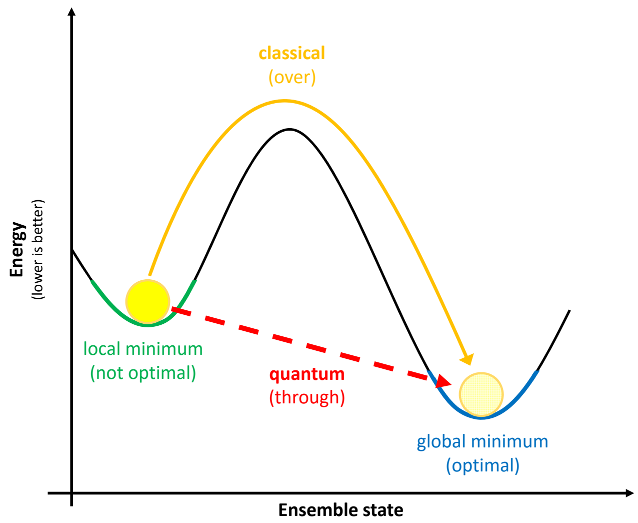


FIG. 9. Quantum annealing (QA) uses quantum tunneling to burrow *through* energy peaks (yellow). Classical simulated annealing (SA) instead generates a sequence of states to scale the peak (blue). The figures shows that a local minimum has trapped the estimate (green). SA would require a sequence of unlikely jumps to scale the potential energy hill. This may not be realistic at low SA temperatures. So the estimate gets trapped in the suboptimal valley. QA uses quantum tunneling to *burrow* through the mountain. This explains why QA can give better estimates than SA gives while optimizing complex potential energy surfaces that contain many high energy states.

would be obtained by plain classical annealing in about 30 years.”

A. The Noisy Quantum Simulated Annealing Algorithm

This section develops a noise-boosted version of path-integral simulated QA. Algorithm 3 presents pseudo-code for the Noisy QA Algorithm.

The Noisy QA Algorithm describes how to advance the state of a quantum Ising model forward in time in a heat bath and under the effect of a perturbing transverse magnetic field. The algorithm represents the qubit system at a given time with state variable X_t. A noise power parameter captures the action of excess quantum effects in the system. This lets external noise produce further spin transitions along coupled Trotter slices. The noise power parameter is similar to the temperature parameter in classical simulated annealing. It describes an increase in the chance of temporary transitions to higher energy states.

The Noisy QA Algorithm uses subscripts to denote the spin index s within the Ising lattice and slice number l between the Trotter lattices. We restrict the fully indexed value $X_t[s,l]$ to -1 and +1 to represent spin-up and spin-down alignments in the spin network. The algorithm advances the time index t to follow the simulation in time. The algorithm updates the transverse magnetic field strength Γ and Trotter-slice lattice coupling $J^{\perp}$ at each step. These proxy values describe the quantum coupling inherent in the system. High values ensure that the system will tunnel through high-energy intermediate states. These constants decrease as the simulation advances. So they resemble the decreasing temperature in classical simulated annealing.

The Noisy QA Algorithm computes the energy of each spin on each Trotter slice as in the standard path-integral quantum annealing. The algorithm does this for each time step. The algorithm computes the local energy between the spin and each of its neighbors in the Ising spin network. It does this for each spin along the Trotter slices in accord with the Hamiltonian

$$H = -\sum_k \left(\sum_{i,j} J_{i,j} s_i^k s_j^k + J^{\perp} \sum_i s_i^k s_i^{k+1} \right). \tag{42}$$

The Noisy QA algorithm then flips spins under one of three conditions:

1. if $E > 0$
2. if $\alpha < e^{E/T}$ where $\alpha = Uniform[0,1]$
3. if the energies satisfy a noise-based inequality

Conditions 1 and 2 describe the standard path-integral quantum annealing spin-flip conditions. The algorithm flips only the currently indexed spin under these two conditions. Condition 3 enables spin-flips among Trotter neighbors. The probability of the flip depends on the relative energy of the Trotter neighbors and on a noise-based inequality. The system then accepts either toggles if they reduce the overall system energy.

Condition 3 is analogous to generating candidate “jump” values in classical simulated annealing. The spin flip along the Trotter slices in Figure 10 is analogous to accepting the candidate jump state in classical simulated annealing. The algorithm checks the noise-based inequality as follows. It first draws a uniform random variable. It then compares this uniform value to the simulation-wide threshold parameter called the `NoisePower`. Standard path-integral quantum annealing corresponds to `NoisePower = 0`.

The Noisy Quantum Annealing Algorithm uses Trotter neighbors to bias an operator average toward lower energy configurations. The Trotter formalism treats each particle in a physical system by a ring of P equivalent particles that interact through *harmonic* springs. The average of an observable $\mathscr{O}$ becomes an average of the operator $\mathscr{O}$ on each Trotter slice. Standard path-integral quantum annealing computes local energies *within* each Trotter slice and then updates the particle state according to conditions 1 and 2 above. The noisy QA algorithm biases the operator average by allowing nodes in meta-stable energy configurations to affect Trotter neighbors *between* slices.

