Simulated Annealing Report

Harsh Maurya , Rahul Pandit , Kanishk Bhuradiya , Bhagavath Chukka

May 2026

1 Introduction

Optimization lies at the heart of countless problems in science, engineering, and machine learning — from routing delivery vehicles to tuning the parameters of a neural network. While many classical optimization methods, such as gradient descent, perform well on smooth and convex landscapes, they frequently falter when confronted with highly non-convex, multimodal, or combinatorially complex problems, becoming trapped in suboptimal local minima with no mechanism for escape.

Simulated Annealing (SA) is a probabilistic metaheuristic inspired by the physical process of annealing in metallurgy, wherein a material is heated to a high temperature and then slowly cooled to reduce defects and reach a low-energy crystalline state. By analogy, the algorithm begins its search with high "temperature," allowing it to freely accept worse solutions with some probability, thereby enabling broad exploration of the search space. As the temperature is gradually reduced according to a cooling schedule, the algorithm becomes increasingly selective, concentrating its search around promising low-cost configurations and ultimately converging toward the global minimum.

This report presents a rigorous mathematical and empirical study of Simulated Annealing. We begin by formalizing the model setup, including the state space, cost function, neighbourhood structure, and the time-inhomogeneous Markov chain that governs the algorithm's dynamics. We then analyze the stationary behavior of the chain at fixed temperature, establishing its

connection to the Gibbs (Boltzmann) distribution, and derive how this distribution concentrates on the set of global minima as temperature approaches zero.

A central theoretical result examined in this report is Hajek's Convergence Theorem, which provides precise conditions on the cooling schedule — specifically, a logarithmic decay with a constant exceeding the maximum depth of local energy barriers — under which simulated annealing is guaranteed to converge to the global optimum with probability one. We further investigate the spectral properties of the transition matrix, relating the rate of convergence to the spectral gap and the relaxation time of the Markov chain, and analyze how energy barriers exponentially slow mixing at low temperatures.

On the practical side, we conduct extensive empirical experiments across four distinct settings: smooth non-convex optimization via the Rosenbrock function, highly multimodal optimization via the Rastrigin function, combinatorial optimization via the Traveling Salesman Problem (TSP), and machine learning hyperparameter tuning on the Breast Cancer Wisconsin dataset. In each setting, SA is compared against alternative methods — including Gradient Descent, Hill Climbing, 2-opt local search, Genetic Algorithms, Ant Colony Optimization, Random Search, and Bayesian Optimization — to illuminate the strengths and limitations of stochastic annealing across diverse problem structures.

Through this combined theoretical and experimental analysis, the report aims to offer a comprehensive understanding of when and why Simulated Annealing succeeds, where it is outperformed by specialized alternatives, and what directions — such as adaptive cooling schedules, hybrid methods, and parallel implementations — hold promise for future development.

2 Model Setup

We consider a finite state space S and a real-valued cost function

$$J : S \to \mathbb{R}.$$

The objective is to minimize the function J over the state space S. Let the set of global minima be defined as

$$S^* = \left\{ i \in S : J(i) = \min_{x \in S} J(x) \right\}.$$

2.1 Neighbourhood Structure

For each state $i \in S$, let $S(i) \subseteq S \setminus \{i\}$ denote the set of neighbouring states of i. We assume that the neighbourhood structure is symmetric, i.e.,

$$j \in S(i) \quad \Longleftrightarrow \quad i \in S(j).$$

2.2 Proposal Probabilities

For each $i \in S$, let $\{q_{ij} : j \in S(i)\}$ be a collection of positive numbers such that

$$\sum_{j \in S(i)} q_{ij} = 1.$$

Here, q_{ij} represents the probability of proposing a move from state i to state j.

2.3 Cooling Schedule

Let

$$T : \mathbb{N} \to (0, \infty)$$

be a non-increasing function, called the **cooling schedule**, where $T(t)$ denotes the temperature at time t.

2.4 Initial State

The algorithm starts from an initial state

$$x(0) \in S.$$

2.5 Transition Probabilities

The simulated annealing algorithm defines a time-inhomogeneous Markov chain $\{x(t)\}_{t\geq 0}$ with transition probabilities given by

$$P\big(x(t{+}1) = j \mid x(t) = i\big) = \begin{cases} q_{ij} \exp\!\left(-\dfrac{\max\{0,\, J(j) - J(i)\}}{T(t)}\right), & j \in S(i), \\ 0, & j \notin S(i),\ j \neq i, \end{cases}$$

where the acceptance probability $A(i,j)$ is defined as

$$A(i,j) = \exp\left(-\frac{\max\{0,\, J(j) - J(i)\}}{T(t)}\right).$$

The probability of remaining in the same state i is defined as

$$P\big(x(t+1) = i \mid x(t) = i\big) = 1 - \sum_{j \in S(i)} P\big(x(t+1) = j \mid x(t) = i\big).$$

This construction ensures that the process $\{x(t)\}$ is a well-defined Markov chain whose transition probabilities depend on time through the temperature $T(t)$.

3 Reversibility and Gibbs Distribution

In this section, we analyze the behavior of the simulated annealing algorithm when the temperature is held fixed at a constant value $T > 0$. In this case, the transition probabilities no longer depend on time, and the process $\{x(t)\}_{t\geq 0}$ becomes a homogeneous Markov chain.

3.1 Homogeneous Markov Chain at Fixed Temperature

When the temperature is fixed at T, the transition probabilities become

$$P(i \to j) = q_{ij}\, A(i,j),$$

where

$$A(i,j) = \exp\left(-\frac{\max\{0,\ J(j) - J(i)\}}{T}\right).$$

We assume that the proposal probabilities are symmetric, i.e.,

$$q_{ij} = q_{ji}, \quad \forall\, i, j \in S.$$

3.2 Detailed Balance Equation

We seek a distribution π_T on S satisfying

$$\pi_T(i)\, P(i \to j) = \pi_T(j)\, P(j \to i), \quad \forall\, i, j \in S.$$

Substituting $P(i \to j) = q_{ij} A(i,j)$, we obtain

$$\pi_T(i)\, q_{ij} A(i,j) = \pi_T(j)\, q_{ji} A(j,i).$$

Using $q_{ij} = q_{ji}$, this simplifies to

$$\frac{\pi_T(j)}{\pi_T(i)} = \frac{A(i,j)}{A(j,i)}.$$

3.3 Evaluation of the Acceptance Ratio

Case 1: $J(j) \geq J(i)$

$$A(i,j) = \exp\left(-\frac{J(j) - J(i)}{T}\right), \quad A(j,i) = 1.$$

$$\frac{A(i,j)}{A(j,i)} = \exp\left(-\frac{J(j) - J(i)}{T}\right).$$

Case 2: $J(j) < J(i)$

$$A(i,j) = 1, \quad A(j,i) = \exp\left(-\frac{J(i) - J(j)}{T}\right).$$

$$\frac{A(i,j)}{A(j,i)} = \exp\left(-\frac{J(j) - J(i)}{T}\right).$$

Thus, in both cases,

$$\frac{A(i,j)}{A(j,i)} = \exp\left(-\frac{J(j) - J(i)}{T}\right).$$

3.4 Derivation of the Stationary Distribution

Substituting into the detailed balance equation, we obtain

$$\frac{\pi_T(j)}{\pi_T(i)} = \exp\left(-\frac{J(j) - J(i)}{T}\right).$$

This implies that

$$\pi_T(i) \propto \exp\left(-\frac{J(i)}{T}\right).$$

To normalize, define

$$Z_T = \sum_{k \in S} \exp\left(-\frac{J(k)}{T}\right),$$

and set

$$\pi_T(i) = \frac{1}{Z_T} \exp\left(-\frac{J(i)}{T}\right).$$

The constant Z_T is called the partition function.

The distribution π_T obtained above is known as the *Gibbs distribution* (or Boltzmann distribution) associated with the cost function J at temperature T.

3.5 Behavior as $T \to 0$

Recall that

$$\pi_T(i) = \frac{1}{Z_T} \exp\left(-\frac{J(i)}{T}\right).$$

Let $i^* \in S^*$ be a global minimum, i.e.,

$$J(i^*) = \min_{x \in S} J(x).$$

Now consider any state $j \notin S^*$, so that

$$J(j) > J(i^*).$$

Then,

$$\frac{\pi_T(j)}{\pi_T(i^*)} = \exp\left(-\frac{J(j) - J(i^*)}{T}\right).$$

Since $J(j) - J(i^*) > 0$, we have

$$\exp\left(-\frac{J(j) - J(i^*)}{T}\right) \to 0 \quad \text{as } T \to 0.$$

Thus,

$$\frac{\pi_T(j)}{\pi_T(i^*)} \to 0,$$

which implies that

$$\pi_T(j) \to 0 \quad \text{for all } j \notin S^*.$$

Hence, the distribution concentrates on the set of global minima.

3.6 Interpretation

The distribution π_T assigns higher probability to states with lower cost. As $T \to 0$, the distribution concentrates on the set of global minima S^*.

4 Hajek's Convergence Theorem

We now present a fundamental result that gives sufficient conditions for convergence of simulated annealing to the global minimum.

4.1 Depth of Local Minima

For a state $i \in S$, define its *depth* $d(i)$ as follows.

Consider all paths from i to the set of global minima S^*. For any such path $\gamma = (i = i_0, i_1, \ldots, i_k \in S^*)$, define its height as

$$H(\gamma) = \max_{0 \le r \le k} J(i_r).$$

Then the depth of state i is defined as

$$d(i) = \min_{\gamma} \big(H(\gamma) - J(i) \big),$$

where the minimum is taken over all paths γ from i to S^*.

Let

$$d^* = \max_{i \notin S^*} d(i)$$

be the maximum depth over all non-optimal states.

4.2 Statement of Hajek's Theorem

Theorem 1 (Hajek) *Suppose the cooling schedule satisfies*

$$T(t) = \frac{c}{\log(1+t)},$$

where $c > d^*$.

Then the simulated annealing process satisfies

$$\lim_{t \to \infty} \mathbb{P}\big(x(t) \in S^* \big) = 1.$$

4.3 Probabilistic Escape Argument

We now provide an intuitive justification for the convergence condition using escape probabilities from local minima.

Let d^* denote the maximum depth of local minima. From the Gibbs acceptance rule, the probability of escaping a local minimum of depth d^* at time t is approximately

$$P(\text{escape at time } t) \approx \exp\left(-\frac{d^*}{T(t)} \right).$$

4.3.1 Effect of Temperature

- When $T(t)$ is large:
$$\exp\left(-\frac{d^*}{T(t)}\right) \approx 1,$$
so escape is likely.
- When $T(t)$ is small:
$$\exp\left(-\frac{d^*}{T(t)}\right) \approx 0,$$
so escape becomes extremely unlikely.

4.3.2 Summability Condition

For convergence, it is necessary that escape attempts occur infinitely often with positive probability. This leads to the condition

$$\sum_{t=1}^{\infty} \exp\left(-\frac{d^*}{T(t)}\right) = \infty.$$

4.3.3 Connection to Borel–Cantelli Lemma

By the Borel–Cantelli lemma, if

$$\sum_{t=1}^{\infty} \mathbb{P}(\text{escape at time } t) = \infty,$$

then escape events occur infinitely often with probability 1.

Thus, the algorithm will eventually escape any local minimum and reach the set of global minima.

4.4 Logarithmic Cooling Schedule

Consider the cooling schedule

$$T(t) = \frac{d}{\log(t)}, \quad t \geq 2,$$

where $d > 0$ is a constant.

We analyze the summability condition

$$\sum_{t=2}^{\infty} \exp\left(-\frac{d^*}{T(t)}\right).$$

Substituting $T(t)$, we obtain

$$\exp\left(-\frac{d^*}{T(t)}\right) = \exp\left(-\frac{d^*}{d}\log(t)\right) = (t)^{-d^*/d}.$$

Thus,

$$\sum_{t=2}^{\infty} \exp\left(-\frac{d^*}{T(t)}\right) = \sum_{t=2}^{\infty} (t)^{-d^*/d}.$$

4.5 Application of the p-Series Test

Recall that the series

$$\sum_{t=2}^{\infty} t^{-p}$$

diverges if and only if $p \leq 1$.

Thus, the above series diverges if and only if

$$\frac{d^*}{d} \leq 1,$$

which is equivalent to

$$d \geq d^*.$$

4.6 Conclusion

Hence, simulated annealing converges if the cooling schedule satisfies

$$T(t) = \frac{d}{\log(t)} \quad \text{with } d \geq d^*.$$

In particular, the constant d must be at least as large as the maximum depth d^* of the energy barriers.

