Markov Chain Monte Carlo and Simulated Annealing with Applications and Implementations

Hongda Li

December 9, 2022

Abstract

In this report, we prove the fundamentals for the convergence of the Metropolis Hasting Chain(MHC) under the discrete case; we will mention and list resources for the continuous case. We discuss the Simulated Annealing algorithm(SA) as a particular case of the Metropolis Hasting and use both algorithms for numerical experiments in Julia [1]. The first experiment is sampling from complicated distribution functions on 2D, the second is applying Simulated Annealing for the knapsack problem, and in the third experiment, we test simulated Annealing on the Rastrigin function using different base chains. We collect data and illustrate the behaviors of these algorithms. The code that produces the results for the numerical experiments is on my GitHub: here.

1 Introduction

Metropolis Hasting is a Markov Chain Monte Carlo (MCMC) method. It converges to a targeted distribution. For motivations, it is easy to sample from a 1D distribution function if we have the inverse of the CDF; however, it is generally tough to sample from a high-dimension distribution even if we have the PDF function.

The Metropolis Hasting Chain (MHC) is a Markov chain whose stationary distribution equals the targeted distribution function. This report is interested in the theoretical foundations for MHC and its applications, and we are also interested in understanding Simulated Annealing (SA) using stochastic processes.

Definition 1 (Stationary distributions)**.** Let $p(x, y)$ be a transition kernel for a Markov chain with state space S that is countable, then π is said to be a stationary distribution for p if it satisfies:

$$\pi(y) = \sum_{x \in S} p(x, y)\pi(x) \quad \forall y \in S.$$

Definition 2 (Detailed balance)**.** Let $p(x, y)$ be the transition kernel for a Markov chain with state space S which is countable, then the distribution π satisfies detailed balance if:

$$\pi(y)p(y, x) = \pi(x)p(x, y) \quad \forall x, y \in S.$$

Remark 1.0.1. If a Markov chain has a distribution that satisfies detailed balance, then the distribution will be the stationary distribution for the Markov chain.

Definition 3 (Support of a distribution)**.** Let f be a probability mass function with domain S, then the support of the PDF is defined as:

$$\text{supp}(f) := \{x \in S : f(x) > 0\}.$$

2 Preliminaries

Theorem 1 (Convergence to stationary distributions)**.** Let $(X_n)_{n\geq 0}$ be a discrete Markov chain with countable/finite state space S. Assuming it is irreducible, aperiodic with a stationary distribution π, then as $n \to \infty$, we have $p^n(x, y) = \pi(y)$; the stationary distribution is also unique.

Proof. The theorem is listed as theorem 1.19 in Rick's book [2]. □

Remark 2.0.1. This theorem plays a central role in understanding the regularity conditions for the convergence properties of MHC. Moreover, we skip the discussion regarding Markov chains with an uncountable state space. For more details about the ergodic theorem, see chapter 6 of the book by Robert and Casella [5].

2.1 Metropolis Hasting chain and its convergence

In this subsection, we present the proof and theoretical foundations for the convergence of the MHC when the underlying state space is countable. We consider the following questions:

1. What is the MHC?
2. For what regularity conditions can we assert that the stationary distribution is equal to the targeted distribution? If so, does it imply convergence to the distribution?
3. What we the regularity conditions imply about the choices of a base chain when it comes to actual applications?

We should apply the convergence theorem and analyze the MHC to answer the questions.

2.2 The algorithm

The quantities in algorithm 1 are listed below:

1. $q(x|y)$ is the base chain defined on S and it has to be doubly stochastic meaning that $q(x|y) = q(y|x)$.
2. $f(x) : S \mapsto \mathbb{R}_+$ is a probability mass function on the state space S.
3. ρ is the acceptance function, given $X^{(t)}$, it decides whether to accept $Y^{(t+1)}$ from q.
4. $X^{(t)}$ is a state given and it has to be the case that $f(X^{(t)}) > 0$.

Algorithm 1 Metropolis Chain

Input: $X^{(t)}$

$Y^{(t)} \sim q(\cdot|x^{(t)})$

$\rho(x,y) := \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\}$

$$X^{(t+1)} := \begin{cases} Y^{(t)} & \text{w.p} : \rho(X^{(t)}, Y^{(t)}) \\ X^{(t)} & \text{otherwise} \end{cases}$$

2.3 Transition kernel of MHC

Firstly, the transition kernel of the MHC is given by:

$$K(x,y) = \rho(x,y)q(y|x) + \left(1 - \underbrace{\sum_{z \in S\setminus\{y\}} \rho(x,z)q(z|x)}_{=:r(x)}\right)\mathbb{1}\{y = x\}.$$

The above is direct from algorithm 1.