Algorithm 3 The Noisy Quantum Annealing Algorithm

```
procedure NoisySimulatedQuantumAnnealing($X$, $\Gamma_0$, $P$, $T$)
    $x_0 \leftarrow$ Initial $(X)$
    for $t \leftarrow 0, N$ do
        $\Gamma \leftarrow TransverseField(\Gamma_0, t)$
        $J^{\perp} \leftarrow TrotterScale(P, T, \Gamma)$
        for all Trotter slices l do
            for all spins s do
                $x_{t+1}[l, s] \leftarrow$ Sample $(x_t, J^{\perp}, s, l)$
procedure TrotterScale($P$, $T$, $\Gamma$)
    return $\frac{PT}{2} \ln \tanh\left(\frac{\Gamma}{PT}\right)$
procedure Sample($x_t$, $J^{\perp}$, $s$, $l$)
    $E \leftarrow LocalEnergy(J^{\perp}, x_t, s, l)$
    if $E > 0$ then
        return $-x_t[l, s]$
    else if Uniform $[0, 1] < \exp(E/T)$ then
        return $-x_t[l, s]$
    else
        if $Uniform[0, 1] < NoisePower$ then
            $E^{+} \leftarrow LocalEnergy(J^{\perp}, x_t, s, l + 1)$
            $E^{-} \leftarrow LocalEnergy(J^{\perp}, x_t, s, l - 1)$
            if $E > E^{+}$ then
                $x_{t+1}[l + 1, s] \leftarrow -x_t[l + 1, s]$
            if $E > E^{-}$ then
                $x_{t+1}[l - 1, s] \leftarrow -x_t[l - 1, s]$
        return $x_t[l, s]$
```

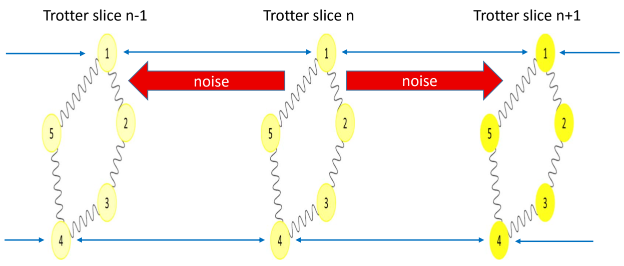


FIG. 10. The noisy quantum annealing algorithm propagates noise along a Trotter ring. The algorithm inspects the local energy landscape after each time step. It then injects noise by conditionally flipping the spin of neighbors. These spin flips diffuse the noise across the network because quantum correlations between neighbors encourage convergence to the optimal solution.

B. Noise improves quantum MCMC

The third simulation shows a noise benefit in simulated quantum annealing[2,6,47]. It shows that noise that obeys a condition similar to the N-MCMC theorem improves the ground-state energy estimate.

We used path-integral Monte Carlo quantum annealing[27] to calculate the ground state of a randomly coupled 1024-bit (32x32) Ising quantum spin system. The simulation used 20 Trotter slices to approximate the quantum coupling at temperature $T = 0.01$. It used 2-D periodic horizontal and vertical boundary conditions (toroidal boundary conditions) with coupling strengths J_{ij} drawn from Uniform $[-2, 2]$.

Each trial used random initial spin states ($s_i \in -1, 1$). We used 100 pre-annealing steps to cool the simulation from an initial temperature $T_0 = 3$ to $T_q = 0.01$. The quantum annealing linearly reduced the transverse magnetic field $\Gamma_0 = 1.5$ to $\Gamma_{final} = 10^{-8}$ over 100 steps. We performed a Metropolis-Hastings pass for each lattice across each Trotter slice after each update. We maintained $T_q = 0.01$ for the entirety of the quantum annealing. The simulation used the standard slice coupling between Trotter lattices:

$$J^{\perp} = \frac{PT}{2} \ln \tanh\left(\frac{\Gamma_t}{PT}\right) \tag{43}$$

where Γ_t is the current transverse field strength, P is the number of Trotter slices, and $T = 0.01$.

The simulation injected noise into the model using a *power parameter* $0 < p < 1$. The algorithm extended the Metropolis-Hastings test to each lattice-site by conditionally flipping the corresponding site on coupled trotter slices.

We benchmarked the results against the true ground state $E_0 = -1591.92$[36]. Figure 3 shows that noise that obeys the N-MCMC benefit condition improved the ground-state solution by 25.6%. This reduced simulation time by several orders of magnitude since the estimated ground state largely converged by the end of the simulation. We did not quantify the decrease in convergence time because the non-noisy quantum annealing algorithm did not converge near the noisy quantum annealing estimate during any trial.

Figure 3 also shows that the noise benefit is not a simple diffusive benefit. We computed for each trial the result of *blind* noise by injecting noise identical to the above but noise that did not have to satisfy the N-MCMC condition. Figure 3 shows that such blind noise *reduced* the accuracy of the ground-state estimate by 41.6%.

VII. CONCLUSION

Noise can speed MCMC convergence in reversible Markov chains that are aperiodic and irreducible. The noise must satisfy an inequality that depends on the reversibility of the Markov chain. Simulations showed that such noise also improved the breadth of such simulation searches for deep local minima. This noise-boosting of the Metropolis-Hastings algorithm does not require symmetric jump densities. Nor need the jump densities have finite variance.

Carefully injected noise can also improve quantum annealing where the noise flips spins among Trotter neighbors. Other forms of quantum noise injection should produce a noise boost if the N-MCMC or noisy-annealing inequalities hold at least approximately.

The proofs that the noise boosts hold for Gaussian and Cauchy jump densities suggest that the more general family of symmetric stable thick-tailed bell-curve densities[35,55] should also produce noise-boosted MCMC and annealing with varying levels of jump impulsiveness.

REFERENCES

[1] Olaoluwa Adigun and Bart Kosko. Using noise to speed up video classification with recurrent backpropagation. In *2017 International Joint Conference on Neural Networks (IJCNN)*, pages 108–115. IEEE, 2017.

[2] Tameem Albash and Daniel A Lidar. Demonstration of a scaling advantage for a quantum annealer over simulated annealing. *Physical Review X*, 8(3):031016, 2018.