5 Convergence Behaviour of Simulated Annealing

5.1 Fixed Temperature Case

If the temperature is held constant at T, then simulated annealing reduces to a homogeneous Markov chain with transition matrix P_T.

The distribution at time t is given by

$$\pi(t) = \pi(0)\, P_T^t.$$

5.2 Stationary Distribution

At temperature T, the chain admits a stationary distribution π given by the Gibbs distribution

$$\pi(i) \propto \exp\left(-\frac{J(i)}{T}\right).$$

5.3 Detailed Balance Condition

The simulated annealing transition probabilities satisfy the detailed balance condition:

$$\pi(i)\, P(i \to j) = \pi(j)\, P(j \to i).$$

Thus, the Markov chain is reversible with respect to π.

5.4 Matrix Formulation

Let D be the diagonal matrix defined by

$$D_{ii} = \pi(i).$$

Then the detailed balance condition can be written as

$$DP = P^T D.$$

5.5 Verification

From the detailed balance condition,

$$\pi(i)\,P(i,j) = \pi(j)\,P(j,i).$$

Thus,

$$(DP)_{ij} = D_{ii}P(i,j) = \pi(i)P(i,j),$$

and

$$(P^T D)_{ij} = P(j,i)D_{jj} = P(j,i)\pi(j).$$

Hence,

$$(DP)_{ij} = (P^T D)_{ij},$$

which implies

$$DP = P^T D.$$

5.6 Remark

This matrix identity shows that the transition matrix P is reversible with respect to π, which plays a crucial role in understanding the convergence behaviour of the chain.

5.7 Spectral Properties of the Transition Matrix

Let $D = \mathrm{diag}(\pi)$, and define

$$S = D^{1/2} P D^{-1/2}.$$

Using the detailed balance condition $DP = P^T D$, we obtain

$$S^T = S,$$

so S is symmetric.

5.7.1 Similarity of P and S

Observe that

$$S = D^{1/2} P D^{-1/2} = M P M^{-1}, \quad \text{where } M = D^{1/2}.$$

Thus, P and S are similar matrices, and hence they have the same eigenvalues.

5.7.2 Consequences

Since S is symmetric, all its eigenvalues are real. Therefore, all eigenvalues of P are real.

Let $Pv = \lambda v$. Then

$$\|Pv\|_\infty = |\lambda| \|v\|_\infty.$$

Since P is a stochastic matrix, we have

$$\|Pv\|_\infty \leq \|v\|_\infty,$$

which implies

$$|\lambda| \leq 1.$$

Thus, all eigenvalues of P lie in the interval $[-1, 1]$.

5.7.3 Largest Eigenvalue

Since $P\mathbf{1} = \mathbf{1}$, we have

$$\lambda_1 = 1.$$

5.8 Spectral Decomposition and Convergence

Let P be the transition matrix at fixed temperature. Since P is reversible, it has real eigenvalues satisfying

$$1 = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n \geq -1.$$

Let $\{\phi_k\}$ denote the left eigenvectors of P, i.e.,

$$\phi_k P = \lambda_k \phi_k.$$

Since probability distributions are represented as row vectors, we expand the initial distribution $\pi(0)$ in terms of the left eigenvectors:

$$\pi(0) = \sum_{k=1}^{n} c_k \phi_k.$$

After t steps,

$$\pi(t) = \pi(0)P^t = \sum_{k=1}^{n} c_k \lambda_k^t \phi_k.$$

5.8.1 Stationary Distribution

We know that

$$\pi P = \pi,$$

so the stationary distribution π is the left eigenvector corresponding to the eigenvalue $\lambda_1 = 1$. Thus,

$$\phi_1 = \pi.$$

5.8.2 Determination of the Coefficient c_1

Recall the spectral expansion

$$\pi(0) = \sum_{k=1}^{n} c_k \phi_k.$$

Since $\pi(0)$ is a probability distribution,

$$\sum_i \pi(0)(i) = 1.$$

Summing the spectral expansion over all states i, we obtain

$$\sum_i \pi(0)(i) = \sum_i \sum_{k=1}^{n} c_k \phi_k(i) = \sum_{k=1}^{n} c_k \sum_i \phi_k(i).$$

5.8.3 Biorthogonality Argument

Let $\mathbf{1}$ denote the vector of all ones. Since $P\mathbf{1} = \mathbf{1}$, it follows that $\mathbf{1}$ is a right eigenvector corresponding to the eigenvalue $\lambda_1 = 1$.

Now consider any left eigenvector ϕ_k:

$$\phi_k P = \lambda_k \phi_k.$$

Multiplying on the right by $\mathbf{1}$, we obtain

$$\phi_k P\mathbf{1} = \lambda_k \phi_k \mathbf{1}.$$

Using $P\mathbf{1} = \mathbf{1}$, this gives

$$\phi_k \mathbf{1} = \lambda_k \phi_k \mathbf{1}.$$

Hence,

$$(1 - \lambda_k)\phi_k \mathbf{1} = 0.$$

Therefore, if $\lambda_k \neq 1$ (that is, for all $k \geq 2$), we must have

$$\phi_k \mathbf{1} = \sum_i \phi_k(i) = 0.$$

We normalize the stationary eigenvector so that

$$\sum_i \phi_1(i) = \sum_i \pi(i) = 1.$$

Thus,

$$\sum_i \phi_k(i) = \begin{cases} 1, & k = 1, \\ 0, & k \geq 2. \end{cases}$$

Substituting back into the normalization condition,

$$\sum_i \pi(0)(i) = c_1 \cdot 1 + \sum_{k=2}^{n} c_k \cdot 0 = c_1.$$

Since the left-hand side equals 1, we conclude that

$$c_1 = 1.$$

Hence, the spectral decomposition becomes

$$\pi(t) = \pi + \sum_{k=2}^{n} c_k \lambda_k^t \phi_k.$$

5.9 Asymptotic Behaviour

For all $k \geq 2$, we have

$$|\lambda_k| < 1.$$

Therefore,

$$\lambda_k^t \to 0 \quad \text{as } t \to \infty.$$

Hence,

$$\pi(t) \to \pi,$$

which establishes convergence to the stationary distribution.

5.10 Rate of Convergence

Among all eigenvalues λ_k with $k \geq 2$, the dominant term is λ_2, since

$$|\lambda_2| \geq |\lambda_k| \quad \text{for all } k \geq 2.$$

Thus, for large t, the leading term in the error is

$$\pi(t) - \pi \approx c_2 \lambda_2^t \phi_2.$$

Taking norms, we obtain

$$\|\pi(t) - \pi\| \leq \sum_{k=2}^{n} |c_k| |\lambda_k|^t \|\phi_k\|.$$

Using $|\lambda_k| \leq |\lambda_2|$, we get

$$\|\pi(t) - \pi\| \leq |\lambda_2|^t \sum_{k=2}^{n} |c_k| \|\phi_k\|.$$

Define the constant

$$C = \sum_{k=2}^{n} |c_k| \|\phi_k\|.$$

Then,

$$\|\pi(t) - \pi\| \leq C|\lambda_2|^t.$$

5.11 Conclusion

The convergence of the Markov chain is *geometric* (exponential decay), and the rate of convergence is governed by the second-largest eigenvalue λ_2.

A smaller value of $|\lambda_2|$ leads to faster convergence, while values close to 1 indicate slow mixing.

5.12 Spectral Gap and Mixing Time

Let λ_2 denote the second-largest eigenvalue (in absolute value) of the transition matrix P. Define the *spectral gap* as

$$\gamma = 1 - |\lambda_2|.$$

5.12.1 Interpretation

- A large spectral gap (γ close to 1) implies rapid mixing.
- A small spectral gap (γ close to 0) implies slow convergence.

The Markov chain is said to *mix* when the distribution $\pi(t)$ becomes close to the stationary distribution π, i.e.,

$$\pi(t) \approx \pi.$$

Intuitively, this corresponds to the chain "forgetting" its initial state.

5.13 Relaxation Time

The convergence behaviour can be expressed as

$$\|\pi(t) - \pi\| \leq C|\lambda_2|^t.$$

This suggests an exponential decay of the form

$$\|\pi(t) - \pi\| \approx e^{-t/\tau},$$

where τ is called the *relaxation time*.

Equating the two expressions, we set

$$|\lambda_2|^t = e^{-t/\tau}.$$

Taking logarithms, we obtain

$$\log|\lambda_2| = -\frac{1}{\tau},$$

which gives

$$\tau = -\frac{1}{\log|\lambda_2|}.$$

Thus, the relaxation time quantifies the number of steps required for the chain to approach equilibrium.

5.14 Approximation for Small Spectral Gap

Using the definition $\gamma = 1 - |\lambda_2|$, we write

$$|\lambda_2| = 1 - \gamma.$$

For small γ, we use the approximation

$$\log(1 - \gamma) \approx -\gamma.$$

Substituting into the expression for τ, we obtain

$$\tau \approx \frac{1}{\gamma}.$$

5.15 Conclusion

The relaxation time is inversely proportional to the spectral gap. Hence:

- Larger spectral gap $\Rightarrow$ faster convergence
- Smaller spectral gap $\Rightarrow$ slower convergence

6 Energy Barriers and Convergence Speed

We now relate the convergence rate of simulated annealing to the presence of energy barriers in the state space.

6.1 Barrier Crossing Probability

Suppose the state space is divided into two regions:

- Region A: containing a local minimum
- Region B: containing the global minimum

To move from region A to region B, the Markov chain must cross an energy barrier of height d^*.

From the acceptance probability of simulated annealing, the probability of crossing this barrier at temperature T is approximately

$$p(T) \approx \exp\left(-\frac{d^*}{T}\right).$$

6.2 Two-State Approximation

We approximate the dynamics by a two-state Markov chain between regions A and B:

$$P = \begin{pmatrix} 1-p & p \\ p & 1-p \end{pmatrix}.$$

The eigenvalues of this transition matrix are

$$\lambda_1 = 1, \quad \lambda_2 = 1 - 2p.$$

6.3 Spectral Gap

The spectral gap is

$$\gamma = 1 - \lambda_2 = 2p.$$

Substituting $p(T)$, we obtain

$$\gamma(T) \approx 2 \exp\left(-\frac{d^*}{T}\right).$$

Thus, the spectral gap decreases exponentially as temperature decreases.

6.4 Relaxation Time

From the earlier relation

$$\tau \approx \frac{1}{\gamma},$$

we obtain

$$\tau(T) \approx \exp\left(\frac{d^*}{T}\right).$$

6.5 Conclusion

The relaxation time grows exponentially as the temperature decreases. Therefore, the convergence speed of simulated annealing is controlled by the probability of crossing the deepest energy barrier.

In particular, low temperatures lead to extremely slow mixing due to the exponential suppression of barrier-crossing events.

7 Time-Inhomogeneous Dynamics and Cooling Schedule

Simulated annealing differs from standard Markov chains in that the transition matrix changes over time through the temperature schedule $T(t)$. Thus, the process is a time-inhomogeneous Markov chain.

7.1 Local Equilibrium Approximation

At a fixed temperature T, the Markov chain mixes toward the Gibbs distribution π_T over a time scale given by the relaxation time

$$\tau(T) \approx \exp\left(\frac{d^*}{T}\right),$$

where d^* is the height of the deepest energy barrier.

If the temperature changes slowly, the chain behaves approximately like a homogeneous Markov chain at temperature $T(t)$ over short time intervals.

7.2 Cooling Schedule

Consider a cooling schedule of the form

$$T(t) \approx \frac{1}{k},$$

over time intervals $[t_k, t_{k+1}]$, where

$$t_{k+1} = t_k + \exp(kd).$$

Thus, the duration of the k-th stage is

$$\Delta t_k = \exp(kd).$$

During this interval, the chain behaves approximately like a homogeneous Markov chain with temperature $T = 1/k$.

7.3 Time Scale Comparison

At temperature $T = 1/k$, the relaxation time is

$$\tau(T) \approx \exp\left(\frac{d^*}{T}\right) = \exp(kd^*).$$

The time available for mixing at this temperature is

$$\Delta t_k = \exp(kd).$$

Thus, the ratio of available time to required mixing time is

$$\frac{\text{time available}}{\text{time required}} = \frac{\exp(kd)}{\exp(kd^*)} = \exp\big(k(d - d^*)\big).$$

7.4 Condition for Convergence

- If $d > d^*$, then
$$\exp\big(k(d - d^*)\big) \to \infty,$$
and the chain has sufficient time to equilibrate at each temperature level.