Theorem 2 (Stationary distribution for the MHC)**.** The distribution f satisfies the detailed balance conditions when $x, y \in S$; consequently, the MHC has f as the stationary distribution.

Proof. Consider any $x, y \in \text{supp}(f)$ with $x \neq y$ then

$$\begin{aligned} \rho(x,y) &= \min\left\{\frac{f(y)}{f(x)}\frac{q(x|y)}{q(y|x)}, 1\right\} \\ (*) \implies \rho(x,y)f(x) &= \min\left\{f(y), f(x)\right\}, \rho(y,x)f(y) = \min\left\{f(x), f(y)\right\} \\ \implies \rho(x,y)f(x) &= \rho(y,x)f(y) \\ \implies \rho(x,y)q(y|x)f(x) &= \rho(y,x)q(x|y)f(y) \\ \implies K(x,y)f(x) &= K(y,x)f(y), \end{aligned}$$

when $x = y$, we have $1 - r(x) = 1 - r(y)$, which is just trivial. At $(*)$ we used the doubly stochastic property of the base chain and $f(x), f(y)$ are strictly positive by the definition of the support set of f. Finally, we can never travel to a state that is not in $\text{supp}(f)$ by the definition of ρ. □

However, the above theorem does not guarantee that the Metropolis chain will converge to the distribution f. To converge, it has to be the case that the chain is irreducible on all states in the support set of f (f-irreducible). To be irreducible on the set of $\text{supp}(f)$ means that for all $x, y \in \text{supp}(f), x \to y$, where $x \to y$ means that it is possible to communicate from the state x to y. In general, it is better to judge f-irreducible conditions on a case-by-case basis to determine the choice of the base chain. Nonetheless, the below theorem quantifies a stronger condition to assert the convergence.

Theorem 3 (Regularity conditions for the MHC)**.** For a base chain that is non-negative, e.g., $q(x|y) > 0 \; \forall x, y \in S \times S$, the MHC is f-irreducible.

Proof. For all $f(x), f(y) > 0$, we know $f(y)/f(x) > 0$, then we have $\rho(x, y)$ being non-negative and therefore, it is possible to jump between any state in the support set for f. □

Remark 2.3.1. Weaker conditions for the regularity conditions exist. For more detail, review Robert and Casella's book [5] in chapters 6 and 7, where it discusses the regularity conditions for the Markov chain with a continuous state space.

In practice, we have to assess the irreducibility on a case-by-case basis for the best choice for the base chain. In most cases, we have no idea what the set $\text{supp}(f)$ even is and how it would be described. And if that is the case, one will have to use the convex combinations of several base chains of different types together to make a base chain such that it satisfies f-irreducibility.

Theorem 4 (Convergence of the Metropolis Chain)**.** The MHC described in algorithm 1 that satisfies theorem 3 and has a countable state space will converge to the stationary distributions: $\text{supp}(f)$.

Proof. Observe that MHC in algorithm 1 is aperiodic because there exists $x, y \in \text{supp}(f)$ with $f(x) \neq f(y)$ by non-negativity property of q, it allows $\rho(x, y) < 1$ and giving us none zero probability for staying at state $X^{(t)}$. If this is not the case, it has to be $f(x) = f(y)$ for all $x, y \in S$, meaning that the uniform distribution is distribution f. In that case, it is the same stationary distribution as the doubly stochastic base chain q. Under both cases, we have some state that can loop back to itself, we also know that it's f-irreducible, therefore the chain is aperiodic

Because it satisfies theorem 3 and $\text{supp}(f)$ satisfies detailed balance (theorem 2), using theorem 1 MHC will converge to $\text{supp}(f)$. □

2.4 Numerical experiments

In this experiment, we consider sampling from the following functions that we made:

$$\begin{aligned} D &:= \{(x_1, x_2) : -\sin(4\pi x_1) + 2\sin(2\pi x_2)^2 > 1.5\} \\ f(x) &:= \mathbb{1}_D(\sin(x_1 4\pi) + \cos(x_2 4\pi) + 2), \end{aligned}$$

observe that this is a distribution function up to a constant.

