NATIONAL UNIVERSITY OF SINGAPORE

A STUDY OF GO-POLARS: THEORY AND PRACTICE

by

GAO ZHENG

In partial fulfillment of the requirements for the Degree of
Bachelor of Engineering

in the

Faculty of Engineering
Department of Industrial and Systems Engineering

Dec 2014

NATIONAL UNIVERSITY OF SINGAPORE

Abstract

Faculty of Engineering
Department of Industrial and Systems Engineering

Bachelor of Engineering

by GAO ZHENG

Gradient Oriented Polar Random Search (GO-POLARS) is a recently developed stochastic search method which incorporates gradient information. We introduce an acceptance/rejection rule to the GO-POLARS algorithm in hope to achieve better performance in terms of global convergence.

Some theoretical results that justifies this modified algorithm in practice are established. We show under suitable conditions the existence of Cooling Schedules which guarantee convergence in probability to the global optimum.

We also includes numerical illustrations of the convergence process, and numerical experiments that evaluate the performance of the algorithm for some standard test functions. Convergence of quasi-stationary distribution to the optimum is observed in numerical examples. Finite-time behavior of the algorithm is found to be sensitive to initial conditions. Comparisons are drawn with standard Simulated Annealing algorithm as benchmark. Examples also illustrate situations where the modified algorithm are superior or inferior to the benchmark.

Acknowledgements

I would like to express my very great appreciation to Prof. Lee Loo Hay and Prof Chew Ek Peng for their patience and guidance over the course of this project, and for their tolerance towards my ignorance and temper. Their advice and suggestions has been immensely helpful.

I would also like to offer my thanks to Dr. Li Haobin for his wonderful advice on the numerical analysis part of the project.

Special thanks are extended to the staff of the Department of Industrial and Systems Engineering for their administrative support.

I am particularly grateful for the support from my friends and family.

Contents

List of Figures

Chapter 1

Introduction

Two major classes of algorithms have been developed to solve optimization problems. The first class of algorithms, based on local behaviour of the objective function, seeks to maximize (minimize) the objective value in an iterative process. Prototypical examples include Steepest Descent (sometimes called Gradient Descent method) and Newton's Method. These algorithms explore a space of admissible configurations, S, in a deterministic fashion. Often the search terminates at a local maximum (minimum) due to the 'greedy' nature of the algorithms.

To avoid being trapped in local optima, a class of randomized algorithms which generate the next configuration randomly have been devised, and are able to 'climb hills', i.e., moves that generate configurations of lower (higher) than the present one are accepted. The second class consists of stochastic search methods and other randomized heuristic algorithms including Metropolis-Hastings algorithm, simulated annealing, and Tabu search.

The effectiveness of the first type search methods, such as gradient-based search, in finding global optimal requires improvement; while the efficiency of the second type algorithms also invites new developments as the time taken for obtaining the solutions is sometimes prohibitive. It is therefore of interest to develop search algorithms to find global optimum with a high probability at a fast rate.

Gradient Oriented Polar Random Search (GO-POLARS) is a recently developed stochastic search method which incorporates gradient information.[1] Randomness is injected under a hyper-spherical coordinates system. Numerical experiments have been designed and conducted to verify the effectiveness proposed strategies in modifying the level of exploration and exploitation dynamically. Steepest Descent (SD) search algorithm was

used as a benchmark for performance analysis. A predetermined amount of computational budget is applied as a control of iterations. Performance based on the given budget has been analyzed.

1.1 In pursuit of convergence

Convergence properties is less studied for the GO-POLARS and it is of interest to explore the convergence properties of the algorithm.

It is established that Simulated Annealing with a proper Cooling Schedule lead to the global optimal while the Steepest Descent to the local optimal. Before we delve into the study of convergence properties we compare the GO-POLARS to the Steepest Descent and Simulated Annealing method here and briefly discuss their similarities and differences.

Resemblance of GO-POLARS and Steepest Descent can be seen from their use of gradient information in generating the next candidate. The direction of search are oriented in the most 'greedy' way locally. On the other hand, GO-POLARS randomizes the search direction and creates a neighbourhood structure that is not present in the Steepest Descent approach.

Similarities of GO-POLARS with Simulated Annealing is also evident in the description of the algorithm. Generation of the next candidate is specialized to a particular form which utilized the gradient information and a polar orientation. Acceptance probability is 1 whenever the proposed new solution surpasses the old one; and is 0 whenever the the proposed new solution is inferior. The GO-POLARS can be in fact identified as a limiting case of the Simulated Annealing with temperature fixed at $T = 0$ (and therefore a dichotomous acceptance/rejection rule), with a specialized neighbourhood structure for proposal of new solutions. However it is of interest to know whether a globally optimal solution can be found with a guaranteed high probability due to its somewhat 'greedy' nature.

1.2 Contents of this thesis

We will explore a modification that brings about better performance in terms of global convergence.