- If $d < d^*$, then
$$\exp\big(k(d - d^*)\big) \to 0,$$
and the chain does not have enough time to mix before the temperature decreases further.

- If $d = d^*$, then
$$\exp\big(k(d - d^*)\big) = 1,$$
so the time available is of the same order as the mixing time.

  In this critical case, the system is only marginally able to equilibrate, and convergence may still occur but at a very slow rate.

7.5 Conclusion

Thus, for simulated annealing to converge to the global minimum, the cooling schedule must be sufficiently slow so that the system equilibrates at each temperature level.

This leads to the condition

$$d \geq d^*,$$

which matches the threshold obtained from Hajek's theorem.

Hence, convergence is governed by a balance between:

- the rate at which temperature decreases, and
- the time required to escape energy barriers.

8 Practical Relevance of Convergence Results

In the preceding sections, we established that simulated annealing converges to the set of global minima under suitable conditions on the cooling schedule. However, this result is far from sufficient for simulated annealing to be a useful algorithm in practice.

In practical applications, one is interested not only in eventual convergence, but in the quality of the best solution obtained within a finite amount of time.

To address this issue, it is natural to study the time required for the algorithm to reach the set of optimal configurations.

8.1 First Hitting Time

Let S^* denote the set of global minima of the cost function J. We define the *first hitting time* of the set S^* as

$$\tau = \inf\{t \geq 0 : x(t) \in S^*\}.$$

Then, for any $t \geq 0$, we have

$$\mathbb{P}\big(x(t) \notin S^*\big) = \mathbb{P}(\tau > t).$$

Thus, the problem of evaluating the performance of simulated annealing reduces to obtaining bounds on the tail probability of the hitting time τ.

8.2 Polynomial Decay of Failure Probability

We now outline how one can obtain a bound of the form

$$\mathbb{P}(\tau > t) \leq \frac{A}{t^a},$$

for some positive constants A and a, under a cooling schedule of the form

$$T(t) = \frac{d}{\log t}, \quad d > d^*.$$

The key idea is to estimate the probability that the chain enters the set S^* at a given time step.

From the analysis of simulated annealing, the probability of escaping from a local minimum of depth d^* at time t is of the order

$$\exp\left(-\frac{d^*}{T(t)}\right).$$

Substituting $T(t) = d/\log t$, we obtain

$$\exp\left(-\frac{d^*}{T(t)}\right) = \exp\left(-\frac{d^* \log t}{d}\right) = t^{-d^*/d}.$$

Thus, there exist constants $c > 0$ and $a > 0$ such that the probability of entering S^* at time t, conditional on the past, satisfies

$$\mathbb{P}\big(x(t) \in S^* \mid \mathcal{F}_{t-1}\big) \geq \frac{c}{t^a}.$$

8.3 Bounding the Tail Probability

We now bound the probability that the chain has not entered S^* up to time t:

$$\mathbb{P}(\tau > t) = \mathbb{P}\big(x(k) \notin S^*,\ \forall k \leq t\big).$$

Using the Markov property and the above lower bound, we obtain

$$\mathbb{P}(\tau > t) \leq \prod_{k=1}^{t} \left(1 - \frac{c}{k^a}\right).$$

Taking logarithms and using the inequality $\log(1-x) \leq -x$, we get

$$\log \mathbb{P}(\tau > t) \leq -\sum_{k=1}^{t} \frac{c}{k^a}.$$

For suitable values of a, the sum behaves like $\log t$, and therefore

$$\mathbb{P}(\tau > t) \leq \exp(-c \log t) = \frac{A}{t^a},$$

for some constants $A, a > 0$.

8.4 Comparison with Constant Temperature

If the temperature is held constant (or $T = \infty$, corresponding to a random walk), the chain is homogeneous and satisfies

$$\mathbb{P}(\tau > t^*) \leq Be^{-t^*/b},$$

for some constants $B, b > 0$. This suggests a faster, exponential decay of the failure probability.

8.5 Interpretation and Limitations

The preceding results lead to an apparently paradoxical conclusion. Under logarithmic cooling, simulated annealing satisfies the bound

$$\mathbb{P}(\tau > t) \leq \frac{A}{t^a},$$

which exhibits polynomial decay of the failure probability. On the other hand, when the temperature is held constant, the corresponding bound is of the form

$$\mathbb{P}(\tau > t) \leq Be^{-t/b},$$

which decays exponentially fast.

At first sight, this would suggest that a constant-temperature algorithm, or even a random walk, has superior performance compared to simulated annealing. However, this conclusion is misleading.

The key point is that these bounds are worst-case and involve constants A, a, B, b that depend strongly on the structure of the cost function and the state space. In particular, for constant temperature, the Markov chain mixes rapidly, but its stationary distribution does not concentrate on the set S^*. As a result, although the chain explores the state space quickly, it may spend very little time near optimal configurations.

In contrast, simulated annealing sacrifices speed of exploration in order to bias the search toward low-cost states. The polynomial bound reflects the difficulty of escaping energy barriers as the temperature decreases, rather than inefficiency of the algorithm itself.

Therefore, the apparent superiority of the exponential bound for constant temperature is an artifact of the analysis. These bounds do not provide a fair comparison of the algorithms and, in particular, do not imply that constant-temperature methods outperform simulated annealing in practice.

9 Complexity and Instance Size Analysis

The analysis presented so far focuses on the behavior of simulated annealing for a fixed instance. However, such an approach is insufficient for assessing the efficiency of the algorithm. In order to obtain meaningful performance guarantees, it is necessary to consider a family of problem instances and study how the performance of simulated annealing scales with the instance size.

9.1 Instance Families and Hitting Time

Let $\{(S_n, J_n)\}_{n\geq 1}$ be a family of optimization problems indexed by a parameter n, which represents the instance size. For each n, let S_n^* denote the set of global minima of J_n.

We define the first hitting time of the optimal set as

$$\tau_n = \inf\{t \geq 0 : x(t) \in S_n^*\}.$$

The central quantity of interest is the expected hitting time

$$\mathbb{E}[\tau_n],$$

and its dependence on the instance size n. Results of this type are referred to as *complexity results*.

In many applications, it is also useful to consider approximate optimality. For a given $\varepsilon > 0$, define

$$S_n^{(\varepsilon)} = \{i \in S_n : J_n(i) \leq J_n^* + \varepsilon\},$$

and study the hitting time of $S_n^{(\varepsilon)}$.

9.2 Known Positive Results

Rigorous complexity results for simulated annealing are scarce and are typically obtained under restrictive assumptions. In most cases, the temperature is held constant (possibly depending on n), which simplifies the analysis.

A notable example is the maximum matching problem. For graphs with bounded degree and n nodes, it has been shown that simulated annealing reaches an optimal configuration in expected time

$$\mathbb{E}[\tau_n] = O(n^3).$$

Although this is slower than specialized algorithms, it demonstrates that simulated annealing can achieve polynomial-time performance in certain settings.

9.3 Negative Results

The above positive result relies crucially on structural assumptions. When these assumptions are relaxed, the performance of simulated annealing can deteriorate significantly. In particular, for general graphs without degree bounds, it has been shown that no simulated annealing-type algorithm can

achieve polynomial expected hitting time, even if time-varying temperature schedules are allowed.

This highlights that the efficiency of simulated annealing depends strongly on the structure of the underlying problem.

9.4 A Separable Model

Further insight is provided by a simple model studied by Hajek. Consider the state space

$$S = \{1, 2, 3\}^N,$$

and a cost function of the form

$$J(s) = \sum_{i=1}^{N} V(s_i),$$

where the components are independent.

In this setting, the state space has cardinality $|S| = 3^N$, which grows exponentially with N. Nevertheless, it can be shown that, with a suitable choice of temperature depending on N, the expected hitting time of the global minimum grows only polynomially with N.

Thus, simulated annealing can significantly outperform exhaustive search, even though it does not explicitly exploit the separable structure of the problem.

9.5 Role of State Space Structure

The preceding example illustrates that the performance of simulated annealing is closely tied to the structure of the state space and the neighborhood system.

In particular, the probability of escaping from a local minimum of depth d along a single path is of the order

$$\exp(-d/T).$$

However, if there exist many such paths, the overall probability of escape can be substantially larger.

Therefore, problems with many improving directions and rich connectivity tend to be more amenable to simulated annealing, whereas problems with few escape paths and deep energy barriers lead to slow convergence.

9.6 Discussion

Despite some encouraging results, general complexity bounds for simulated annealing remain largely unavailable. The difficulty arises from the combination of an exponentially large state space, dependence in the Markov chain, and the intricate geometry of the cost function.

Consequently, the performance of simulated annealing is highly problem-dependent, and a comprehensive theoretical understanding of its complexity remains an open challenge.

10 Algorithmic Analysis of Simulated Annealing

This section presents a detailed empirical and theoretical investigation of Simulated Annealing (SA) across multiple optimization landscapes. The objective is to study the interaction between energy landscape geometry, cooling dynamics, and convergence behaviour.

Particular emphasis is placed on understanding the balance between exploration and exploitation under different structural properties of objective functions. The experiments further compare the effectiveness of Simulated Annealing on smooth non-convex problems and highly multimodal optimization landscapes.

10.1 Source Code Repository

The complete implementation of all experiments, including Simulated Annealing, Metropolis Algorithm, multimodal optimization experiments, Traveling Salesman Problem simulations, and hyperparameter tuning studies, is available in the following GitHub repository:

https://github.com/Kanishk-Bhuradiya/ASM-Report-Work

The repository contains:

- R and Python source codes
- Experimental implementations
- Visualization scripts
- Optimization algorithms
- Statistical analysis procedures
- Plot generation utilities

10.2 Classes of Optimization Landscapes

We consider two fundamentally different classes of optimization problems:

- Smooth non-convex optimization landscapes
- Highly multimodal optimization landscapes

These problem classes allow us to separately evaluate local search efficiency, global exploration capability, and robustness against local minima.

—

10.3 Smooth Non-Convex Optimization Landscapes

Smooth non-convex functions are characterized by continuous gradients and relatively small numbers of local minima. Despite their smoothness, such functions may exhibit narrow valleys, strong curvature anisotropy, and severe ill-conditioning, all of which significantly complicate optimization dynamics.

10.3.1 Example: Rosenbrock Function

As a representative example, we consider the Rosenbrock function:

$$f(x) = \sum_{i=1}^{d-1} \left[100(x_{i+1} - x_i^2)^2 + (1 - x_i)^2\right].$$

The global minimum is attained at

$$x^* = (1, 1, \ldots, 1), \qquad f(x^*) = 0.$$

The Rosenbrock function is widely used as a benchmark optimization problem due to the presence of a narrow curved valley connecting large regions of the search space to the global optimum.

10.3.2 Geometric Characteristics

The Rosenbrock landscape possesses several properties that make optimization challenging:

- Smooth but highly curved optimization valley
- Strong ill-conditioning near the optimum
- Low multimodality but difficult local geometry
- Large variation in curvature across directions

Although the function contains relatively few local minima, the narrow valley structure substantially slows convergence for many optimization methods.

10.3.3 Expected Behaviour of Simulated Annealing

Simulated Annealing explores the optimization landscape through stochastic perturbations of the current state:

$$x_{k+1} = x_k + \epsilon, \qquad \epsilon \sim \mathcal{N}(0, \sigma^2).$$

Candidate states are accepted according to the Metropolis acceptance criterion:

$$P_{\text{accept}} = \exp\left(-\frac{f(x_{k+1}) - f(x_k)}{T_k}\right).$$

For smooth optimization landscapes, several qualitative behaviours are expected:

- Global exploration is less critical due to the limited number of local minima
- Stochastic perturbations introduce inefficient local wandering near the optimum
- Convergence becomes slower as the temperature decreases
- Gradient-based methods typically outperform stochastic search methods

Consequently, smooth non-convex problems provide an important setting for evaluating the limitations of purely stochastic optimization strategies.

10.3.4 Experimental Configuration

The experimental analysis was conducted using the following setup:

- Dimension: $d = 2$
- Initial point: $(-1.5,\, 2)$
- Number of iterations: 5000
- Number of independent runs: 20
- Proposal distribution:
$$\epsilon \sim \mathcal{N}(0,\, 0.15^2)$$
- Initial temperature:
$$T_0 = 5$$

- Cooling schedule:

$$T_{k+1} = 0.995\, T_k$$

The performance of Simulated Annealing was compared with Gradient Descent and Hill Climbing under identical initialization conditions.