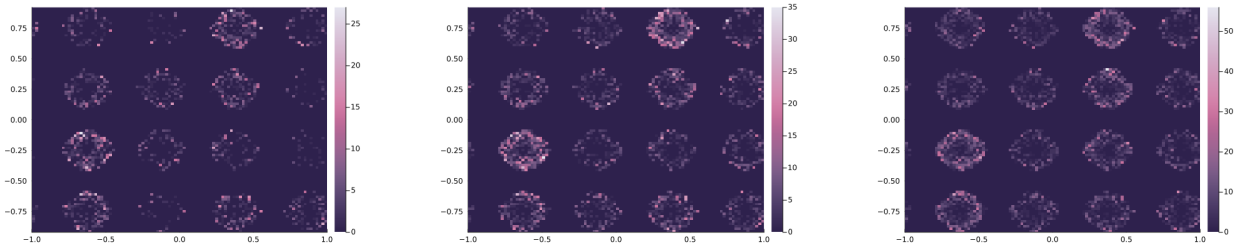
For our case, the function f has a disconnected support set, and the sample space is continuous. Please take for granted that the convergence theorem we proved is applicable for Markov chains with a bounded continuous state space. For our experiment, we will be sampling from f using this list of base chains:

1. A wrapped Gaussian random walks in $[0, 1] \times [0, 1]$ centered at the previous state, or in other words, a Gaussian random walks with periodic boundary conditions on the domain: $[0, 1] \times [0, 1]$ and the standard deviation for the Gaussian distribution is 0.1.

2. A uniform distribution in $[0, 1] \times [0, 1]$ is disregarding the previous state. This is a trivial Markov chain.

We present the results for both choices of base chain in figure 1 and 2. We take snapshots of all the existing sampled points every 5000 times. Going from left to right in both cases, we have the approximated distributions when there are 5000, 10000, and 15000 sampled points. These two chains converge very differently for this example, which makes sense. For a uniform random sampler, the previous point does not correlate with the next point from q, making it possible to jump between the discrete sets in $\text{supp}(f)$ and sample from them evenly. On the contrary, the Gaussian random walks samplers have difficulties sampling evenly from each of the discrete sets in $\text{supp}(f)$. The argument makes intuitive sense because a wrapped Gaussian random walk is more likely to sample from the "island" close to the one where the current state resides, and it needs more attempts to jump between the islands. This experiment focuses on non-negative double stochastic chain chan q; however, it also tells us that it won't be able to sample from all the disconnected components of $\text{supp}(f)$ when the probability of traversing between states in the support set of f when the base chain does not allow for MHC to be f-irreducible.

Finally, the reader should notice that a Monte Carlo method, without the Markov chain, is just sampling using a uniform base chain where the function f is an indicator function for some sets. Moreover, this observation should reveal the intuitions that MCMC is a type of MC that can sample locally.



(a) Firt snapshot. (b) The second snapshot. (c) The third snapshot.

Figure 1: Snapshots of accumulated samples when a wrapped Gaussian random walk base chain.

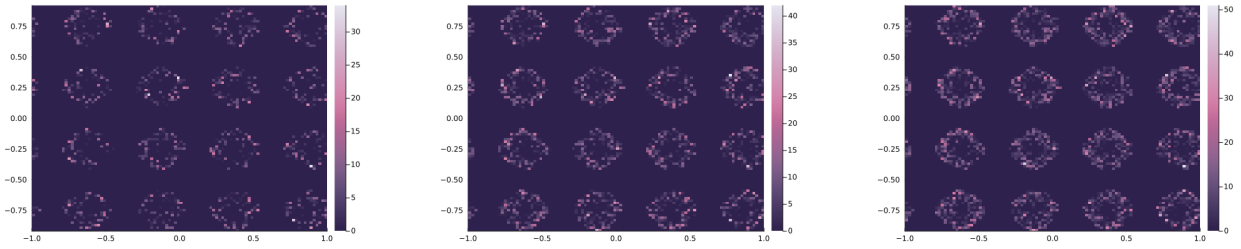


Figure 2: Snapshot of samples for a uniform random base chain.