In Chapter 2 we will survey and describe a number of related algorithms. Important existing results on convergence properties for the algorithms will be discussed.

We formally introduce GO-POLARS and the modification in Chapter 3. We introduce the concept of temperature from Simulated Annealing to the GO-POLARS algorithm, and show that under suitable conditions the existence of a Cooling Schedule that leads to convergence towards the global optimal solution.

Chapter 4 provides some Numerical examples to illustrate the convergence of quasi-stationary distribution, and to demonstrate finite time behaviors of the modified algorithm. We compare the performance against standard Simulated Annealing as a benchmark, comment on the findings and discuss the characteristics of GO-POLARS with Cooling.

Summary of the project is included in Chapter 5. We point out some limitations and possible directions for further investigation.

Chapter 2

Survey of Literature

We survey the three algorithms that are intimately related to the modified GO-POLARS algorithm: Steepest Descent algorithm, Metropolis-Hastings algorithm, Simulated Annealing algorithm. Convergence results of the Simulated Annealing algorithm will be summarized.

2.1 Survey of Methods

Three optimization methods are surveyed: Steepest Descent algorithm, Metropolis-Hastings algorithm, Simulated Annealing algorithm; the first gradient-based, Metropolis-Hastings and Simulated Annealing stochastic search methods. The connections of GO-POLARS algorithm with the gradient-based methods and the stochastic search methods will be briefly discussed when we formally introduce GO-POLARS in the next Chapter.

2.1.1 Steepest Descent

Gradient descent is based on the observation that if the multivariable function $F(\mathbf{x})$ is defined and differentiable in a neighborhood of a point $\mathbf{a}$, then $F(\mathbf{x})$ decreases fastest if one goes from $\mathbf{a}$ in the direction of the negative gradient of F at $\mathbf{a}$, $-\nabla F(\mathbf{a})$. It follows that, if

$$\mathbf{b} = \mathbf{a} - \gamma \nabla F(\mathbf{a})$$

for γ small enough, $F(\mathbf{a}) \geq F(\mathbf{b})$. With this observation in mind, one starts with a guess $\mathbf{x}_0$for a local minimum of F, and considers the sequence $\mathbf{x}_0, \mathbf{x}_1, \mathbf{x}_2, \ldots$ such that

$$\mathbf{x}_{n+1} = \mathbf{x}_n - \gamma_n \nabla F(\mathbf{x}_n),\ n \geq 0$$

We have

$$F(\mathbf{x}_0) \geq F(\mathbf{x}_1) \geq F(\mathbf{x}_2) \geq \cdots$$

so hopefully the sequence $(\mathbf{x}_n)$ converges to the desired local minimum. Note that the value of the step size γ is allowed to change at every iteration. With certain assumptions on the function F (for example, F convex and ∇F Lipschitz) and particular choices of γ, convergence to a local minimum can be guaranteed. When the function F is convex, all local minima are also global minima, so in this case gradient descent can converge to the global solution.

A proof of convergence in the case of quasi-convex objective objective functions and a step size schedule can be found in Kiwiel and Murty [2].

2.1.2 Metropolis-Hastings

The Metropolis-Hastings algorithm draws unweighted samples from probability distribution $f(x)$, if the density f can be identified up to a normalizing constant as $\pi(x)$. The Metropolis-Hastings algorithm constructs a Markov chain that has the desired distribution $f(x)$ as the stationary distribution. The transition probability is determined by a proposal distribution $q(\cdot|\cdot)$ depending possibly on the current sample point, and an acceptance rule specified in the following description of the algorithm.

1. Choose an starting candidate x_0 for the Markov chain in the sample space.
2. Generate y from the proposal distribution $q(\cdot|x_i)$ and u from uniform distribution on $[0, 1]$.
3. Set $x_{i+1} = y$ if $u < \min\{\frac{\pi(y)q(y|x_i)}{\pi(x_i)q(x_i|y)}, 1\}$, otherwise set $x_{i+1} = x_i$.
4. Increment i. Go to step 2.

An intuitive interpretation of the algorithm is that the acceptance-rejection process adjust for the number of times the Markov chain visits the points in the sample space by weighting the acceptance probability.

The requirement that $f(x)$ is only identified up to a normalizing constant rather than its normalized form makes the Metropolis–Hastings algorithm particularly useful, because calculating the necessary normalization factor is often difficult in practice.

A good introductory text on Markov Chain Monte Carlo and its applications in Bayesian inference can be found in [3]. Proofs for convergence to the target distribution when sample space is discrete can also be found in the reference.

The algorithm, although not tailored for optimization problems, has its later reincarnation with a slight modification as a stochastic search algorithm which we introduce next.