[3] Kartik Audhkhasi, Osonde Osoba, and Bart Kosko. Noise benefits in backpropagation and deep bidirectional pre-training. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–8. IEEE, 2013.

[4] Kartik Audhkhasi, Osonde Osoba, and Bart Kosko. Noisy hidden Markov models for speech recognition. In *Neural Networks (IJCNN), The 2013 International Joint Conference on*, pages 1–6. IEEE, 2013.

[5] Kartik Audhkhasi, Osonde Osoba, and Bart Kosko. Noise-enhanced convolutional neural networks. *Neural Networks*, 78:15–23, 2016.

[6] Sergio Boixo, Troels F Rønnow, Sergei V Isakov, Zhihui Wang, David Wecker, Daniel A Lidar, John M Martinis, and Matthias Troyer. Evidence for quantum annealing with more than one hundred qubits. *Nature Physics*, 10(3):218, 2014.

[7] Steve Brooks, Andrew Gelman, Galin Jones, and Xiao-Li Meng. *Handbook of Markov Chain Monte Carlo*. CRC press, 2011.

[8] François Chapeau-Blondeau and David Rousseau. Noise-enhanced performance for an optimal bayesian estimator. *IEEE Transactions on Signal Processing*, 52(5):1327–1334, 2004.

[9] Mary Kathryn Cowles and Bradley P. Carlin. Markov chain Monte Carlo convergence diagnostics: A comparative review. *Journal of the American Statistical Association*, 91(434):883–904, June 1996.

[10] Arthur P Dempster, Nan M Laird, and Donald B Rubin. Maximum likelihood from incomplete data via the em algorithm. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(1):1–22, 1977.

[11] Johannes M. Dieterich. Empirical Review of Standard Benchmark Functions Using Evolutionary Global Optimization. *Applied Mathematics*, 03(October):1552–1564, 2012.

[12] Bradley Efron and Trevor Hastie. *Computer age statistical inference*, volume 5. Cambridge University Press, 2016.

[13] William Feller. *An introduction to probability theory and its applications*, volume 2. John Wiley & Sons, Inc., New York, NY, USA, 2nd edition, 2008.

[14] Brandon Franzke and Bart Kosko. Noise can speed convergence in Markov chains. *Physical Review E*, 84(4):041112, 2011.

[15] Luca Gammaitoni, Peter Hänggi, Peter Jung, and Fabio Marchesoni. Stochastic resonance. *Reviews of modern physics*, 70(1):223, 1998.

[16] S. Geman and D. Geman. Stochastic relaxation, Gibbs distributions, and the Bayesian restoration of images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-6:721–741, 1984.

[17] Charles J Geyer. Practical markov chain monte carlo. *Statistical science*, pages 473–483, 1992.

[18] W. R. Gilks, Walter R. Gilks, Sylvia Richardson, and D. J. Spiegelhalter. *Markov chain Monte Carlo in practice*. CRC Press, 1996.

[19] W. K. Hastings. Monte Carlo sampling methods using Markov chains and their applications. *Biometrika*, 57:97–109, 1970.

[20] Tosio Kato. On the adiabatic theorem of quantum mechanics. *Journal of the Physical Society of Japan*, 5(6):435–439, 1950.

[21] Claude Kipnis and SR Srinivasa Varadhan. Central limit theorem for additive functionals of reversible markov processes and applications to simple exclusions. *Communications in Mathematical Physics*, 104(1):1–19, 1986.

[22] Scott Kirkpatrick, Mario P. Vecchi, and C. D. Gelatt. Optimization by simulated annealing. *Science*, 220(4598):671–680, 1983.

[23] John Edward Lennard-Jones. On the determination of molecular fields. i. from the variation of the viscosity of a gas with temperature. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 106(738):441–462, 1924.

[24] John Edward Lennard-Jones. On the determination of molecular fields. ii. from the equation of state of a gas. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 106(738):463–477, 1924.

[25] Andrew Lucas. Ising formulations of many np problems. *Frontiers in Physics*, 2(February):1–15, 2014.

[26] Roman Martoňák, Giuseppe E Santoro, and Erio Tosatti. Quantum annealing by the path-integral monte carlo method: The two-dimensional random ising model. *Physical Review B*, 66(9):094203, 2002.

[27] Roman Martoňák, Giuseppe Santoro, and Erio Tosatti. Quantum annealing by the path-integral Monte Carlo method: The two-dimensional random Ising model. *Physical Review B*, 66(9):1–8, 2002.

[28] Mark D McDonnell, Nigel G Stocks, Charles EM Pearce, and Derek Abbott. *Stochastic resonance*. 2008.

[29] N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. Equations of state calculations by fast computing machines. *Journal of Chemical Physics*, 21:1087–1091, 1953.

[30] Sean Meyn and Richard L. Tweedie. *Markov Chains and Stochastic Stability*. Cambridge University Press, 2nd edition, 2009.

[31] Sanya Mitaim and Bart Kosko. Adaptive stochastic resonance. *Proceedings of the IEEE*, 86(11):2152–2183, 1998.

[32] Sanya Mitaim and Bart Kosko. Noise-benefit forbidden-interval theorems for threshold signal detectors based on cross correlations. *Physical Review E*, 90(5):052124, 2014.

[33] H. Mühlenbein, M. Schomisch, and J. Born. The parallel genetic algorithm as function optimizer. *Parallel Computing*, 17(6-7):619–632, September 1991.

[34] K Murali, Sudeshna Sinha, William L Ditto, and Adi R Bulsara. Reliable logic circuit elements that exploit nonlinearity in the presence of a noise floor. *Physical review letters*, 102(10):104101, 2009.

[35] Chrysostomos L Nikias and Min Shao. *Signal processing with alpha-stable distributions and applications*. Wiley-Interscience, 1995.