As a remark, choosing a base chain is much more complicated in higher dimensions, and using MCMC has overwhelming advantages. For example, as the dimension increases, a unit sphere's volume over the unit hypercube approaches zero. As a result, it poses challenges for sampling from the hypersphere using Monte Carlo due to a high rejection rate. However,

if we have an initial point in the hypersphere and a carefully designed base chain, we can sample from a high-dimensional sphere using MCMC much more efficiently.

3 Simulated annealing

Simulated Annealing is an optimization method, and it is also a specific case of the MHC. Firstly, observe that if a distribution f with a bounded domain and a set X^* in the domain that denotes the maximizers and X^* that makes a function f close to its maximum, then the probability of sampling from X^* is higher than other sets with the same measure. This inspiration here allows us to design and make an optimization method. More specifically, given any function $g : S \mapsto \mathbb{R} \cup \{-\infty\}$, we have $f(x) := \exp(g(x))$ such that f is a distribution up to a positive constant multiplier. By the monotone property of $\exp(\cdot)$, f and g share the same maximizers.

Compared to the MHC, which is about sampling from a distribution, the Simulated Annealing method aims to look for the maximum of a given function and comes with a slight modification called the temperature. The temperature term is called: T_i where i is the current step for the drawn samples from the MHC, then we have the distribution function to be given as $f(x) := \exp(g(x)/T_i)$. The goal of the temperature is to accentuate the set X^*. The smaller the value T_i, the larger the probability it is to sample from the set X^* compare to other sets.

For readers who are optimizations enthusiasts, this is invigorating news. One can see that this method makes few assumptions about the objective function, and there is also a convergence proof for the method [3]. However, we will show later with the knapsack problem that its performance depends on the problem structure, the base chain, and the temperature schedule. Adding salt to injury, it only converges in distributions. Nonetheless, one can not deny its universality, the fact that it is simple to implement, and it is trivial to parallelize in modern computing platforms.

3.1 The limit of temperature

Theorem 5 (The limit of temperature)**.** Suppose that state space S is a finite set, and it is the state space for an MHC equipped with $f(x) = \exp(g(x)/T_i)$ where $g : S \mapsto \mathbb{R} \cup \{-\infty\}$, then as $T_i \to 0$, f approaches $\mathbb{1}_{X^*}$.

Proof.

$$\begin{aligned}
\lim_{i\to\infty} f(x) &= \lim_{i\to\infty} \frac{\exp(g(x)/T_i)}{\sum_{y\in S}\exp(g(y)/T_i)} \\
&= \lim_{i\to\infty} \frac{1}{\sum_{y\in S}\exp(g(y)-g(x)/T_i)} \\
&= \lim_{i\to\infty} \frac{1}{\exp(0)+\sum_{y\in S\setminus\{x\}}\exp(g(y)-g(x)/T_i)} \\
&= \lim_{T\to 0} \frac{1}{1+\sum_{y\in S\setminus\{x\}}\exp(g(y)-g(x)/T)} \\
&= \mathbb{1}\{X^+\}.
\end{aligned}$$

Take note that the assumption I made here is stronger than it needs to be; it might still converge for MHC with a continuous distribution. For more detail about the limiting behavior of the distribution function as $T \to 0$, see chapter 3 of the book by Levin et al. [4] for more details. □

To illustrate the process of changing temperature, we make an example involving $g(x) = \operatorname{sinc}(x)$ and choose temperatures from 1 to 10^{-2} geometrically distributed. Then we plot the normalized function $f(x) := \exp(\operatorname{sinc}(x)/T_i)$ on the interval $[-3, 3]$ and it produces figure 3. Observe that the function's maximum accentuates as the temperature gets lower, which means that the MHC is more likely to reject states with lower values and climb more aggressively than when the temperature is still high.