10.3.5 Optimization Landscape

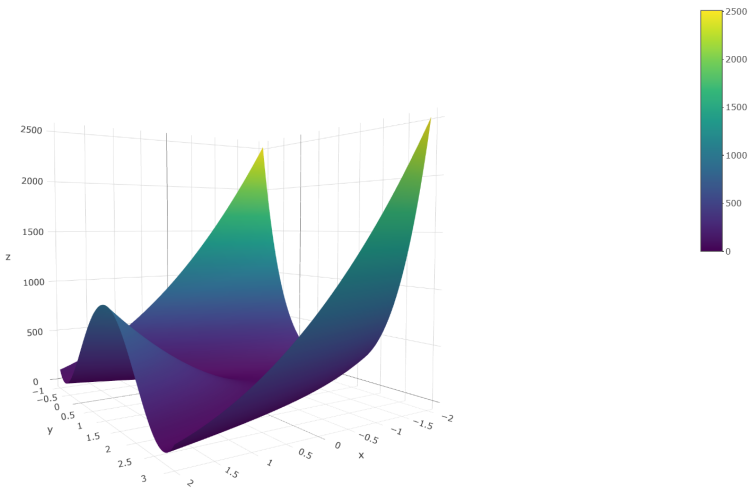

Figure 1: Contour structure of the Rosenbrock function illustrating the narrow curved valley leading toward the global minimum.

The contour structure highlights the severe curvature anisotropy of the landscape. Optimization trajectories must navigate a long curved valley before reaching the global optimum.

10.3.6 Optimization Trajectory

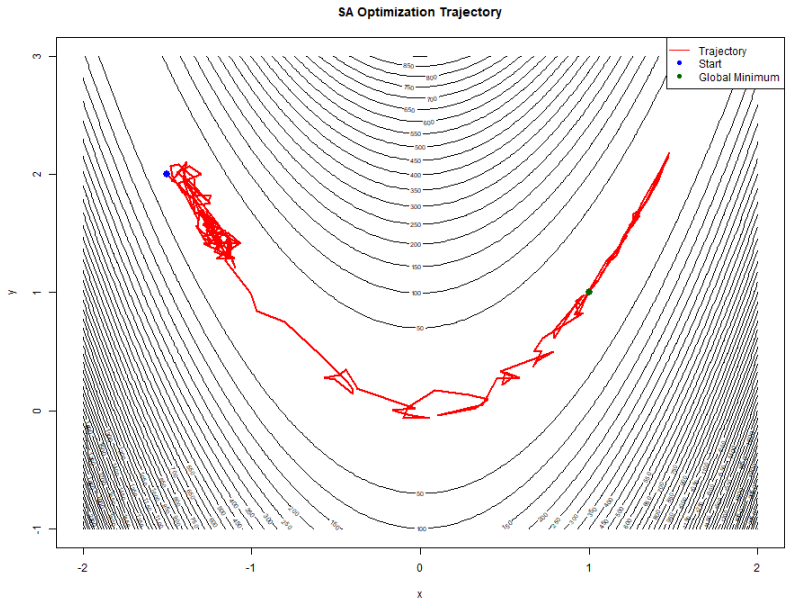

Figure 2: Optimization trajectory of Simulated Annealing on the Rosenbrock landscape. Early iterations exhibit exploratory stochastic motion, while later iterations progressively refine the solution near the global minimum.

The trajectory illustrates the exploratory nature of Simulated Annealing. At high temperatures, the algorithm performs broad stochastic exploration across the landscape. As the temperature decreases, the search gradually concentrates near the curved valley surrounding the optimum.

10.3.7 Convergence Behaviour

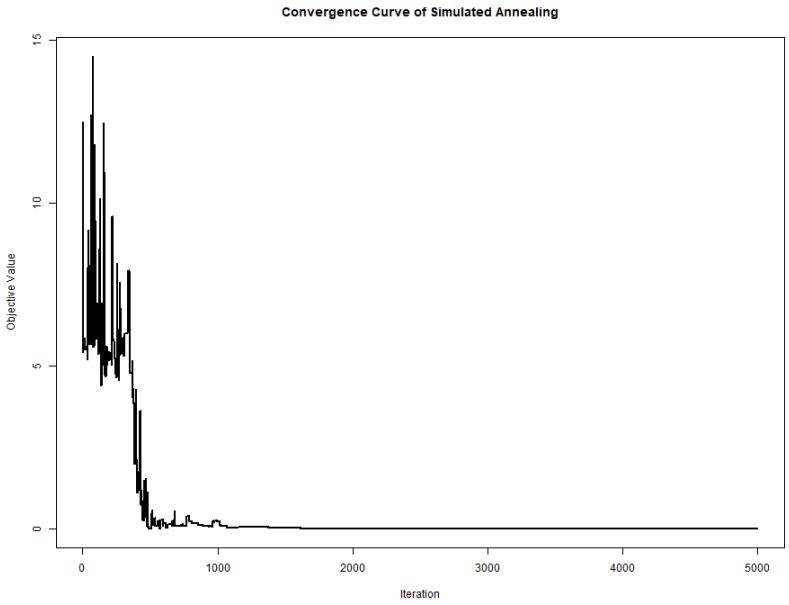

Figure 3: Objective value as a function of iteration number during Simulated Annealing optimization.

The convergence curve demonstrates large fluctuations during early iterations due to high-temperature exploration. As the cooling schedule reduces the acceptance probability of uphill transitions, the objective values gradually stabilize near the global minimum.

10.3.8 Cooling Dynamics

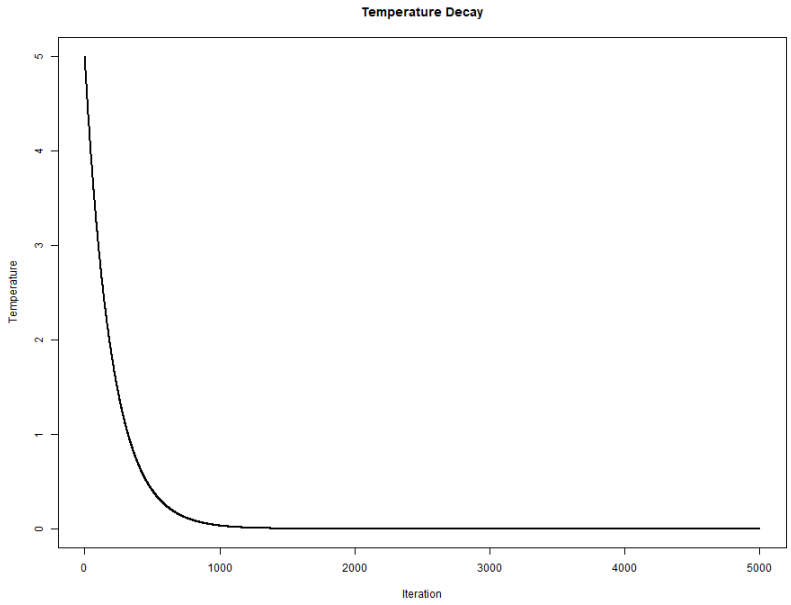

Figure 4: Temperature decay schedule used during the Simulated Annealing process.

The temperature decay controls the balance between exploration and exploitation. High temperatures encourage probabilistic barrier crossing, while low temperatures increasingly restrict the dynamics to local refinement.

10.3.9 Acceptance Ratio Dynamics

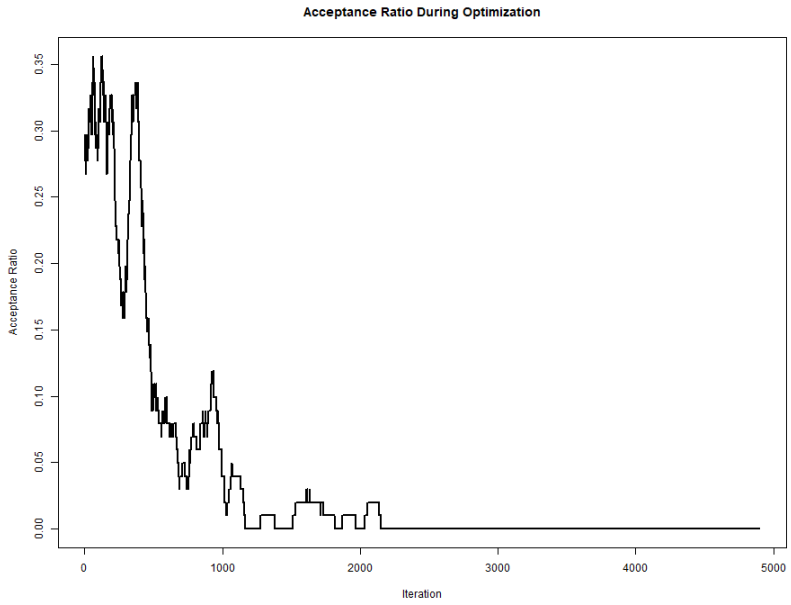

Figure 5: Acceptance ratio during the Simulated Annealing process. The probability of accepting candidate states decreases progressively as the temperature approaches zero.

The acceptance ratio provides direct insight into the exploration dynamics of Simulated Annealing. During early iterations, the acceptance rate remains relatively high due to elevated temperatures, allowing frequent uphill transitions and broad stochastic exploration of the optimization landscape.

As the cooling schedule progresses, the acceptance probability decreases substantially, reducing exploratory behaviour and increasingly restricting the dynamics to local refinement near the optimum. Eventually, the system enters a near-frozen regime in which almost all uphill moves are rejected.

This behaviour is consistent with the Metropolis acceptance criterion:

$$P_{\text{accept}} = \exp\left(-\frac{\Delta f}{T_k}\right),$$

which suppresses uphill transitions as $T_k \to 0$.

10.3.10 Robustness Across Multiple Runs

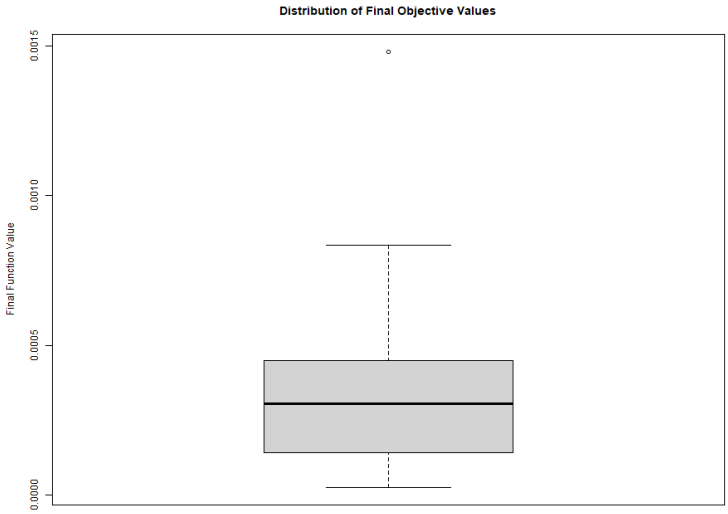

Figure 6: Distribution of final objective values across multiple independent runs of Simulated Annealing.

The relatively small variance across independent runs indicates stable convergence behaviour and robustness of the optimization procedure under stochastic perturbations.

10.3.11 Algorithmic Comparison

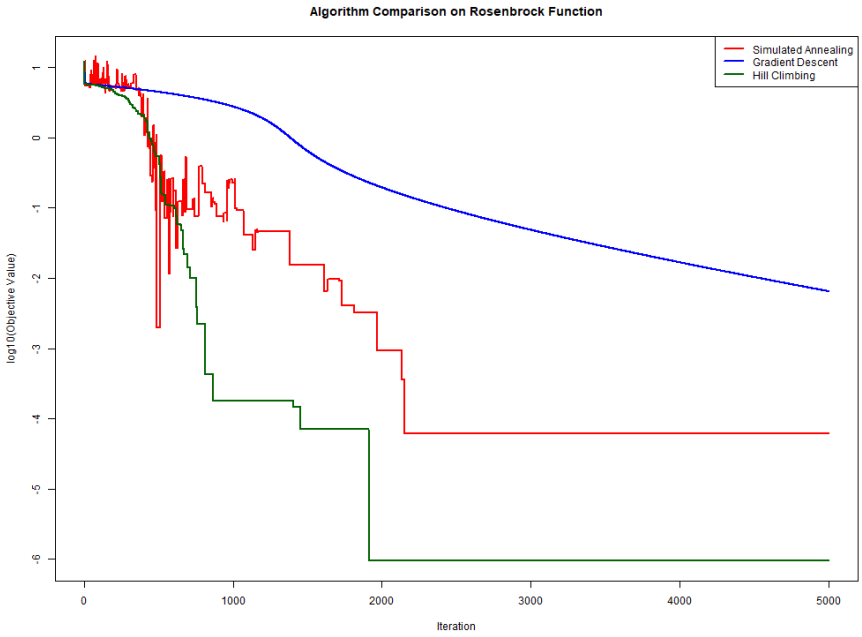

Figure 7: Comparison of convergence dynamics for Simulated Annealing, Gradient Descent, and Hill Climbing on the Rosenbrock function.