[36] University of Cologne. Spin glass server.

[37] Osonde Osoba and Bart Kosko. Noise-enhanced clustering and competitive learning algorithms. *Neural Networks*, 37:132–140, 2013.

[38] Osonde Osoba and Bart Kosko. The noisy expectation-maximization algorithm for multiplicative noise injection. *Fluctuation and Noise Letters*, page 1650007, 2016.

[39] Osonde Osoba, Sanya Mitaim, and Bart Kosko. The noisy expectation–maximization algorithm. *Fluctuation and Noise Letters*, 12(03), 2013.

[40] Ashok Patel and Bart Kosko. Stochastic resonance in continuous and spiking neuron models with levy noise. *IEEE Transactions on Neural Networks*, 19(12):1993–2008, 2008.

[41] Ashok Patel and Bart Kosko. Optimal noise benefits in neyman–pearson and inequality-constrained statistical signal detection. *IEEE Transactions on Signal Processing*, 57(5):1655–1669, 2009.

[42] Ashok Patel and Bart Kosko. Noise benefits in quantizer-array correlation detection and watermark decoding. *IEEE Transactions on Signal Processing*, 59(2):488–505, 2011.

[43] P. Ray, B. K. Chakrabarti, and Arunava Chakrabarti. Sherrington-kirkpatrick model in a transverse field: Absence of replica symmetry breaking due to quantum fluctuations. *Phys. Rev. B*, 39(16):11828–11832, 1989.

[44] Christian P Robert and George Casella. *Monte Carlo statistical methods (Springer Texts in Statistics)*. Springer-Verlag, 2nd edition, 2005.

[45] L. A. Rowley, D. Nicholson, and N. G. Parsonage. Monte Carlo grand canonical ensemble calculation in a gas-liquid transition region for 12-6 argon. *Journal of Computational Physics*, 17(4):401–414, 1975.

[46] Walter Rudin. *Real and complex analysis*. McGraw-Hill, 2006.

[47] Giuseppe E Santoro, Roman Martoňák, Erio Tosatti, and Roberto Car. Theory of quantum annealing of an ising spin glass. *Science*, 295(5564):2427–2430, 2002.

[48] Hans-Paul Schwefel. *Numerical Optimization of Computer Models*. John Wiley & Sons, Inc., New York, NY, USA, 1981.

[49] Troels F. Rnnow Sergio Boixo, Sergei V. Isakov, Zhihui Wang, David Wecker, Daniel A. Lidar, John M. Martinis, and Matthias Troyer. Evidence for quantum annealing with more than one hundred qubits. *Nature Physics*, 10:218–224, 2014.

[50] Adrian FM Smith and Gareth O Roberts. Bayesian computation via the Gibbs sampler and related Markov chain Monte Carlo methods. *Journal of the Royal Statistical Society. Series B (Methodological)*, pages 3–23, 1993.

[51] Yun Ju Sung and Charles J. Geyer. Monte Carlo likelihood inference for missing data models. *The Annals of Statistics*, 35(3):990–1011, 2007.

[52] Harold Szu and Ralph Hartley. Fast simulated annealing. *Physics letters A*, 122(3):157–162, 1987.

[53] Darrell Whitley, Soraya Ranaa, John Dzubera, and Keith E Mathias. Artificial Intelligence Evaluating evolutionary algorithms. *Artificial Intelligence*, 85(1-2):245–276, 1996.

[54] Mark M Wilde and Bart Kosko. Quantum forbidden-interval theorems for stochastic resonance. *Journal of Physics A: Mathematical and Theoretical*, 42(46):465309, 2009.

[55] VM Zolotarev. *One-dimensional stable distributions*, volume 65. American Mathematical Soc., 1986.

PROOFS OF N-MCMC THEOREMS

Theorem 1 (Noisy Markov Chain Monte Carlo Theorem (N-MCMC))**.** *Suppose that $Q(x|x_t)$ is a Metropolis-Hastings jump pdf for time t and that it satisfies the detailed balance condition $\pi(x_t)\,Q(x|x_t) = \pi(x)\,Q(x_t|x)$ for the target equilibrium pdf $\pi(x)$. Then the MCMC noise benefit $d_t(N) \le d_t$ holds on average at time t if*

$$E_{N,X}\left[\ln \frac{Q(x_t+N \mid x)}{Q(x_t \mid x)}\right] \ge E_N\left[\ln \frac{\pi(x_t+N)}{\pi(x_t)}\right] \tag{8}$$

where $d_t = D\Big(\pi(x) \,\Big\|\, Q(x \mid x_t)\Big)$, $d_t(N) = D\Big(\pi(x) \,\Big\|\, Q(x \mid x_t+N)\Big)$, $N \sim f_{N|x_t}(n|x_t)$ is noise that may depend on x_t, and $D\Big(\cdot \,\Big\|\, \cdot\Big)$ is the relative-entropy pseudo-distance: $D\Big(P \,\Big\|\, Q\Big) = \int_X p(x) \ln\left(\frac{p(x)}{q(x)}\right)\, dx$.

Proof. Define the metrical averages (Kullback-Leibler divergences) d_t and $d_t(N)$ as

$$d_t = \int_X \pi(x) \ln \frac{\pi(x)}{Q(x|x_t)}\, dx = E_X\left[\ln \frac{\pi(x)}{Q(x|x_t)}\right] \tag{44}$$

$$d_t(N) = \int_X \pi(x) \ln \frac{\pi(x)}{Q(x_t+N|x)}\, dx = E_X\left[\ln \frac{\pi(x)}{Q(x|x_t+N)}\right]. \tag{45}$$

Take expectations over N: $E_N[d_t] = d_t$ and $E_N[d_t(N)] = E_N[d_t(N)]$. Then $d_t(N) \le d_t$ guarantees that a noise benefit occurs on average: $E_N[d_t(N)] \le d_t$.