2.1.3 Simulated Annealing

Similar to that of the Metropolis-Hastings algorithm, the Simulated Annealing method considers some neighbouring state y of the current state x_i, and probabilistically decides between moving the system to state y or staying in state x_i. These probabilities ultimately lead the system to move to states of lower objective value. Typically this step is repeated until the system reaches a state that is good enough for the application, or until a given computation budget has been exhausted.

The probability of making the transition from the current state s to a candidate new state y is specified by an acceptance probability function $P(f(x_i), f(y), T)$, that depends on the objective values at y and x_i, and on a global time-varying parameter T called the temperature. The probability function P must be positive even when $f(y)$ is greater than $f(x_i)$. This feature prevents the method from becoming stuck at a local minimum that is worse than the global one.

When T tends to zero, the probability $P(f(x_i), f(y), T)$ must tend to zero if $f(y)$ is greater than $f(x_i)$ and to a positive value otherwise. For sufficiently small values of T, the system will then increasingly favor moves that go "downhill" (i.e., to lower objective values), and avoid those that go "uphill." With $T = 0$ the procedure reduces to the 'greedy' algorithm, which makes only the downhill transitions.

The P function is chosen so that the probability of accepting a move decreases when the difference $f(y) - f(x_i)$ increases—that is, small uphill moves are more likely than large ones. We will restrict our attention to version of the algorithm when the conditions are satisfied.

Given these properties, the temperature T plays a crucial role in controlling the evolution of the state of the system with regard to its sensitivity to the variations of objective values. To be precise, for a large T, the evolution of x_i is sensitive to coarser objective value variations, while it is sensitive to finer objective value variations when T is small.

Basic steps in the algorithm are

1. Choose an starting candidate x_0 for the Markov chain in the solution space.

2. Generate y from the proposal distribution $q(\cdot|x_i, T)$ and u from uniform distribution on $[0, 1]$.

3. Set $x_{i+1} = y$ if $u < \min\{\exp\{-\frac{f(y)-f(x_i)}{T_i}\}, 1\}$, otherwise set $x_{i+1} = x_i$.

4. Increment i; update temperature T_{i+1}. Go to step 2.

Conditions of convergence to the global optimal and its proof can be found in Mitra, Romeo and Sangiovanni-Vincentelli [4]. A study of its finite time behavior is also found in the reference.

2.2 Convergence of Simulated Annealing

There are a number of proofs of convergence in the literature when the Cooling Schedule follows a particular parametric form, and under the assumption that a symmetry condition is imposed on the proposal matrix such that the strong reversibility condition is satisfied (to be defined the next chapter). A comprehensive review can be found in D. Henderson's survey and discussion [5].

Chapter 3

GO-POLARS with Cooling and Convergence Properties

We first formally introduce GO-POLARS in its original set-up, and comment on some finite-time behaviors of the algorithm studied in [1] and [6].

We then introduce the concept of a Cooling Schedule similar to that in the Simulated Annealing to the GO-POLARS. The idea is that instead of having a dichotomous acceptance/rejection rule, the acceptance and rejection probability depends on the differential of the objective values as well as the number of iterations run. The modified algorithm moves towards a 'greedier' nature as in the original set-up, but attempts to explore a larger region in the initial search.

A major difference between the modified GO-POLARS with standard Simulated Annealing is that the former does not satisfy the strong reversibility condition. Heuristically speaking, the former's proposal of candidate solution at each iteration may depend on the gradient information as well as the iterations run, further, a quasi-stationary distribution that is independent of the proposal distribution may not exist; while the latter assumes a structure of 'symmetry' in its proposal of candidate solutions.

Nevertheless, proof of Simulated Annealing serves a good starting point of our result. We will state some existing results on global convergence of Simulated Annealing in section 2.3 and point out where the conditions fail for the case of GO-POLARS with cooling. The fix will be provided in section 2.4, where we offer a proof of existence of Cooling Schedules that leads to convergence.

3.1 GO-POLARS

We formally describe the algorithm here.

Gradient-oriented Polar Random Search is a recently proposed stochastic search by Li. et al. [1]

Basic steps in the algorithm are

1. Choose an starting candidate x_0 in the solution space.
2. Obtain gradient information at x_i. If the gradient is zero, terminate the algorithm, otherwise generate a randomized directional vector $\mathbf{d_i}$ in polar coordinates. $||\mathbf{d_i}|| = 1$.
3. Generate y by moving a step size γ_i from x_i in the direction of $\mathbf{d_i}$
   i.e. $y = x_i + \gamma_i \mathbf{d_i}$
4. Set $x_{i+1} = y$ if $f(y) < f(x_i)$, otherwise set $x_{i+1} = x_i$.
5. Increment i. Go to step 2.

Two features are of importance in the algorithm. The amount of randomness injected in the generation of the directional vector and the step sizes jointly determines the extend to which the neighbourhood structure is explored as well as the extend to which the local gradient information is exploited.