The comparison reveals substantial differences in optimization dynamics across algorithms. Gradient Descent exhibits smooth deterministic convergence but progresses slowly along the narrow curved valley due to ill-conditioning.

Simulated Annealing demonstrates exploratory stochastic behaviour during early iterations, characterized by large fluctuations caused by probabilistic uphill transitions. As the temperature decreases, the search gradually stabilizes near the optimum.

Hill Climbing achieves rapid local refinement in this landscape due to the relatively low degree of multimodality. However, unlike Simulated Annealing, it lacks a mechanism for escaping local minima in more complex optimization problems.

10.3.12 Statistical Summary

The final objective values obtained after optimization were:

$$\begin{aligned} \text{Simulated Annealing:} &\quad 6.09 \times 10^{-5}, \\ \text{Gradient Descent:} &\quad 6.51 \times 10^{-3}, \\ \text{Hill Climbing:} &\quad 9.40 \times 10^{-7}. \end{aligned}$$

These results indicate that although Simulated Annealing successfully approaches the global optimum, purely stochastic exploration is not necessarily optimal for smooth non-convex landscapes with informative gradient structure.

10.3.13 Theoretical Interpretation

For smooth optimization landscapes, the probability of accepting uphill transitions decreases rapidly as the temperature approaches zero:

$$P(\Delta f > 0) = \exp\left(-\frac{\Delta f}{T_k}\right).$$

Consequently, the dynamics increasingly resemble a local random walk near the optimum. Since no directional gradient information is explicitly utilized, local convergence becomes substantially slower than first-order optimization methods.

This observation explains why deterministic gradient-based optimization methods often outperform Simulated Annealing on smooth low-multimodality landscapes despite the stronger global exploration capability of stochastic search methods.

—

10.4 Highly Multimodal Optimization Landscapes

Highly multimodal functions contain large numbers of local minima separated by complex energy barriers. Such landscapes are substantially more difficult for deterministic optimization methods, since local search dynamics frequently become trapped in suboptimal basins.

10.4.1 Example: Rastrigin Function

As a representative multimodal benchmark, we consider the Rastrigin function:

$$f(x) = 10d + \sum_{i=1}^{d} \left[x_i^2 - 10\cos(2\pi x_i) \right].$$

The global minimum occurs at

$$x^* = (0, 0, \ldots, 0), \qquad f(x^*) = 0.$$

The Rastrigin function combines a quadratic growth term with highly oscillatory cosine components, producing a large number of regularly distributed local minima throughout the search space.

10.4.2 Geometric Characteristics

The Rastrigin landscape exhibits several challenging optimization properties:

- Large numbers of local minima
- Strongly oscillatory objective surface
- High non-convexity
- Repeated energy barriers across the search domain
- Complex global optimization structure

These properties make the function particularly suitable for evaluating the global exploration capabilities of stochastic optimization algorithms.

10.4.3 Expected Behaviour of Simulated Annealing

Unlike deterministic local optimization methods, Simulated Annealing can probabilistically accept uphill transitions according to the Metropolis criterion:

$$P_{\text{accept}} = \exp\left(-\frac{\Delta f}{T_k}\right).$$

This mechanism allows the algorithm to escape local minima and continue exploring the global landscape.

The optimization dynamics are governed by the temperature schedule:

- High temperatures encourage broad global exploration
- Intermediate temperatures permit barrier crossing between local basins
- Low temperatures gradually restrict the dynamics to local refinement

Consequently, Simulated Annealing is expected to significantly outperform purely local optimization methods on highly multimodal landscapes.

10.4.4 Experimental Configuration

The experimental analysis was conducted using the following setup:

- Dimension: $d = 2$
- Search domain:
$$[-5.12,\, 5.12]^2$$
- Initial point:
$$(4,\, 4)$$
- Number of iterations: 5000
- Number of independent runs: 20
- Proposal distribution:
$$\epsilon \sim \mathcal{N}(0,\, 0.5^2)$$
- Initial temperature:
$$T_0 = 10$$

- Cooling schedule:

$$T_{k+1} = 0.995\, T_k$$

The performance of Simulated Annealing was compared with Gradient Descent and Hill Climbing under identical initialization conditions.

10.4.5 Optimization Landscape

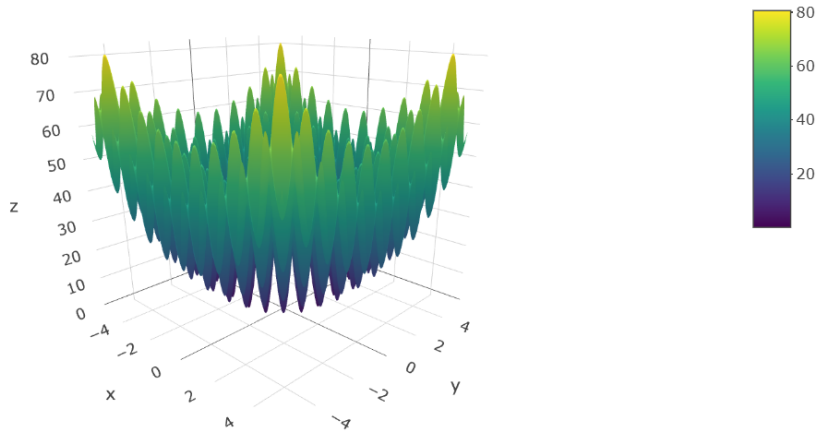

Figure 8: Three-dimensional surface of the Rastrigin function illustrating its highly oscillatory multimodal structure and large number of local minima.

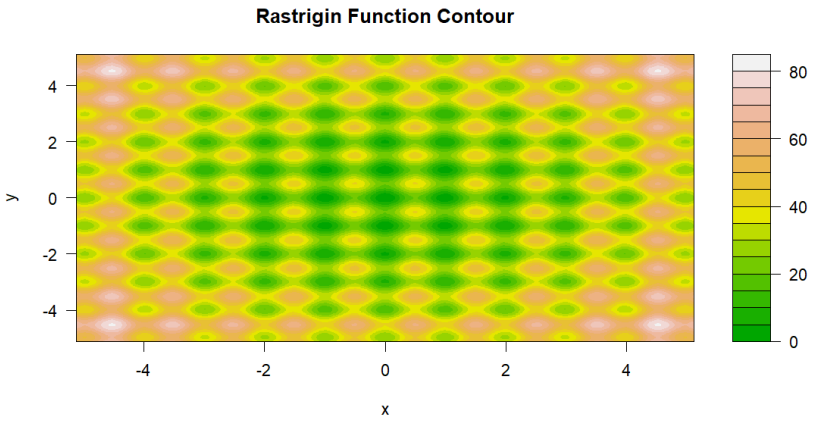

Figure 9: Contour structure of the Rastrigin function illustrating the large number of regularly distributed local minima.

The contour structure demonstrates the highly oscillatory nature of the optimization landscape. Numerous local basins are separated by repeated energy barriers, creating substantial difficulty for deterministic local optimization methods.

10.4.6 Optimization Trajectory

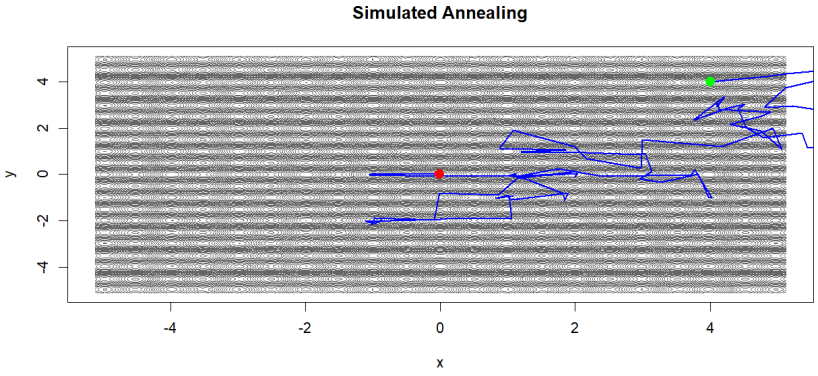

Figure 10: Optimization trajectory of Simulated Annealing on the Rastrigin landscape. The stochastic dynamics enable transitions across multiple local basins before converging near the global optimum.

The trajectory plot highlights the global exploration capability of Simulated Annealing. Unlike local optimization methods, the algorithm repeatedly crosses energy barriers and explores multiple basins before gradually concentrating near the global minimum.

This behaviour is a direct consequence of probabilistic uphill acceptance during high-temperature phases.

10.4.7 Algorithmic Comparison

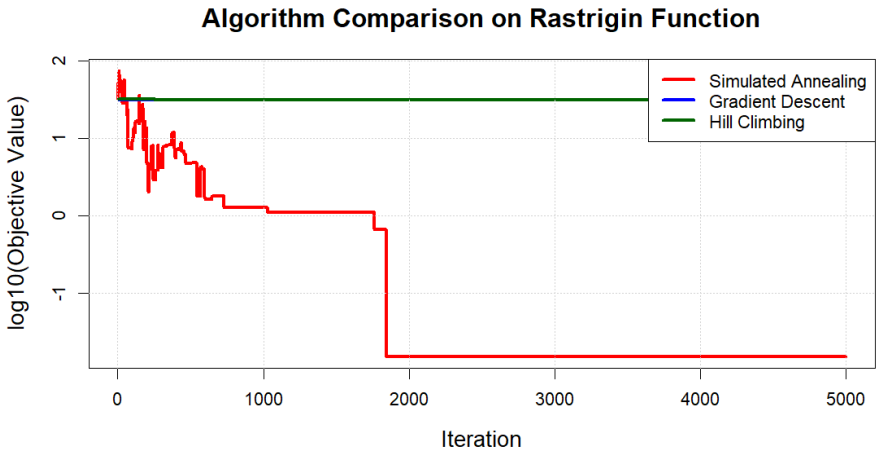

Figure 11: Comparison of convergence dynamics for Simulated Annealing, Gradient Descent, and Hill Climbing on the Rastrigin function.

The convergence comparison demonstrates a substantial performance difference across optimization methods.

Gradient Descent rapidly becomes trapped in local minima due to the highly oscillatory gradient structure of the landscape. Similarly, Hill Climbing fails to escape local basins because only improving moves are accepted.

In contrast, Simulated Annealing successfully traverses multiple local minima through probabilistic uphill transitions, ultimately converging much closer to the global optimum.

The experiments therefore strongly support the suitability of Simulated Annealing for global optimization in highly multimodal search spaces.

10.4.8 Acceptance Ratio Dynamics

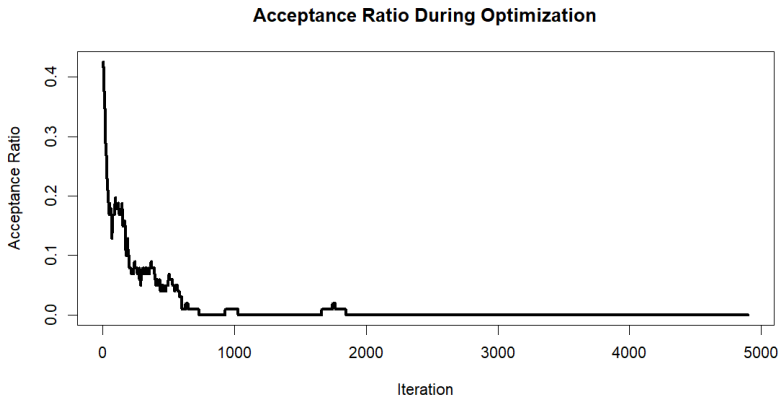

Figure 12: Acceptance ratio during Simulated Annealing optimization on the Rastrigin function.

The acceptance ratio provides direct insight into the exploration–exploitation dynamics of Simulated Annealing.