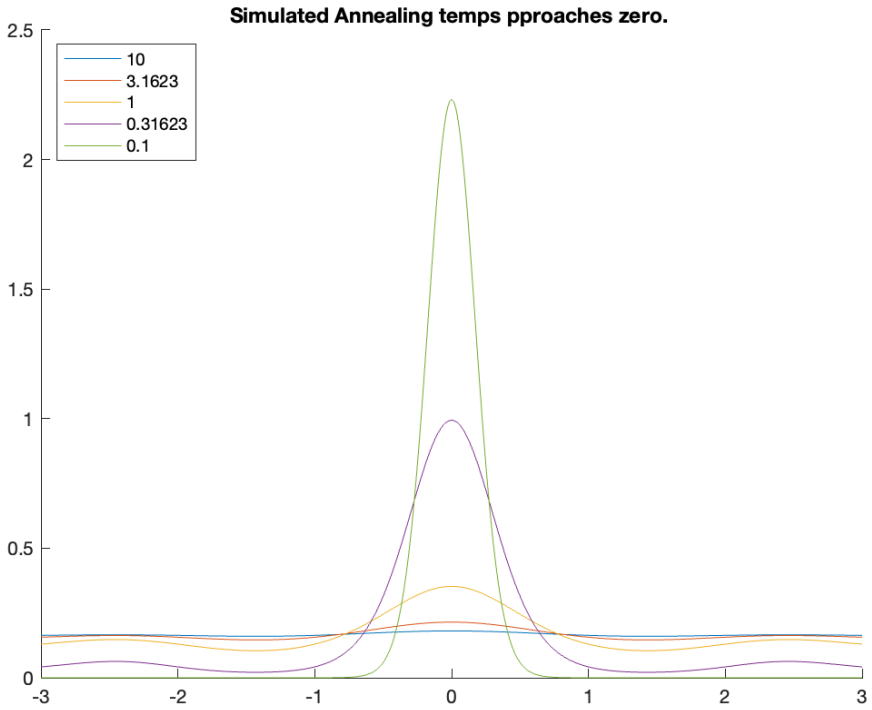


Figure 3: The distribution function of $\text{sinc}(x)$ as the temperature for simulated annealing gets smaller and smaller.

In practice, people use a function that controls the temperature for the SA to converge as the temperature decreases which avoids exploring the whole landscape and wasting all the computing resources. Furthermore, to avoid the limitations of floating points, we consider a customized acceptance function for SA:

$$\rho(x, y) := \min\left\{\mathbb{1}\{f(y) < f(x)\}\exp(f(y) - f(x)) + \mathbb{1}\{f(y) \geq f(x)\}, 1\right\}.$$

The above function eliminates floating point overflow because taking the exponential of a large number can cause overflow in the computer.

3.2 Solving the knapsack problem

The knapsack problem is an integer programming problem:

$$\max\langle w, x\rangle \text{ s.t: } \langle w, x\rangle \leq 1, x \in \{1, 0\}^n.$$

We look for a subset of numbers from the vector w such that it sums up as close to one as possible while not exceeding one. We can solve this problem with SA, and the state space of the MHC will be $S := \{0, 1\}^n$, a binary vector with length n with the objective function

$g(x)$ defined as:

$$g(x) := \begin{cases} \langle w, x \rangle & \text{if } \langle w, x \rangle \leq 1, \\ -\infty & \text{else.} \end{cases}$$
$$f(x) := \exp(g(x)/T_i),$$

where f will be our target distribution.

3.3 Best base chain that avoids the curse of dimensionality

Consider a case where all pairwise sum of elements in x exceeds 1; then it is implied that all feasible solutions x have to have less than one 1 in them. Suppose one chose a uniform random base chain, let $\epsilon_i \sim \text{Bernoulli}(1/2)$ $\forall 1 \leq i \leq n$, $\epsilon \in \{0, 1\}$ and suppose that the $[\![\cdot]\!]$ denotes the equivalent classes under modulo of two. It applies to each element of a vector. Let e_i denote the ith standard basis vector. The doubly stochastic base chain $q(x|y)$ would be given by:

$$q(Y|x) = \mathbb{P}\left(Y = \sum_{i=1}^{n} [\![x + \epsilon]\!] e_i\right)$$

and this is equivalent to choosing each of the bits in the vector x and flipping it. Assuming that we start with a feasible solution $x = \mathbf{0}$ and only solutions x such that it has less than one 1 in it are feasible. As a result, the probability of transition from $\mathbf{0}$ to a feasible solution is $\mathbb{P}(\text{binomial}(1/2, n) \leq 1) = (n+1)(1/2)^n$, because we would need to either remain at state $x = \mathbf{0}$ or change exactly one of the bits to keep the solution feasible. The Probability of obtaining the first feasible solution approaches zero as n increases. Therefore, this choice of the base chain is bad if the set of feasible solutions is small relative to the whole space $\{0, 1\}^n$. To reduce the variance, we propose a random variable η that uniformly distributes in $\{1, \cdots, n\}$, then the chain:

$$q(Y|x) := \mathbb{P}(Y = [\![x + e_\eta]\!]),$$

would be a better choice. It is also f-irreducible because we can mutate every "1" bit in x to get to $\mathbf{0}$, then there is a probability to get from $\mathbf{0}$ to any x.