It was shown in the numerical experimentation that exploitation plays a significant role in improve search efficiency for local optimization problem, while exploration is essential in increasing global convergence rate[1]. The results for local optimization problem are comparable to that of the SD approach, whereas the global convergence rate surpasses that of the SD approach [6].

3.2 Set-up of GO-POLARS with Cooling Schedule

We consider the basic elements of the algorithm and establish notations here:

1. A finite set of candidate solutions S.
2. A real-valued cost function c refined on S. Let $S^* \subset S$ be the set where the global minima of cost function c is attained.

3. An $n \times n$ stochastic matrix $\mathbf{A} = (a_{ij})$, i.e., $a_{ij} \geq 0$ and

$$\sum_{j=1}^{n} a_{ij} = 1 \quad \forall i \in S.$$

   which will be called the proposal matrix.

4. We associate the directed *neighbourhood graph* $G = G(\mathbf{A})$, whose vertices are the elements of S and whose edges are those pairs (i, j) of vertices satisfying $i = j$ or $a_{ij} > 0$. We assume that G is strongly connected, or equivalently, that the homogeneous Markov process is irreducible.

5. For every $i \in S$ we call $N(i) := \{j \in S | a_{ij} > 0\}$ the neighbours of i.

6. A non-decreasing real-valued function $T : \mathbb{N} \to (0, \infty)$ which will be called the **cooling schedule**.

7. An acceptance probability $f_T(c_i, c_j)$.

Basic steps in the algorithm are

1. Choose an starting point $i \in S$ as a candidate in the solution space.

2. Generate j from the proposal distribution $a(\cdot|i)$ and u from uniform distribution over $[0, 1]$.

3. Accept j and set $x_t = j$ if $u < f_T(c_i, c_j)$; otherwise set $x_t = i$.

4. Increment t. Go to step 2.

We concern ourselves with only the acceptance probability that satisfies the following multiplicativity condition:

$$f_T(c_i, c_k) = f_T(c_i, c_j) f_T(c_j, c_k) \quad \forall c_i \leq c_j \leq c_k$$

This condition is satisfied by most commonly used acceptance functions, for example $f_T(c_i, c_j) = \exp\{\min\{-\frac{c_j - c_i}{T}\}, 0\}$. We restrict our discussion to this particular form of acceptance function.

Viewed as a time inhomogeneous Markov chain, The transition stochastic matrix P_t is:

$$p_{ij}(t) = \begin{cases} a_{ij} f_{T_t}(c_i, c_j) & \text{if } i \neq j \\ 1 - \sum_{m \neq i} p_{im}(t) & \text{if } i = j \end{cases}$$

Question arises as to whether the algorithm leads the sequence $(x_t)_{t=1}^{\infty}$ to the optimal solutions S^*, and if so, under what conditions imposed on the acceptance probability f, the proposal matrix $\mathbf{A}$, and the cooling schedules T_m does the algorithm work.

3.3 Convergence Results for a Special Class of Proposal Distribution

We first restrict ourselves to Simulated Annealing, and present the main results for convergence.

There are a number of proofs of convergence in the literature under the assumption that a symmetry condition is imposed on the proposal matrix $\mathbf{A}$ such that the strong reversibility condition is satisfied, i.e., there exists a probability mass function g over S such that $g_i a_{ij} = g_j a_{ji}$. A more comprehensive review can be found in D. Henderson's survey and discussion [5]. We simply state the important theorems that motivates the arguments in our discussion.

3.3.1 Quasi-stationary Distributions and their properties

Quasi-stationary distribution at time t is the stationary distribution associated with the transition matrix at a time t (A chain which stays at the temperature T_t for a infinitely long time converges to the quasi-stationary distribution). It is clear that in practice this distribution is a construct rather than an observable since the chain cannot not stay at a fixed temperature for an infinitely amount of time. However they is important for the purpose of proof of convergence. We will see that convergence of quasi-stationary distribution to the optimal is a prerequisite for the chain to converge to the optimal.

Definition 3.1. The quasi-stationary distribution $\pi(t)$ of the time-inhomogeneous Markov chain at time t is defined as

$$\pi_i(t) := \frac{g_i \exp(-c_i/T_t)}{H(t)}$$

where $H(t)$ is the normalizing constant such that $\|\pi(t)\|_1 = 1$.

Proposition 3.2. *The quasi-stationary distribution $\pi(t)$ of the time-inhomogeneous Markov chain at time t is the stationary distribution associated with the transition matrix $P(t)$, i.e., $\pi(t)P(t) = \pi(t)$, $t = 0, 1, ...$*

Proof. Easily verified via the detailed balance condition. □

The next result states that the quasi-stationary distributions concentrate on the optimal configurations as temperature decreases to 0.