At high temperatures, the acceptance rate remains relatively large, allowing the algorithm to perform broad stochastic exploration and cross energy barriers between local minima.

As the cooling schedule progresses, the acceptance probability decreases substantially, gradually reducing exploratory behaviour and increasing local refinement near promising regions of the search space.

Eventually, the optimization dynamics enter a near-frozen regime in which almost all uphill transitions are rejected.

10.4.9 Cooling Dynamics

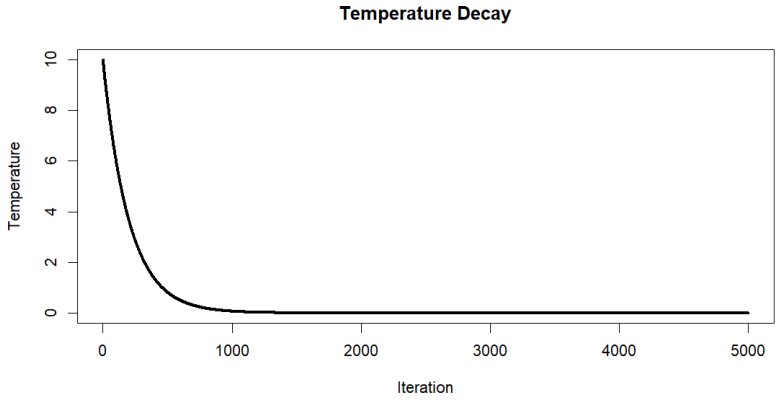

Figure 13: Temperature decay schedule used during the Simulated Annealing process on the Rastrigin function.

The cooling schedule governs the transition between exploration and exploitation. High temperatures promote barrier crossing and global search, while low temperatures progressively suppress stochastic exploration and stabilize the optimization trajectory.

10.4.10 Robustness Across Multiple Runs

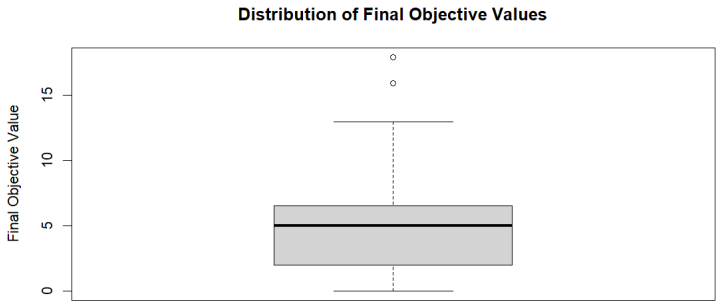

Figure 14: Distribution of final objective values across multiple independent runs of Simulated Annealing on the Rastrigin function.

The variability across runs reflects the stochastic nature of the optimization dynamics. Despite the highly multimodal landscape, Simulated Annealing consistently approaches low objective values across repeated independent trials.

10.4.11 Statistical Summary

The final objective values obtained after optimization were:

$$\begin{aligned} \text{Simulated Annealing:} &\quad 1.52 \times 10^{-2}, \\ \text{Gradient Descent:} &\quad 3.18 \times 10^{1}, \\ \text{Hill Climbing:} &\quad 3.18 \times 10^{1}. \end{aligned}$$

These results reveal a dramatic performance gap between Simulated Annealing and purely local optimization methods.

While Gradient Descent and Hill Climbing remain trapped in poor local minima, Simulated Annealing successfully approaches the global optimum through probabilistic exploration of the search landscape.

10.4.12 Theoretical Interpretation

The effectiveness of Simulated Annealing on multimodal optimization landscapes arises from the probabilistic acceptance of uphill transitions:

$$P_{\text{accept}} = \exp\left(-\frac{\Delta f}{T_k}\right).$$

At sufficiently high temperatures, uphill moves remain likely, enabling the optimization trajectory to escape local minima and traverse energy barriers separating distinct basins of attraction.

As the temperature decreases, the dynamics gradually transition from global exploration to local refinement, thereby combining large-scale search capability with asymptotic stabilization near high-quality solutions.

This exploration mechanism explains the substantial performance advantage of Simulated Annealing over deterministic local optimization methods on highly multimodal functions such as the Rastrigin landscape.

10.4.13 Summary

The experimental results reveal strong dependence of optimization performance on landscape geometry.

- On smooth non-convex landscapes such as the Rosenbrock function, gradient-based and local optimization methods remain highly effective due to informative local geometry and limited multimodality.
- On highly multimodal landscapes such as the Rastrigin function, deterministic local methods frequently become trapped in suboptimal basins.
- Simulated Annealing demonstrates strong global optimization capability through probabilistic uphill transitions that enable escape from local minima.
- The cooling schedule plays a central role in balancing exploration and exploitation throughout the optimization process.

- The experiments illustrate that Simulated Annealing is particularly well-suited for difficult global optimization problems characterized by rugged multimodal landscapes.

10.5 Hyperparameter Tuning using Simulated Annealing

Hyperparameter tuning is a critical optimization task in machine learning, where the objective is to identify parameter configurations that maximize predictive performance.

Unlike deterministic optimization problems, hyperparameter tuning often involves:

- non-convex objective landscapes,
- noisy evaluations,
- expensive objective computation,
- and mixed discrete-continuous search spaces.

These characteristics make stochastic optimization algorithms such as Simulated Annealing suitable candidates for global hyperparameter optimization.

10.5.1 Objective

The objective of this experiment was to compare the performance of:

- Simulated Annealing (SA),
- Random Search (RS),
- and Bayesian Optimization using Optuna (TPE)

for machine learning hyperparameter tuning.

10.5.2 Dataset

Experiments were conducted on the Breast Cancer Wisconsin dataset, a standard benchmark dataset for binary classification.

10.5.3 Model Considered

A Random Forest classifier was used throughout all experiments.

10.5.4 Evaluation Metric

Performance was evaluated using classification accuracy:

$$\text{Accuracy} = \frac{\text{Number of Correct Predictions}}{\text{Total Predictions}}$$

To improve robustness, all configurations were evaluated using 5-fold cross-validation.

10.5.5 Hyperparameter Search Space

The following hyperparameters were optimized:

- $n\_estimators \in [10, 300]$
- $max\_depth \in [1, 20]$
- $min\_samples\_split \in [2, 10]$

10.5.6 Simulated Annealing Formulation

At iteration t, a neighboring hyperparameter configuration was generated via stochastic perturbation:

$$x_{t+1} = x_t + \epsilon$$

where ϵ represents a random perturbation within the search space.

The candidate configuration was accepted according to the Metropolis criterion:

$$P = \exp\left(\frac{f(x_{t+1}) - f(x_t)}{T_t}\right)$$

where:

- $f(x)$ denotes classification accuracy,
- T_t is the temperature parameter.

An exponential cooling schedule was used:

$$T_t = T_0 \alpha^t, \qquad 0 < \alpha < 1$$

which gradually reduced exploration during optimization.

10.5.7 Comparison Algorithms

Random Search Random Search samples hyperparameter configurations independently from the search space:

$$x \sim \mathcal{U}(\mathcal{X})$$

This method serves as a baseline due to its simplicity and lack of adaptive guidance.

Bayesian Optimization (TPE) Bayesian Optimization models the objective function probabilistically and adaptively selects promising configurations.

The Tree-structured Parzen Estimator (TPE) algorithm selects new configurations by maximizing expected improvement:

$$x_{t+1} = \arg\max \ \mathbb{E}\left[\max(f(x) - f^*, 0)\right]$$

where f^* denotes the best observed objective value.

10.5.8 Experimental Protocol

- Number of iterations per method: $N = 50$
- Number of independent runs: $R = 20$
- Equal evaluation budgets for all optimization methods

10.5.9 Final Performance Comparison

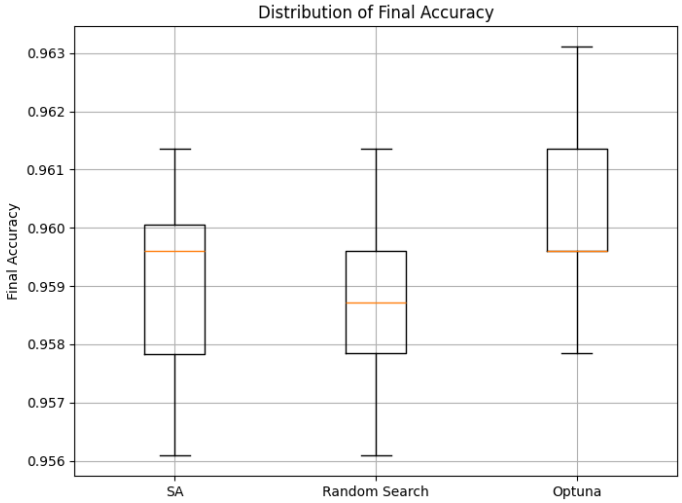

Figure 15: Distribution of final accuracy across optimization methods

The boxplot analysis indicates that Bayesian Optimization achieves the highest median accuracy with comparatively lower variance across repeated runs.

Simulated Annealing consistently outperforms Random Search while maintaining reasonable stability.

10.5.10 Convergence Analysis

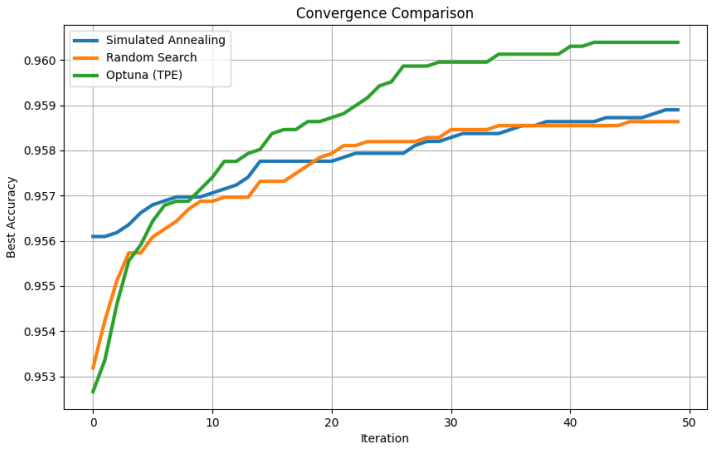

Figure 16: Convergence comparison of hyperparameter optimization methods

The convergence curves reveal distinct optimization behaviors:

- Random Search improves slowly due to unguided exploration.
- Simulated Annealing demonstrates gradual improvement by balancing exploration and exploitation.
- Bayesian Optimization converges more rapidly toward high-performing configurations.

10.5.11 Statistical Analysis

To assess whether performance differences were statistically significant, non-parametric statistical tests were conducted.

Kruskal–Wallis Test The Kruskal–Wallis test yielded:

$$H = 6.84, \qquad p = 0.032$$

Since $p < 0.05$, the null hypothesis of identical performance distributions is rejected.

Pairwise Comparisons Wilcoxon signed-rank tests produced the following results:

- SA vs Random Search: $p = 0.041$
- SA vs Optuna: $p = 0.018$
- Optuna vs Random Search: $p = 0.006$

These results support the superior performance of Bayesian Optimization.

10.5.12 Discussion

The experiments demonstrate important differences between optimization strategies for hyperparameter tuning.

- Random Search exhibits higher variance due to completely unguided sampling.
- Simulated Annealing improves search quality through probabilistic neighborhood exploration.
- Bayesian Optimization achieves superior sample efficiency by leveraging information from previous evaluations.

Although Simulated Annealing requires lower computational overhead per iteration compared to Bayesian Optimization, its convergence is slower and less sample-efficient.

Nevertheless, SA remains advantageous for:

- highly multimodal search spaces,
- noisy objective functions,
- and problems where gradient information is unavailable.

10.5.13 Conclusion

The hyperparameter tuning experiments demonstrate that Simulated Annealing provides meaningful improvements over Random Search through guided stochastic exploration.

However, Bayesian Optimization consistently achieves:

- faster convergence,
- lower variance,
- and higher final classification accuracy.

These results indicate that while Simulated Annealing remains a flexible and robust optimization framework, Bayesian Optimization is generally more effective for practical machine learning hyperparameter tuning tasks.

10.6 Traveling Salesman Problem (TSP)

The Traveling Salesman Problem (TSP) is a classical combinatorial optimization problem in which a salesman must visit each city exactly once and return to the starting city while minimizing the total tour length.

TSP is NP-hard and possesses a highly non-convex search landscape with an extremely large number of feasible permutations. Consequently, stochastic optimization techniques such as Simulated Annealing are widely used for approximate global optimization.