Suppose that the N-MCMC condition holds:

$$E_N\left[\ln \frac{\pi(x_t+N)}{\pi(x_t)}\right] \le E_{N,X}\left[\ln \frac{Q(x_t+N \mid x)}{Q(x_t \mid x)}\right]. \tag{46}$$

Expand the expectations to give

$$\begin{aligned}&\int_N \ln \frac{\pi(x_t+n)}{\pi(x_t)} f_{N|x_t}(n|x_t)\, dn \\ &\qquad \le \iint_{N,X} \ln \frac{Q(x_t+n \mid x)}{Q(x_t \mid x)} \pi(x)\, f_{N|x_t}(n|x_t)\, dx\, dn.\end{aligned} \tag{47}$$

Then split the log ratios:

$$\int_N \ln \pi(x_t+n)\, f_{N|x_t}(n|x_t)\ dn - \int_N \ln \pi(x_t)\, f_{N|x_t}(n|x_t)\ dn$$
$$\leq \iint_{N,X} \ln Q(x_t+n \mid x)\pi(x)\, f_{N|x_t}(n|x_t)\ dx\ dn$$
$$- \iint_{N,X} \ln Q(x_t \mid x)\pi(x)\, f_{N|x_t}(n|x_t)\ dx\ dn. \qquad (48)$$

Rearrange the inequality as follows:

$$\int_N \ln \pi(x_t+n)\, f_{N|x_t}(n|x_t)\ dn$$
$$- \iint_{N,X} \ln Q(x_t+n \mid x)\pi(x)\, f_{N|x_t}(n|x_t)\ dx\ dn$$
$$\leq \int_N \ln \pi(x_t)\, f_{N|x_t}(n|x_t)\ dn$$
$$- \iint_{N,X} \ln Q(x_t \mid x)\pi(x)\, f_{N|x_t}(n|x_t)\ dx\ dn. \qquad (49)$$

Take expectations with respect to $\pi(x)$ in the single integrals:

$$\iint_{N,X} \ln \pi(x_t+n)\pi(x)\, f_{N|x_t}(n|x_t)\ dx\ dn$$
$$- \iint_{N,X} \ln Q(x_t+n \mid x)\pi(x)\, f_{N|x_t}(n|x_t)\ dx\ dn$$
$$\leq \iint_{N,X} \ln \pi(x_t)\pi(x)\, f_{N|x_t}(n|x_t)\ dx\ dn$$
$$- \iint_{N,X} \ln Q(x_t \mid x)\pi(x)\, f_{N|x_t}(n|x_t)\ dx\ dn. \qquad (50)$$

Then the joint integral factors into a product of single integrals from Fubini's Theorem[46] because we assume that all functions are integrable:

$$\iint_{N,X} \pi(x) \ln \frac{\pi(x_t+n)}{Q(x_t+n \mid x)} f_{N|x_t}(n|x_t)\ dx\ dn$$
$$\leq \underbrace{\int_N f_{N|x_t}(n|x_t)\ dn}_{=1} \int_X \pi(x) \ln \frac{\pi(x_t)}{Q(x_t \mid x)}\ dx \qquad (51)$$

since $f_{N|x_t}$ is a pdf.

The next step is the heart of the proof: Apply the MCMC detailed balance condition $\pi(x)\, Q(y|x) = \pi(y)\, Q(x|y)$ to the denominator Q terms in the previous inequality. This gives

$$Q(x_t|x) = \frac{\pi(x_t)\ Q(x|x_t)}{\pi(x)} \qquad (52)$$

and

$$Q(x_t+n|x) = \frac{\pi(x_t+n)\ Q(x|x_t+n)}{\pi(x)}. \qquad (53)$$

Insert these two Q equalities into (51) and then cancel like π terms to give

$$\iint_{N,X} \pi(x) \ln \frac{\cancel{\pi(x_t+n)}}{\frac{\cancel{\pi(x_t+n)}Q(x \mid x_t+n)}{\pi(x)}} f_{N|x_t}(n|x_t)\ dx\ dn$$
$$\leq \int_X \pi(x) \ln \frac{\cancel{\pi(x_t)}}{\frac{\cancel{\pi(x_t)}Q(x \mid x_t)}{\pi(x)}}\ dx. \qquad (54)$$

Rewrite the inequality as

$$\iint_{N,X} \pi(x) \ln \frac{\pi(x)}{Q(x \mid x_t+n)} f_{N|x_t}(n|x_t)\ dx\ dn$$
$$\leq \int_X \pi(x) \ln \frac{\pi(x)}{Q(x \mid x_t)}\ dx. \qquad (55)$$

Then Fubini's Theorem again gives

$$\int_N \left[\int_X \pi(x) \ln \frac{\pi(x)}{Q(x \mid x_t+n)}\ dx \right] f_{N|x_t}(n|x_t)\ dn$$
$$\leq \int_X \pi(x) \ln \frac{\pi(x)}{Q(x \mid x_t)}\ dx. \qquad (56)$$

This inequality holds if and only if (iff)

$$\int_N D\Big(\pi(x)\ \Big\|\ Q(x \mid x_t+n)\Big) f_{N|x_t}(n|x_t)\ dn$$
$$\leq D\Big(\pi(x)\ \Big\|\ Q(x \mid x_t)\Big). \qquad (57)$$