Proposition 3.3. *If the Cooling Schedule is a non-increasing function and $T_t \to 0$ as $t \to \infty$, the quasi-stationary probability vector defined in definition 2.1 converges to the optimal vector* $\mathbf{e}^*$

$$\mathbf{e}_i^* = \begin{cases} g_i / \sum_{j \in S^*} g_j & \text{if } i \in S^* \\ 0 & \text{if } i \notin S^* \end{cases}$$

Proof. See [4]. □

This proposition states that, in very imprecise languages, if the Cooling Schedule is such that the temperature decreases so slow the chain remains at any temperature for an infinite number of steps and the stationary probability is reached, the chain will eventually lead to the optimal solutions. Of course such a Cooling Schedule does not exist. It is natural, then, to ask whether there exists a Cooling Schedule for which the chain does converge. And if it does, will the limiting distribution be e^* as defined above.

The answer to the two questions is affirmative.

We first state the proposition that will lead to the result, and review two theorems in the theory of inhomogeneous Markov chain.

Proposition 3.4. *Monotonicity of Quasi-stationary Distributions*
For $i \in S^*$, $\pi_i(T_{t+1}) - \pi_i(T_t) > 0 \quad \forall m \geq 0$.
For $i \notin S^*$, there exists an unique integer $\hat{m}_i$, such that

$$\pi_i(T_{t+1}) - \pi_i(T_t) > 0 \quad \text{for} \quad 0 \leq m \leq \hat{m}_i - 1$$

and

$$\pi_i(T_{t+1}) - \pi_i(T_t) > 0 \quad \text{for} \quad m \geq \hat{m}_i.$$

Proof: See [4].

3.3.2 Time-inhomogeneous Markov Chains

We establish more notations here. Let $P(t_1, t_1)$ be the identity matrix, and

$$P(t_1, t_2) := \prod_{t=t_1}^{t_2-1} P(t) \quad 0 < t_1 < t_2$$

be the n-step transition matrix. Let $v(t)$ denote the probability vector after t transitions so that

$$v(t_2) = v(t_1)P(t_1, t_2)$$

also denote

$$v(t_1, t_2) = v(0)P(t_1, t_2).$$

Definition 3.5. A time-inhomogeneous Markov chain is *weakly ergodic* if for all t_1,

$$\lim_{t_2 \to \infty} \sup_{v^1(0), v^2(0)} \|v^1(t_1, t_2) - v^2(t_1, t_2)\| = 0 \tag{3.1}$$

where $v^1(0)$ and $v^2(0)$ are two arbitrary state probability vectors and

$$v^1(t_1, t_2) = v(0)^1 P(t_1, t_2)$$
$$v^2(t_1, t_2) = v(0)^2 P(t_1, t_2).$$

It should be noted that the definition of weak ergodicity indicates only a tendency for the rows of $P(t_1, t_2)$ to be equal, that is similar to a "loss of memory" property. It does not imply convergence. The necessary and sufficient condition for a time-inhomogeneous Markov chain to be weakly ergodic involves the definition of coefficient of ergodicity.

Definition 3.6. The *coefficient of ergodicity* of a stichatic matrix P is defined as

$$\tau_1(P) := \frac{1}{2} \max_{i,j} \sum_{k \in S} |P_{ik} - P_{jk}| = 1 - \min_{i,j} \sum_{k \in S} \min(P_{ik}, P_{jk}) \tag{3.2}$$

We are now ready for the theorems on convergence of time-inhomogeneous Markov chains.

Theorem 3.7. *The time-inhomogeneous Markov chain is weakly ergodic if and only if there is a strictly increasing sequence of positive integers* $(t_i)_{i=0}^{\infty}$ *such that*

$$\sum_{i=0}^{\infty} [1 - \tau_1(P(t_i, t_{i+1}))] = \infty. \tag{3.3}$$

Proof. See [7] or [8]. □

In contrast, the *strong ergodicity* is defined as follows.

Definition 3.8. A time-inhomogeneous Markov chain is *strongly ergodic* if there exists a probability vector q, such that for all t_1

$$\lim_{t_2 \to \infty} \sup_{v(0)} \|v(t_1, t_2) - q\| = 0 \tag{3.4}$$

The strong ergodicity dictates that the chain converges in addition to the requirement of weak ergodicity. Now a sufficient condition for strong ergodicity:

Theorem 3.9. *If the time-inhomogeneous Markov chain is weakly ergodic, and for all t there is a stationary distribution $\pi(t)$, i.e., $\pi(t)P(t) = \pi(t)$ such that*

$$\sum_{t=0}^{\infty} \|\pi(t) - \pi(t+1)\| < \infty$$

then the chain is strongly ergodic. Moreover if $\mathbf{e}^* = \lim_{t\to\infty} \pi(t)$, then for all t,

$$\lim_{t_2\to\infty} \sup_{v(0)} \|v(t_1, t_2) - \mathbf{e}^*\| = 0.$$

Proof. See [7] or [8]. □

3.3.3 Convergence Criterion under Strong Reversibility Condition

Using Theorem 3.7, Mitra, Romeo and Sangiovanni-Vincentelli has shown that the setup of the algorithm gives rise to a weakly ergodic Markov chain under the conditions in the following theorem:

Theorem 3.10. *The Markov chain associated with the algorithm with the following Cooling Schedule*

$$T_t = \frac{\gamma}{\log(t + t_0 + 1)}, \quad t = 0, 1, 2, ... \tag{3.5}$$

is weakly ergodic for any $t_0 \geq 1$ and some finite γ

Proof. See [4]. □

Using Proposition 3.4 and Theorem 3.9, we have as a corollary

Theorem 3.11. *The Markov chain associated with the algorithm is strongly ergodic if it is weakly ergodic. Further,*

$$\lim_{t_2\to\infty} \sup_{v(0)} \|v(t_1, t_2) - \mathbf{e}^*\| = 0 \tag{3.6}$$

where $\mathbf{e}^*$ is as defined in Proposition 3.3.

In particular, the chain is strongly ergodic under the Cooling Schedule specified in Theorem 3.10.

3.4 Convergence Results for GO-POLARS with Cooling Schedule

One immediately sees that the strong reversibility condition fails to hold in more general conditions. In particular, in GO-POLARS where gradient information is used to propose the candidate solutions, finding the quasi-stationary distribution will involve solving a system of linear equations with rank $|S|$, and solutions are in general dependent on the proposal matrix.

This section addresses the following question: what happens if we relax the the strong reversibility condition imposed on the proposal matrix? And in particular, does GO-POLARS with cooling possess similar properties in terms of convergence?

Definition of quasi-stationary distribution in the previous section no longer applies; it is instead understood as the stationary distribution of the chain at a certain temperature. Note that it is again a construct and not an observable.

Hajek has shown that the convergence result still holds, although somewhat differently from in previous theorems, under the assumption that the proposal matrix is time-homogeneous and *weakly reversible* [9]. He did not attempt to show that the chain is strongly ergodic (since it is possible that the chain is not!), but rather took a probabilistic approach in showing that the chain will eventually escape the local minima when the Cooling Schedule is suitably designed.

It was not known whether the same result holds if we relax the condition on the proposal matrix even further. It is of interest to know what happens if the proposal matrix itself is time-inhomogeneous. For example we perform the search on a slightly different neighbourhood or a neighbourhood with a different proposal distribution at each step, as is done in GO-POLARS where the algorithm utilizes gradient information in generating candidate solutions. We seek to establish some results based on an observation on convergence by Faigle and Kern [10].

3.4.1 Time-homogeneous and weakly reversible proposal matrix

Some sort of reversibility condition is still needed for us to proceed with the quest of convergence. For if in the extreme one could enter a local optimal but proposes no solution that 'climbs hills', the algorithm will fail. It turns out that the following condition is useful.

We define the *weak reversibility* condition (WR) as follows

Definition 3.12. A proposal matrix is said to satisfy *weak reversibility* condition (WR) if each connected component of $G(c_i)$ is strongly connected, where $G(c_i)$ is the graph induced by G on the vertex set $S(c_i) = \{j \in S : c_j \leq c_i\}$. A directed graph is called *strongly connected* if every pair of vertices (u, v) is connected by a path from u to v.

A necessary and sufficient condition for the chain to converge to the optima is presented in the main result by Hajek.

Theorem 3.13. *We say that state i communicates with S^* at height h if there exists a path in G (with each vertex of the path being a neighbour of the previous vertex) that starts at i and ends at some element of S^* and such that the largest cost along the path is no larger than $c_i + h$. Let d^* be the smallest number that every $i \in S$ communicates with S^* at height d^*. Then a SA algorithm that satisfies (WR) gives rise to a chain such that*

$$\lim_{t\to\infty} P(x_t \in S^*) = 1$$

if and only if

$$\sum_{t=1}^{\infty} \exp\{-d^*/T_t\} = +\infty.$$

Proof. See [9]. □

3.4.2 Time-inhomogeneous and weakly reversible proposal matrix

Weak-reversibility alone is not enough. When the proposal matrix itself is time-inhomogeneous, we need some condition to guarantee that the probability of escaping as time passes is bounded from below.

We define the ε-condition (EC) as follows,

Definition 3.14. A family of stochastic matrices $\{\mathbf{A}(t)\}$ is said to satisfy the ε-condition if there exists an ε such that for each t and matrix $\mathbf{A}(t) = (a_{ij}(t))$,

$$a_{ij} \geq \varepsilon \quad \text{if} \quad a_{ij} > 0$$

whenever $i \neq j$, and all $\mathbf{A}(t)$ have the same neighbourhood structure, or equivalently, are associated with the same directed graph G.

Faigle and Kern observed that under the conditions (EC) and (WR), results analogous to that of Proposition 3.3 hold. Their main result is presented here.