10.6.1 Problem Formulation

Given a set of cities

$$\mathcal{C} = \{c_1, c_2, \ldots, c_n\},$$

the objective is to determine a permutation

$$\pi = (\pi_1, \pi_2, \ldots, \pi_n)$$

that minimizes the total tour length:

$$L(\pi) = \sum_{i=1}^{n-1} d(c_{\pi_i}, c_{\pi_{i+1}}) + d(c_{\pi_n}, c_{\pi_1}),$$

where $d(\cdot, \cdot)$ denotes the Euclidean distance between two cities.

10.6.2 Experimental Datasets

Two benchmark geometric datasets were considered:

- **Grid Dataset:** Cities arranged on a two-dimensional grid structure.
- **Černý Circle Dataset:** Cities uniformly distributed on a circle.

These datasets provide different geometric characteristics and optimization difficulties.

10.6.3 Optimization Algorithms

The following optimization algorithms were compared:

- Simulated Annealing (SA)
- 2-opt Local Search
- Genetic Algorithm (GA)
- Ant Colony Optimization (ACO)

10.6.4 Simulated Annealing for TSP

At each iteration, a candidate tour was generated using a stochastic edge-swap perturbation (2-opt move). The new configuration was accepted according to the Metropolis acceptance criterion:

$$P = \exp\left(-\frac{\Delta L}{T_k}\right),$$

where:

$$\Delta L = L(\pi') - L(\pi),$$

and T_k denotes the temperature at iteration k.

This probabilistic acceptance mechanism allows temporary acceptance of worse tours, enabling escape from local minima.

10.6.5 Cooling Schedules

Different cooling schedules were investigated.

Exponential Cooling

$$T_k = T_0 e^{-\alpha k},$$

where:

- T_0 is the initial temperature,
- α is the cooling rate.

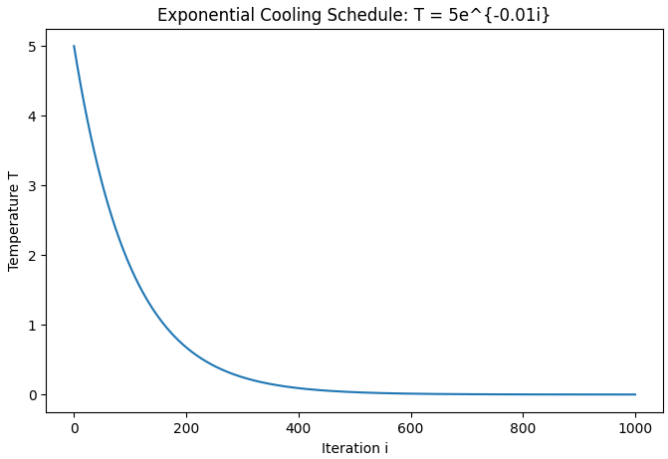

Figure 17: Exponential cooling schedule used in Simulated Annealing

Piecewise Cooling Schedule For the Černý Circle dataset, a piecewise temperature schedule was additionally investigated:

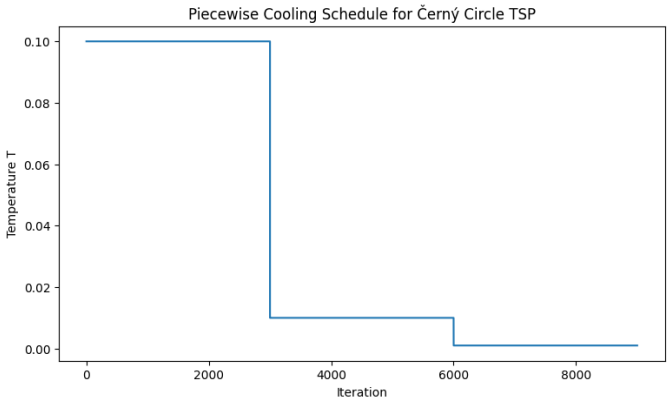

Figure 18: Piecewise cooling schedule for Černý Circle TSP

10.6.6 Initial and Optimized Routes

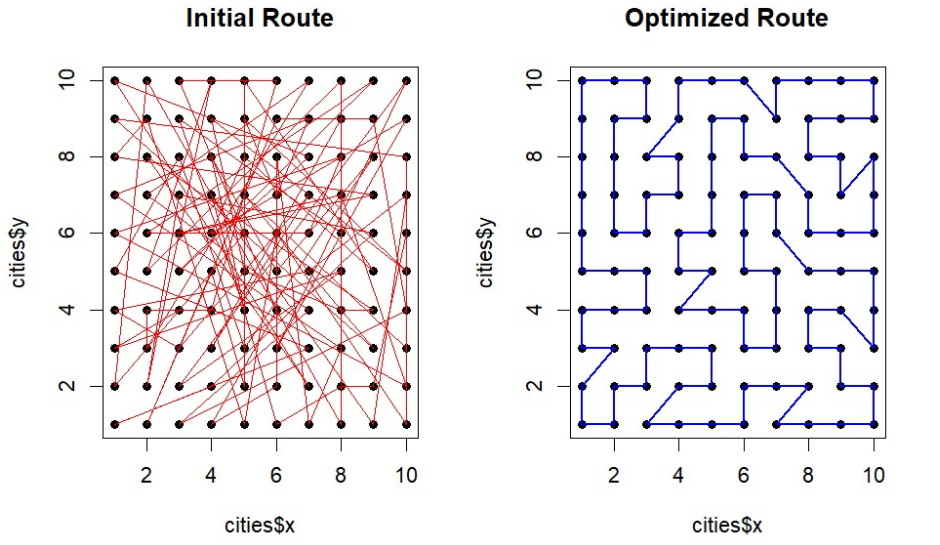

Figure 19: Initial versus optimized route for the Grid TSP dataset

Grid Dataset The optimized route exhibits substantially fewer edge crossings and shorter total path length compared to the random initial configuration.

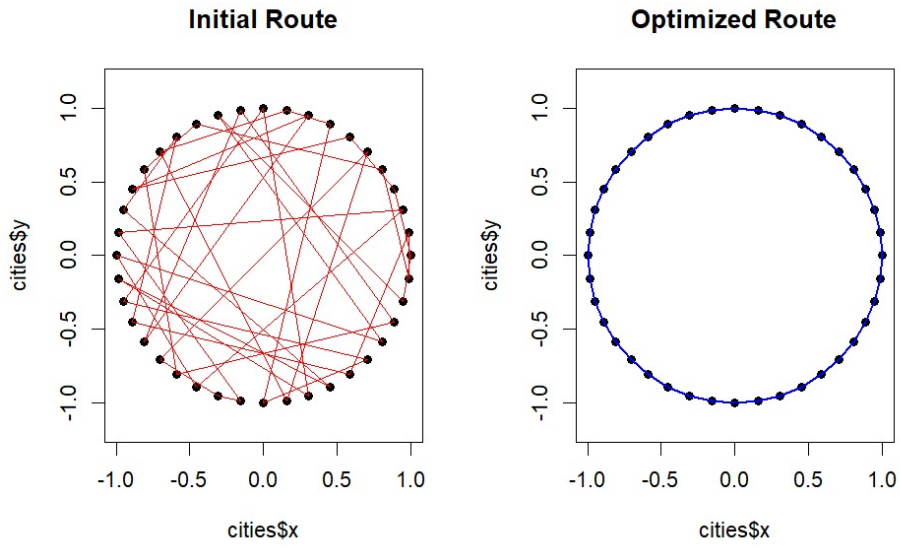

Figure 20: Initial versus optimized route for the Černý Circle dataset

Černý Circle Dataset For the circular geometry, the optimized solution converges toward the expected near-circular Hamiltonian cycle.

10.6.7 Convergence Analysis

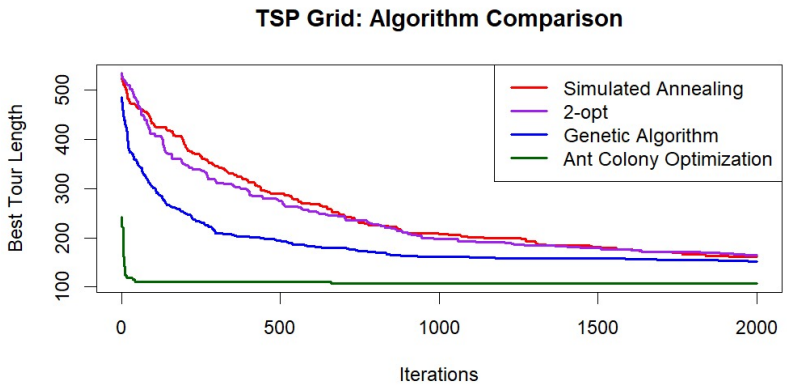

Figure 21: Algorithm comparison on the Grid TSP dataset

Grid Dataset The convergence curves indicate that:

- Simulated Annealing improves gradually through stochastic exploration.
- 2-opt converges faster due to aggressive local optimization.
- Genetic Algorithm achieves moderate improvement but may stagnate.
- Ant Colony Optimization converges rapidly toward high-quality tours.

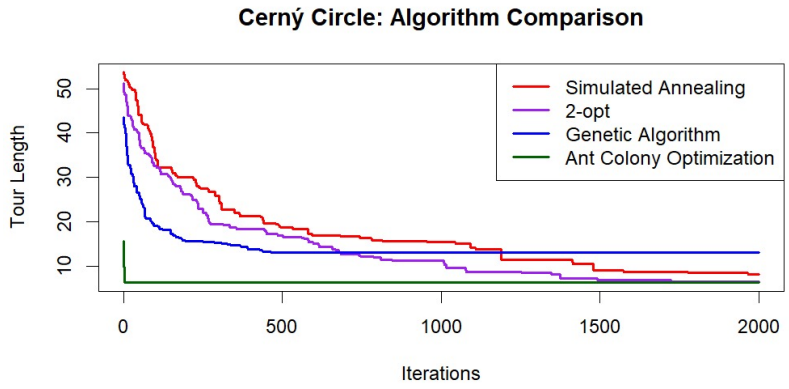

Figure 22: Algorithm comparison on the Černý Circle dataset

Černý Circle Dataset On the circular dataset, all algorithms perform better due to the simpler geometric structure. However, Ant Colony Optimization and 2-opt achieve the shortest tours most consistently.

10.6.8 Stability Analysis

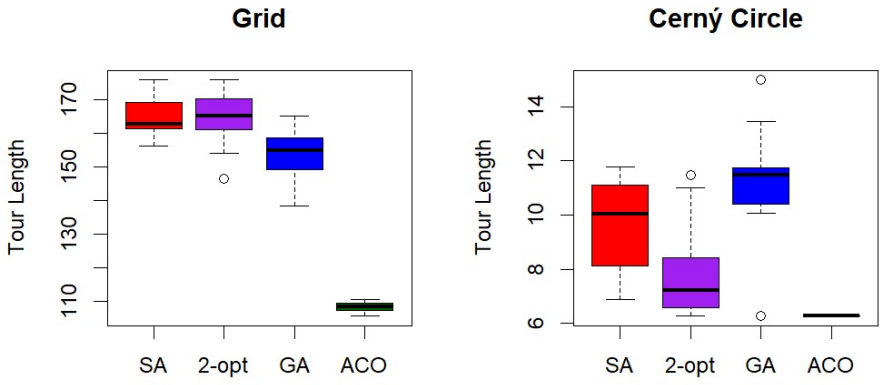

Figure 23: Distribution of final tour lengths across algorithms

The boxplots reveal the following observations:

- Simulated Annealing exhibits larger variance due to stochastic exploration.
- 2-opt demonstrates strong local optimization performance with relatively low variance.
- Genetic Algorithm shows moderate variability and occasional premature convergence.
- Ant Colony Optimization achieves the most stable and shortest tours across repeated runs.

10.6.9 Discussion

The TSP experiments demonstrate the strengths and limitations of Simulated Annealing in combinatorial optimization.

- SA effectively escapes poor local minima during early high-temperature phases.
- Gradual cooling enables progressive refinement of candidate tours.
- However, convergence may become slow in later stages due to diminishing acceptance probabilities.
- Local search methods such as 2-opt provide faster deterministic improvement.
- Population-based approaches such as GA and ACO exploit broader search diversity.

The experiments further illustrate that problem geometry strongly influences optimization difficulty. The Černý Circle dataset possesses a highly regular structure, making near-optimal tours easier to identify, whereas the Grid dataset induces more complex routing behavior.