Then the metrical averages in (44) - (45) give

$$\int_N d_t(N)\, f_{N|x_t}(n|x_t)\ dn \leq d_t. \qquad (58)$$

This noise inequality is just the desired average result:

$$E_N\left[d_t(N)\right] \leq d_t. \qquad (59)$$

□

Corollary 1. *The N-MCMC noise benefit condition holds if*

$$Q(x_t+n|x) \geq e^A\ Q(x_t|x) \qquad (9)$$

for almost all x and n where

$$A = E_N\left[\ln \frac{\pi(x_t+N)}{\pi(x_t)}\right]. \qquad (10)$$

Proof. The following inequalities need hold only for almost all x and n. The first inequality is just the N-MCMC condition:

$$Q(x_t+n|x) \geq e^A\ Q(x_t|x) \qquad (60)$$

iff

$$\ln\left[Q(x_t+n|x)\right] \geq A + \ln\left[Q(x_t|x)\right] \qquad (61)$$

iff

$$\ln\left[Q(x_t+n|x)\right] - \ln\left[Q(x_t|x)\right] \geq A \qquad (62)$$

iff

$$\ln \frac{Q(x_t+n|x)}{Q(x_t|x)} \geq A. \qquad (63)$$

Then taking expectations gives the desired noise-benefit inequality:

$$E_{N,X}\left[\ln \frac{Q(x_t+N \mid x)}{Q(x_t \mid x)}\right] \geq E_N\left[\ln \frac{\pi(x_t+N)}{\pi(x_t)}\right]. \qquad (64)$$

□

Corollary 3. *Suppose $Q(x_t|x) \sim \mathcal{N}\left(x,\sigma^2\right)$. Then the sufficient noise benefit condition (9) holds if*

$$n\left(n-2\left(x_t-x\right)\right) \le -2\sigma^2 A \tag{13}$$

for A in (12).

Proof. Assume that the normal hypothesis holds: $Q(x_t|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t-x)^2}{2\sigma^2}}$. Then $Q(x_t+n|x) \ge e^A \; Q(x_t|x)$ holds iff

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t+n-x)^2}{2\sigma^2}} \ge e^A \; \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t-x)^2}{2\sigma^2}} \tag{65}$$

iff

$$e^{-\frac{(x_t+n-x)^2}{2\sigma^2}} \ge e^{A-\frac{(x_t-x)^2}{2\sigma^2}} \tag{66}$$

iff

$$-\frac{(x_t+n-x)^2}{2\sigma^2} \ge A - \frac{(x_t-x)^2}{2\sigma^2} \tag{67}$$

iff

$$-\left(x_t+n-x\right)^2 \ge 2\sigma^2 A - \left(x_t-x\right)^2 \tag{68}$$

iff

$$-x_t^2+2x_tx-2x_tn-x^2+2xn-n^2 \ge 2\sigma^2 A - x_t^2+2x_tx-x^2 \tag{69}$$

iff

$$2xn-2x_tn-n^2 \ge 2\sigma^2 A \tag{70}$$

$$\tag{71}$$

iff

$$n\left(n-2\left(x_t-x\right)\right) \le -2\sigma^2 A. \tag{72}$$

□

Corollary 4. *Suppose $Q(x_t|x) \sim \mathcal{N}\left(x,\sigma^2\right)$ and $g(x_t,n) = nx_t$. Then the sufficient noise benefit condition (11) holds if*

$$nx_t\left(nx_t-2x\right) - x_t\left(x_t-2x\right) \le -2\sigma^2 A \tag{15}$$

for A in (12).

Proof. Assume the normality condition that $Q(x_t|x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t-x)^2}{2\sigma^2}}$. Then $Q(nx_t|x) \ge e^A \; Q(y|x_t)$ holds iff

$$\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(nx_t-x)^2}{2\sigma^2}} \ge e^A \; \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x_t-x)^2}{2\sigma^2}} \tag{73}$$

iff

$$e^{-\frac{(nx_t-x)^2}{2\sigma^2}} \ge e^{A-\frac{(x_t-x)^2}{2\sigma^2}} \tag{74}$$

iff

$$-\frac{(nx_t-x)^2}{2\sigma^2} \ge A - \frac{(x_t-x)^2}{2\sigma^2} \tag{75}$$

iff

$$-\left(nx_t-x\right)^2 \ge 2\sigma^2 A - \left(x_t-x\right)^2 \tag{76}$$

iff

$$-x^2+2xnx_t-n^2x_t^2 \ge 2\sigma^2 A - x^2+2xx_t-x_t^2 \tag{77}$$

iff

$$2xnx_t-n^2x_t^2-2xx_t+x_t^2 \ge 2\sigma^2 A \tag{78}$$

iff

$$nx_t\left(nx_t-2x\right) - x_t\left(x_t-2x\right) \le -2\sigma^2 A. \tag{79}$$

□

Corollary 5. *Suppose $Q(x_t|x) \sim Cauchy(x,d)$. Then the sufficient condition (9) holds if*

$$n^2+2n\left(x_t-x\right) \le \left(e^{-A}-1\right)\left(d^2+\left(x_t-x\right)^2\right) \tag{16}$$

for A in (12).