Theorem 3.15. *Let* $\{\mathbf{A}(t)\}$ *be a family of irreducible stochastic matrices satisfying the conditions (EC) and (WR), then there exists a constant* K *and some* t_0 *such that*

$$\pi_j(t) \leq K f_t(c_i, c_j)\pi_i(t)$$

whenever $t > t_0$, $c_j > c_i$.

Proof. See [10]. □

Corollary 3.16. *Assume that (EC) and (WR) holds for* $\{\mathbf{A}(t)\}$, *and further,* $\lim_{t\to\infty} f_t(c_i, c_j) = 0$ *for all* $c_j > c_i$, *then*

$$\lim_{t\to\infty} \pi_j(t) = 0$$

for all $j \in S \backslash S^*$.

Similar to Proposition 3.3, this result states that if the temperature decreases so slow the chain remains at any temperature for an infinite number of steps, and the stationary probability is reached, the chain will eventually lead to the optimal solutions. Again such a Cooling Schedule does not exist. The first of the two questions remains: is there a Cooling Schedule for which the chain does converge? However, the second question, i.e., will the limiting distribution be e*, is not well-defined since the quasi-stationary distribution may no longer be of the particular form in the case of a strongly reversible proposal matrix, and a limiting distribution may not exist at all.

The answer to the first question is again affirmative.

We will construct a Cooling Schedule for which the chain converges under the conditions (EC), (WR), and the condition that the global optimal is unique.

3.4.3 Existence of a Cooling Schedule when global optimal is unique

The strategy of he proof is to first show that there is a Cooling Schedule with temperatures decreasing fast enough such that a monotonicity condition similar to that of proposition 3.4 is guaranteed. We then "slow down" the Cooling Schedule to keep the temperature constant at each step for a finite number of iterations such that the chain is guaranteed to be weakly ergodic. Strong ergodicity will follow from weak ergodicity and monotonicity of the quasi-stationary distributions.

Proposition 3.17. *Under (EC) and (WR), for any sequence of Cooling Schedule* $(T_t)_{t=0}^{\infty}$, $\lim_{t\to\infty} T_t = 0$ *there is a subsequence* $(T_{t_n})_{n=0}^{\infty}$ *such that*

$$\sum_{i\in S^*} \pi_i(t_{n+1}) - \sum_{i\in S^*} \pi_i(t_n) > 0 \quad \forall n$$

and for all $i \in S\backslash S^*$,

$$\pi_i(t_{n+1}) - \pi_i(t_n) < 0 \quad \forall n$$

Proof. Since the set S is finite, corollary 3.3 implies that

$$\lim_{t\to\infty} \sum_{i\in S\backslash S^*} \pi_i(t) = 0$$

consequently,

$$\lim_{t\to\infty} \sum_{i\in S^*} \pi_i(t) = 1.$$

By Bolzano–Weierstrass theorem there is a subsequence (t_n) such that the sequence $\sum_{i\in S^*} \pi_i(t_n)$ is monotone increasing and has the same limit, i.e., $\lim_{n\to\infty} \sum_{i\in S^*} \pi_i(t_n) = 1$. The first part is shown.

By the same argument we can sieve the sequence (t_n) to obtain a subsequence (t'_n) such that the quasi-stationary probability on the first element of the set $S\backslash S^*$ is monotone decreasing. Since the set S is finite, the desired subsequence is obtained after a finite number of operations. □

In particular, if the global optimal is unique, there is a Cooling Schedule such that the quasi-stationary probability is monotone increasing on the global optimal and monotone decreasing on all $i \in S\backslash S^*$.

We need some definitions related to the structure of the graph the cost function. The definitions follow [4].

We call $S_m := \{i \in S | c_j \leq c_i, \forall j \in N(i)\}$ the set of local maxima of the cost function. and let

$$r := \min_{i\in S\backslash S_m} \max_{j\in S} d(j,i)$$

be the radius of the graph, where $d(j,i)$ is the distance from j to i measured by the minimum number of edges of a path connecting j to i. let $\hat{i}$ be a vertex where the minimum is attained, call $\hat{i}$ the center of the graph.

A Lipschitz-like constant bounding the local slope of the cost function is given by

$$L := \max_{i \in S} \max_{j \in N(i)} |c_j - c_i|.$$

Similarly a lower bound on the local slope of the cost function is given by

$$l := \min_{i \in S \setminus S_m} \Big\{ \min_{\substack{j \in N(i) \\ c_j > c_i}} |c_j - c_i| \Big\}$$

Proposition 3.18. *When the proposal matrix satisfies conditions (EC) and (WR), the Markov chain associated with the algorithm is weakly ergodic if*

$$\sum_{k=k_0}^{\infty} \exp\Big\{ - \frac{rL}{T_{kr-1}} \Big\} = \infty. \tag{3.7}$$

Proof. We first give an estimate of the lower bound of the transition probability to the center of the graph, and then provide an upper bound for the coefficient of ergodicity.