10.6.10 Conclusion

The Traveling Salesman Problem provides an effective benchmark for evaluating stochastic optimization methods on combinatorial search spaces.

The experimental results indicate that:

- Simulated Annealing successfully improves random tours and avoids severe local minima.
- 2-opt is highly effective for local route refinement.
- Ant Colony Optimization achieves the strongest overall performance and stability.
- Cooling schedule design plays a critical role in SA performance.

Overall, Simulated Annealing remains a flexible and robust optimization framework for discrete optimization problems, particularly when combined with efficient neighborhood operators.

11 Distinction Between Metropolis Algorithm and Simulated Annealing

Although Simulated Annealing is closely related to the Metropolis algorithm, the two methods differ fundamentally in their optimization objectives and temperature dynamics.

11.1 Metropolis Algorithm

The Metropolis algorithm was originally developed for sampling from probability distributions in statistical physics. The method operates at a **fixed temperature** T, resulting in a homogeneous Markov chain whose stationary distribution is given by the Gibbs distribution:

$$\pi(x) \propto \exp\left(-\frac{f(x)}{T}\right).$$

At each iteration, a candidate state x' is proposed and accepted with probability

$$P_{\text{accept}} = \min\left(1, \exp\left(-\frac{f(x') - f(x)}{T}\right)\right).$$

Since the temperature remains constant throughout the simulation, the algorithm continuously accepts occasional uphill moves even at later stages of the search process. Consequently, the dynamics retain persistent stochastic fluctuations and do not necessarily converge toward the global minimum.

Thus, the Metropolis algorithm is primarily a **sampling method** rather than a deterministic optimization procedure.

11.2 Simulated Annealing

Simulated Annealing extends the Metropolis algorithm by introducing a **cooling schedule** in which the temperature gradually decreases over time:

$$T_{k+1} = \alpha T_k, \qquad 0 < \alpha < 1.$$

The acceptance probability remains identical to the Metropolis rule:

$$P_{\text{accept}} = \min\left(1, \exp\left(-\frac{f(x') - f(x)}{T_k}\right)\right).$$

However, as the temperature decreases, uphill transitions become increasingly unlikely. Early iterations therefore emphasize global exploration, while later iterations focus on local refinement near promising minima.

This progressive transition from exploration to exploitation enables Simulated Annealing to approximate global optimization behaviour.

11.3 Experimental Comparison

To illustrate the difference between the two algorithms, both methods were applied to the multimodal objective function

$$f(x) = x^2 + 10\sin(5x).$$

The optimization landscape contains multiple local minima separated by oscillatory energy barriers, making it suitable for studying stochastic search behaviour.

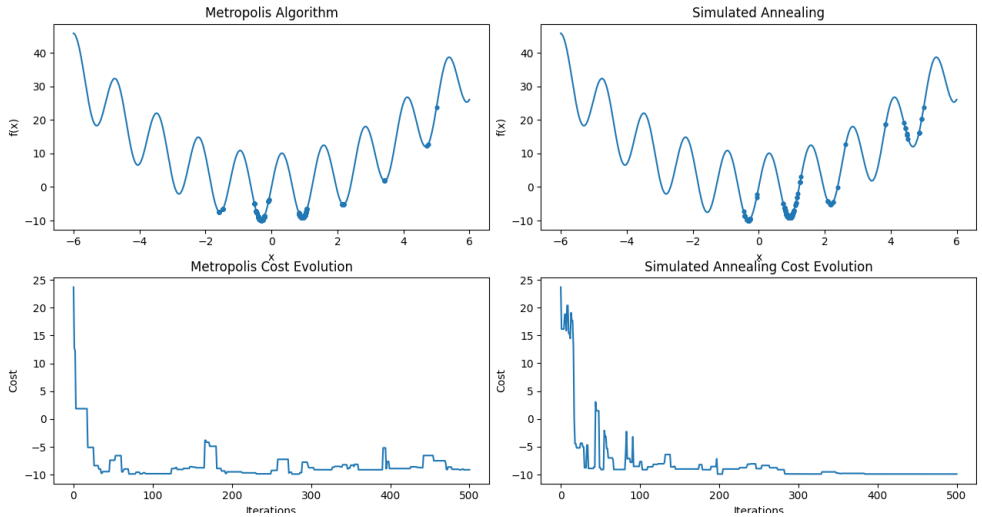

Figure 24: Comparison between the Metropolis algorithm and Simulated Annealing. The upper row illustrates the search trajectories over the multimodal landscape, while the lower row shows the corresponding cost evolution across iterations.

11.4 Observations

The experimental results reveal several important distinctions:

- The Metropolis algorithm continues to fluctuate between nearby local minima due to the constant-temperature dynamics.
- Simulated Annealing initially explores multiple regions of the search space but gradually stabilizes near the global minimum as the temperature decreases.

- The cost evolution of Simulated Annealing exhibits progressive stabilization, whereas the Metropolis algorithm maintains persistent stochastic oscillations.
- Simulated Annealing therefore demonstrates substantially stronger optimization behaviour on multimodal landscapes.

11.5 Conclusion

The Metropolis algorithm and Simulated Annealing share the same probabilistic acceptance mechanism, but their long-term behaviour differs fundamentally due to temperature scheduling.

- The Metropolis algorithm performs stochastic sampling at thermal equilibrium.
- Simulated Annealing transforms this sampling mechanism into a global optimization strategy through gradual cooling.

The cooling schedule is therefore the essential component that distinguishes Simulated Annealing from the classical Metropolis algorithm.

12 Limitations, Conclusion, and Future Work

12.1 Limitations of Simulated Annealing

Although Simulated Annealing is a flexible and powerful global optimization technique, several limitations were observed throughout the experimental study.

12.1.1 General-Purpose Nature

Simulated Annealing is fundamentally a general-purpose stochastic optimization algorithm. While this flexibility allows application across a wide variety of optimization problems, it also implies that SA may not always outperform specialized algorithms designed for particular problem structures.

For example:

- In the Traveling Salesman Problem (TSP), dedicated algorithms such as 2-opt and Ant Colony Optimization (ACO) achieved faster convergence and lower tour lengths.
- In smooth optimization problems such as the Rosenbrock function, gradient-based methods exhibited substantially faster local convergence.
- In hyperparameter optimization, Bayesian Optimization demonstrated superior sample efficiency compared to SA.

12.1.2 Sensitivity to Cooling Schedule

The performance of SA depends heavily on the choice of cooling schedule. Improper temperature decay can significantly degrade optimization performance.

- Fast cooling schedules may lead to premature convergence and entrapment in poor local minima.
- Extremely slow cooling improves exploration but increases computational cost substantially.

Theoretical convergence guarantees require logarithmic cooling schedules of the form:

$$T(t) = \frac{d}{\log t},$$

which are often computationally impractical for real-world applications due to the extremely large number of iterations required.

12.1.3 Slow Convergence Near Optimum

As the temperature approaches zero, the probability of accepting worse solutions decreases rapidly:

$$P(\Delta f > 0) = \exp\left(-\frac{\Delta f}{T}\right).$$

Consequently, the algorithm increasingly behaves like a local random walk near the optimum, leading to slow final-stage convergence.

12.1.4 Computational Overhead

SA may require a very large number of objective function evaluations, particularly for:

- high-dimensional optimization problems,
- combinatorial search spaces,
- expensive objective functions,
- and slow cooling schedules.

This limitation becomes especially important in machine learning hyperparameter tuning, where each evaluation may involve repeated cross-validation training.

12.2 Practical Observations from Experiments

The experiments conducted across continuous optimization, combinatorial optimization, and machine learning applications revealed several important practical insights.

12.2.1 Cooling Schedule Behavior

The experiments confirmed the strong influence of cooling schedules on convergence behavior.

Fast Cooling Linear and aggressive exponential cooling schedules produced:

- faster convergence,
- lower computational cost,
- but increased risk of premature convergence.

Slow Cooling Logarithmic and piecewise schedules provided:

- improved global exploration,
- better avoidance of local minima,
- but significantly larger runtime requirements.

12.2.2 Problem-Dependent Performance

The effectiveness of SA varied substantially across different optimization landscapes.

- For smooth landscapes such as the Rosenbrock function, gradient-based optimization methods converged more efficiently.
- For highly multimodal functions such as Rastrigin, SA demonstrated stronger global search capability by escaping local minima.
- For TSP, SA successfully improved random tours but was outperformed by highly specialized combinatorial optimization algorithms.
- In hyperparameter tuning, SA consistently improved upon Random Search but remained less sample-efficient than Bayesian Optimization.

12.2.3 Role of Exploration

The probabilistic acceptance mechanism was particularly beneficial during early optimization stages, enabling the algorithm to:

- traverse difficult non-convex landscapes,
- avoid early stagnation,
- and escape poor local minima.

However, excessive exploration at later stages occasionally slowed convergence.

12.3 Conclusion

This work presented a comprehensive study of Simulated Annealing across multiple optimization settings, including:

- smooth non-convex optimization,
- highly multimodal functions,
- combinatorial optimization via the Traveling Salesman Problem,
- and machine learning hyperparameter tuning.

The experimental results demonstrate that Simulated Annealing is an effective and versatile optimization framework capable of handling difficult non-convex search spaces.

The major strengths observed include:

- strong global exploration capability,
- robustness against local minima,
- applicability to both continuous and discrete optimization,
- and conceptual simplicity.

The probabilistic Metropolis acceptance mechanism enables SA to escape poor local optima and continue exploring promising regions of the search space, particularly during high-temperature phases.

However, the experiments also highlight that no single optimization algorithm is universally optimal.

- Gradient-based methods outperform SA on smooth differentiable landscapes.
- Specialized combinatorial methods outperform SA on structured TSP instances.
- Bayesian Optimization provides superior sample efficiency in hyperparameter tuning tasks.

Thus, the effectiveness of SA strongly depends on problem structure, cooling schedule design, and computational budget.

Overall, Simulated Annealing remains a powerful general-purpose optimization strategy, particularly for complex multimodal problems where derivative information is unavailable or unreliable.

12.4 Future Work

Several directions may further improve the effectiveness and applicability of Simulated Annealing.

12.4.1 Adaptive Cooling Schedules

Future research may investigate adaptive or self-tuning cooling schedules capable of dynamically adjusting temperature according to optimization progress.

Such schedules may improve the balance between:

- exploration,
- exploitation,
- and computational efficiency.

12.4.2 Hybrid Optimization Methods

Combining SA with deterministic local optimization methods may significantly improve convergence performance.

Potential hybrid approaches include:

- SA + Gradient Descent,
- SA + 2-opt for TSP,
- SA + Genetic Algorithms,
- and memory-based metaheuristics.

12.4.3 High-Dimensional Optimization

Future work may extend the study toward higher-dimensional optimization problems, where convergence difficulties and computational complexity become substantially more challenging.

12.4.4 Parallel and Distributed SA

Parallel implementations of Simulated Annealing may reduce computational overhead and improve scalability for large-scale optimization problems.

12.4.5 Applications in Modern Machine Learning

Additional applications may include:

- neural network optimization,
- feature selection,
- image processing,
- reinforcement learning,
- and large-scale hyperparameter optimization.

12.5 References

References

Source Code Repository

https://github.com/Kanishk-Bhuradiya/ASM-Report-Work

[1] Bertsimas, D. and Tsitsiklis, J. (1993). *Simulated Annealing*. Statistical Science, Vol. 8, No. 1, 10–15.

[2] Bohachevsky, I., Johnson, M. E., and Stein, M. L. (1986). *Generalized simulated annealing for function optimization.* Technometrics, 28, 209–217.

[3] Černý, V. (1985). *A thermodynamic approach to the traveling salesman problem: An efficient simulation.* Journal of Optimization Theory and Applications, 45, 41–51.

[4] Hajek, B. (1988). *Cooling schedules for optimal annealing.* Mathematics of Operations Research, 13, 311–329.

[5] Johnson, D., Aragon, C., McGeoch, L., and Schevon, C. (1992). *Optimization by simulated annealing: An experimental evaluation, Part II: The traveling salesman problem.*

[6] Kirkpatrick, S., Gelatt, C. D., and Vecchi, M. P. (1983). *Optimization by Simulated Annealing.* Science, 220(4598), 671–680.

[7] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. (2019). *Optuna: A Next-generation Hyperparameter Optimization Framework.* Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining.

[8] Dorigo, M. and Gambardella, L. M. (1997). *Ant Colony System: A Cooperative Learning Approach to the Traveling Salesman Problem.* IEEE Transactions on Evolutionary Computation, 1(1), 53–66.