Proof. Assume the Cauchy-pdf condition that

$$Q(x_t|x) = \frac{1}{\pi d\left[1+\left(\frac{x_t-x}{d}\right)^2\right]}. \tag{80}$$

Then

$$Q(x_t+n|x) \ge e^A \; Q(x_t|x) \tag{81}$$

iff

$$\frac{1}{\pi d\left[1+\left(\frac{x_t+n-x}{d}\right)^2\right]} \ge e^A \; \frac{1}{\pi d\left[1+\left(\frac{x_t-x}{d}\right)^2\right]} \tag{82}$$

iff

$$1+\left(\frac{x_t+n-x}{d}\right)^2 \le e^{-A} \; \left[1+\left(\frac{x_t-x}{d}\right)^2\right] \tag{83}$$

$$\le e^{-A}+e^{-A} \; \left(\frac{x_t-x}{d}\right)^2 \tag{84}$$

iff

$$\left(\frac{x_t+n-x}{d}\right)^2 - e^{-A} \; \left(\frac{x_t-x}{d}\right)^2 \le e^{-A}-1 \tag{85}$$

iff

$$\left(x_t+n-x\right)^2 - e^{-A} \; \left(x_t-x\right)^2 \le d^2\left(e^{-A}-1\right) \tag{86}$$

$$\left(x_t-x\right)^2+n^2+2n\left(x_t-x\right) - e^{-A} \; \left(x_t-x\right)^2 \le d^2\left(e^{-A}-1\right) \tag{87}$$

$$\left(1-e^{-A}\right)\left(x_t-x\right)^2+n^2+2n\left(x_t-x\right) \le d^2\left(e^{-A}-1\right) \tag{88}$$

iff

$$n^2 \le d^2\left(e^{-A}-1\right)+\left(e^{-A}-1\right)\left(x_t-x\right)^2-2n\left(x_t-x\right) \tag{89}$$

$$\le \left(e^{-A}-1\right)\left(d^2+\left(x_t-x\right)^2\right)-2n\left(x_t-x\right). \tag{90}$$

□

PROOF OF N-SA THEOREM

Theorem 2 (Noisy Simulated Annealing Theorem (N-SA))**.** *Suppose $C(x)$ is an energy surface with occupancy probabilities $\pi(x;T) \propto \exp\left(-\frac{C(x)}{T}\right)$. Then the simulated-annealing noise benefit*

$$E_N\left[\alpha_N(T)\right] \geq \alpha(T) \tag{17}$$

holds on average if

$$E_N\left[\ln \frac{\pi(x_t+N;T)}{\pi(x_t;T)}\right] \geq 0 \tag{18}$$

where $\alpha(T)$ is the simulated annealing acceptance probability from state x_t to the candidate x^*_{t+1} that depends on a temperature T (with cooling schedule $T(t)$):

$$\alpha(T) = \min\left\{1, \exp\left(-\frac{\Delta E}{T}\right)\right\} \tag{19}$$

and $\alpha_N(T)$ is the noisy simulated annealing acceptance probability from state x_t to the candidate $x^*_{t+1}+N$:

$$\alpha_N(T) = \min\left\{1, \exp\left(-\frac{\Delta E_N}{T}\right)\right\} \tag{20}$$

where $\Delta E = E^*_{t+1} - E_t = C\left(x^*_{t+1}\right) - C(x_t)$ is the energy difference of states x^*_{t+1} and x_t and $\Delta E_N = E^*_{N,t+1} - E_t = C\left(x^*_{t+1}+N\right) - C(x_t)$ is the energy difference of states $x^*_{t+1}+N$ and x_t .

Proof. The proof uses Jensen's inequality for a concave function g[13,46]:

$$g(E[X]) \geq E[g(x)] \tag{91}$$

where X is a real integrable random variable. Then Jensen's inequality gives

$$\ln E[X] \geq E[\ln X] \tag{92}$$

because the natural logarithm is a concave function.

We first expand $\alpha(T)$ in terms of the π densities:

$$\alpha(T) = \min\left\{1, \exp\left(-\frac{\Delta E}{T}\right)\right\} \tag{93}$$

$$= \min\left\{1, \exp\left(-\frac{E^*_{t+1} - E_t}{T}\right)\right\} \tag{94}$$

$$= \min\left\{1, \frac{\exp\left(-\frac{E^*_{t+1}}{T}\right)}{\exp\left(-\frac{E_t}{T}\right)}\right\} \tag{95}$$

$$= \min\left\{1, \frac{\frac{\cancel{1}}{\cancel{Z}} \cdot \pi\left(x^*_{t+1};T\right)}{\frac{\cancel{1}}{\cancel{Z}} \cdot \pi(x_t;T)}\right\} \tag{96}$$

$$= \min\left\{1, \frac{\pi\left(x^*_{t+1};T\right)}{\pi(x_t;T)}\right\} \tag{97}$$

for the normalizing constant

$$Z = \int_X \exp\left(-\frac{C(x)}{T}\right)\, dx. \tag{98}$$

We next let N be an integrable noise random variable that perturbs the candidate state x^*_{t+1}. We want to show the inequality

$$E_N\left[\alpha_N(T)\right] = E_N\left[\min\left\{1, \frac{\pi\left(x^*_{t+1}+N;T\right)}{\pi(x_t;T)}\right\}\right] \tag{99}$$

$$\geq \min\left\{1, \frac{\pi\left(x^*_{t+1};T\right)}{\pi(x_t;T)}\right\} \tag{100}$$

$$= \alpha(T). \tag{101}$$

So it suffices to show that

$$E_N\left[\frac{\pi\left(x^*_{t+1}+N;T\right)}{\pi(x_t;T)}\right] \geq \frac{\pi\left(x^*_{t+1};T\right)}{\pi(x_t;T)} \tag{102}$$

holds. This inequality holds iff

$$E_N\left[\pi\left(x^*_{t+1}+N;T\right)\right] \geq \pi\left(x^*_{t+1};T\right) \tag{103}$$

because $\pi(x_t) \geq 0$ since π is a pdf.