3.4 Numerical Experiments

Consider a list of i.i.d random variables uniformly sampled from $[0, 1]$, denoted using W_i. For our example, there are 99 of them. Then the sum:

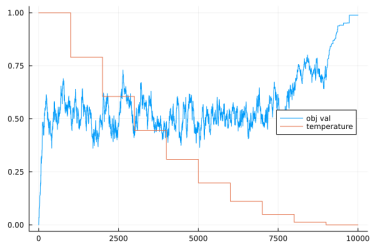
$$W_1 + \left(\sum_{i=1}^{99} w_{i+1} - W_i\right) + 1 - W_{99} = 1,$$

therefore, we make the first element of $w_1 = W_1$ and the last element $w_{100} = 1 - W_{99}$ and $w_i = W_{i+1} - W_i$ for all $1 \leq i \leq 99$. The first 100 elements of w sum up to an optimal solution. Next, we let w_i for all $200 > i > 100$ to be an i.i.d random variable drawn from

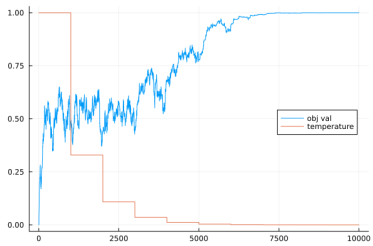
the uniform distribution on the interval $[1, 2]$. This means that any of the w_i for $101 \leq 200$ cannot be part of the solution because adding any of them more than once will make the solution infeasible (This also means that all feasible solutions are not a local maximum.). This 200 elements vector for the knapsack problem is our numerical experiment. In addition, we also consider three different temperature schedules for the SA experiment.

1. The temperature drops linearly from 1 to 0.001 every 1000 iterations, 10k iterations in total.
2. The temperature is $exp(-k)$ where k goes from 0 to 10. It decreases every 1000 iterations in a total of 10k iterations.
3. The temperature is $1/k^2$; it drops whenever the algorithm discovers a solution lower than all previous solutions for the objective function.

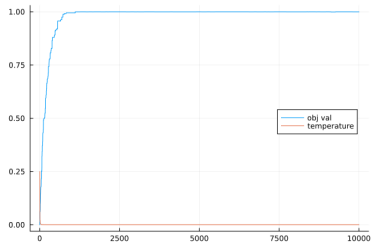
figure 4 shows the results. All the MHC can converge to the unique maximum. Moreover, observe that when the temperature is automatically scheduled, it converges to the optimal fastest. However, we would like to point out that there is no good schedule for every problem. In our experiment, all feasible solutions can be changed into optimal ones without decreasing the objective function's value. Suppose that x^+ is the optimal and $\bar{x}$ is a feasible solution, then mutating any of the first 100 bits of $\bar{x}$ keeps its feasibility and increases its objective value. However, for the knapsack problem, this is not true in general; the algorithm could be stuck in a local optimal instead.



(a) Temperature with a linear schedule.



(b) Temperature with an exponential schedule.



(c) When a better solution is identified, the temperature decreases.

Figure 4: The objective values for 10k iterations with different types of temperature schedules.

There are ways to combat the problem, and I would like to point out that if $q_1, 1_2$ are the kernels of two doubly stochastic chains, then their weighted sum $\alpha_1 q_1 + \alpha_2 q_2$ where $1 = \alpha_1 + \alpha_2$ is also doubly stochastic. This can be exploited to make a base chain that is robust to the curse of dimensionality while, at the same time, able to do short random walks locally.

References

[1] Jeff Bezanson, Stefan Karpinski, Viral B Shah, and Alan Edelman. Julia: A fast dynamic language for technical computing. *arXiv preprint arXiv:1209.5145*, 2012.

[2] R. Durrett. *Essentials of Stochastic Processes*. Springer Texts in Statistics. Springer New York, 2012.

[3] V. Granville, M. Krivanek, and J.-P. Rasson. Simulated annealing: A proof of convergence. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 16(6):652–656, 1994.

[4] David A. Levin, Yuval Peres, and Elizabeth L. Wilmer. *Markov chains and mixing times*. American Mathematical Society, 2006.

[5] Christian P. Robert and George Casella. *Monte Carlo Statistical Methods*. Springer, 2005.