Suppose now that the N-SA condition holds:

$$E_N\left[\ln \frac{\pi(x_t+N)}{\pi(x_t)}\right] \geq 0. \tag{104}$$

Then

$$E_N\left[\ln \pi(x_t+N) - \ln \pi(x_t)\right] \geq 0 \tag{105}$$

iff

$$E_N\left[\ln \pi(x_t+N)\right] \geq E_N\left[\ln \pi(x_t)\right]. \tag{106}$$

Then Jensen's inequality gives the inequality

$$\ln E_N\left[\pi(x_t+N)\right] \geq E_N\left[\ln \pi(x_t)\right]. \tag{107}$$

This inequality holds iff

$$\ln E_N\left[\pi(x_t+N)\right] \geq \int_N \ln \pi(x_t) f_N(n|x_t)\; dn \tag{108}$$

iff

$$\ln E_N\left[\pi(x_t+N)\right] \geq \ln \pi(x_t) \underbrace{\int_N f_N(n|x_t)\; dn}_{=1} \tag{109}$$

iff

$$\ln E_N\left[\pi(x_t+N)\right] \geq \ln \pi(x_t). \tag{110}$$

Then taking exponentials gives the desired average noise benefit:

$$E_N\left[\pi(x_t+N)\right] \geq \pi(x_t). \tag{111}$$

□

Corollary 6. *Suppose m is a convex increasing function. Then the N-SA Theorem noise benefit*

$$E_N\left[\beta_N(T)\right] \geq \beta(T) \tag{21}$$

holds on average if

$$E_N\left[\ln \frac{\pi(x_t+N;T)}{\pi(x_t;T)}\right] \geq 0 \tag{22}$$

where β is the acceptance probability from state x_t to the candidate x^*_{t+1}:

$$\beta(T) = \min\left\{1, m\left(\frac{\pi\left(x^*_{t+1};T\right)}{\pi(x_t;T)}\right)\right\}. \tag{23}$$

and β_N is the noisy acceptance probability from state x_t to the candidate $x^*_{t+1}+N$:

$$\beta_N(T) = \min\left\{1, m\left(\frac{\pi\left(x^*_{t+1}+N;T\right)}{\pi(x_t;T)}\right)\right\}. \tag{24}$$

Proof. We want to show that

$$\begin{aligned} E_N\left[\beta_N(T)\right] &= E_N\left[\min\left\{1, m\left(\frac{\pi\left(x^*_{t+1}+N;T\right)}{\pi(x_t;T)}\right)\right\}\right] && (112)\\ &\geq \min\left\{1, m\left(\frac{\pi\left(x^*_{t+1};T\right)}{\pi(x_t;T)}\right)\right\} && (113)\\ &= \beta(T). && (114)\end{aligned}$$

So it suffices to show that

$$E_N\left[m\left(\frac{\pi\left(x^*_{t+1}+N;T\right)}{\pi(x_t;T)}\right)\right] \geq m\left(\frac{\pi\left(x^*_{t+1};T\right)}{\pi(x_t;T)}\right). \tag{115}$$

Suppose that the N-SA condition holds:

$$E_N\left[\ln \frac{\pi(x_t+N;T)}{\pi(x_t;T)}\right] \geq 0. \tag{116}$$

Then

$$E_N\left[\pi(x_t+N)\right] \geq \pi(x_t) \tag{117}$$

holds as in the proof of the N-SA Theorem. This inequality implies that

$$\frac{E_N\left[\pi(x_t+N)\right]}{\pi(x_t;T)} \geq \frac{\pi(x_t)}{\pi(x_t;T)} \tag{118}$$

because $\pi(x_t) \geq 0$ since π is a pdf and because

$$E_N\left[\frac{\pi(x_t+N)}{\pi(x_t;T)}\right] \geq \frac{\pi(x_t)}{\pi(x_t;T)}. \tag{119}$$

Then

$$m\left(E_N\left[\frac{\pi(x_t+N)}{\pi(x_t;T)}\right]\right) \geq m\left(\frac{\pi(x_t)}{\pi(x_t;T)}\right) \tag{120}$$

because m is an increasing function. Then Jensen's inequality for convex functions gives the desired average noise inequality:

$$E_N\left[m\left(\frac{\pi(x_t+N)}{\pi(x_t;T)}\right)\right] \geq m\left(\frac{\pi(x_t)}{\pi(x_t;T)}\right) \tag{121}$$

because m is convex. □

Corollary 7. *Suppose $\pi(x) = Ce^{g(x)}$ if C is the normalizing constant $C = \frac{1}{\int_X e^{g(x)}\,dx}$. Then there is an N-SA noise benefit if*

$$E_N\left[g(x_t+N)\right] \geq g(x_t). \tag{25}$$

Proof. Suppose that the N-SA condition holds:

$$E_N\left[g(x_t+N)\right] \geq g(x_t). \tag{122}$$

Then the equivalent inequality

$$E_N\left[\ln e^{g(x_t+N)}\right] \geq \ln e^{g(x_t)} \tag{123}$$

holds iff the following equivalent inequalities hold:

$$\begin{aligned} E_N\left[\ln\left(e^{g(x_t+N)}\right)\right] &\geq \ln\left(e^{g(x_t)}\right) && (124)\\ E_N\left[\ln \frac{\pi(x_t+N)}{C}\right] &\geq \ln \frac{\pi(x_t)}{C} && (125)\\ E_N\left[\ln \frac{\pi(x_t+N)}{C} - \ln \frac{\pi(x_t)}{C}\right] &\geq 0 && (126)\\ E_N\left[\ln \frac{\frac{\pi(x_t+N)}{\cancel{C}}}{\frac{\pi(x_t)}{\cancel{C}}}\right] &\geq 0 && (127)\\ E_N\left[\ln \frac{\pi(x_t+N)}{\pi(x_t)}\right] &\geq 0. && (128)\end{aligned}$$

□