For i, j such that $j \in N(i)$, $p_{ij}(t) \geq \varepsilon \exp(-L/T_t)$, for all t.

For diagonal elements of P, $p_{ii}(t)$ where $i \in S_m$, we do not seek a lower bound since p_{ii} may be 0.

For diagonal elements of P, $p_{ii}(t)$ where $i \in S \setminus S_m$, $p_{ii}(t)$ is bounded below by the constant $\varepsilon(1 - \exp(-l/T_0))$ for all t, because

$$\begin{aligned} p_{ii}(t) &= 1 - \sum_{j \in N(i)} p_{ij}(t) = 1 - \sum_{j \in N(i)} a_{ij}(t) f_t(c_i, c_j) \\ &= \sum_{j \in N(i)} a_{ij}(t)[1 - f_t(c_i, c_j)] \\ &= \sum_{\substack{j \in N(i) \\ c_j > c_i}} a_{ij}(t)[1 - f_t(c_i, c_j)] \\ &\geq \varepsilon(1 - \exp(-l/T_t)) > 0 \end{aligned}$$

Since the lower bound is increasing in t, $p_{ii}(t)$ is bounded below by $\varepsilon(1 - \exp(-l/T_0))$ for all t. Hence there exists some integer k_0 such that for all $i \in S \setminus S_m$

$$p_{ii}(t) \geq \varepsilon \exp(-L/T_t), \quad t \geq (k_0 - 1)r,$$

since the left-hand side is bounded below by some constant greater than 0, and the right-hand side is monotonically decreasing to 0.

Therefore we obtain the lower bound for the r-step transition probability from any $j \in S$ to the center of the graph for $t \geq k_0 r$, i.e., all the elements of the $\hat{i}$-th column of $p_{j\hat{i}}(t-r,t)$ is bounded below by

$$\begin{aligned} p_{j\hat{i}}(t-r,t) &\geq \prod_{n=t-r}^{t-1} \{\varepsilon \exp(-L/T_n)\} \\ &= \varepsilon^r \exp\{-\sum_{n=t-r}^{t-1} L/T_n\} \\ &\geq \varepsilon^r \exp\{-rL/T_{t-1}\}. \end{aligned}$$

since there is at least one path from j to $\hat{i}$ in r-steps (possibly contains transitions from a vertex to itself), with transition probability at each step bounded below by $\varepsilon \exp(-L/T_n)$.

Hence the coefficient of ergodicity τ_1 for the r-step transition matrix is bounded from above by

$$\begin{aligned} \tau_1(P(kr-r,kr)) &\leq 1 - \min_{ij}\{\min(p_{i\hat{i}}, p_{j\hat{i}})\} \\ &\leq 1 - \varepsilon^r \exp\{-rL/T_{kr-1}\}, \quad \forall k \geq k_0. \end{aligned}$$

Therefore, by Theorem 3.7, the Markov chain is weakly ergodic if

$$\sum_{k=k_0}^{\infty} \exp\Big\{-\frac{rL}{T_{kr-1}}\Big\} = \infty.$$

□

We have in fact arrived at the same sufficient condition for the chain to be weakly ergodic. In particular, the Markov chain associated with the Cooling Schedule in Theorem 3.10

$$T_t = \frac{\gamma}{\log(t+t_0+1)}, \quad t = 0,1,2,... \tag{3.8}$$

where $\gamma \geq rL$, is weakly ergodic.

Theorem 3.19. *There is a Cooling Schedule such that the Markov chain is strongly ergodic, provided that the global optimal is unique, and the proposal matrix satisfies conditions (EC) and (WR). In this case for all* t_1

$$\lim_{t_2 \to \infty} \sup_{v(0)} \|v(t_1,t_2) - \mathbf{e}^*\| = 0 \tag{3.9}$$

where $\mathbf{e}^*$ *is a delta function on the optimal solution.*

Proof. We construct a sequence (T'_t) from (T_t) that satisfies condition in Proposition 3.18, and show that it has the desired property.

By Proposition 3.17, there is a subsequence of (T_t) such that the quasi-stationary probability on S^* (in this case a unique candidate) is monotonically increasing, and quasi-stationary probability on all other candidates monotonically decreases in t_n, and hence, in n.

Let

$$T'_t = \begin{cases} T_t & \text{if } t < t_1 \\ T_{t_n} & \text{if } t_n \leq t < t_{n+1} \end{cases}$$

Since $T'_t \geq T_t$ for all t, by Proposition 3.18 and the fact that (T_t) satisfies its condition, we have $\sum_{k=k_0}^{\infty} \exp\left\{-\frac{rL}{T'_{kr-1}}\right\} = \infty.$, and thus conclude that the chain is weakly ergodic. By construction, the algorithm also satisfies the conditions (EC) and (WR) since the proposal matrix is not altered.

Denote the quasi-stationary distributions of the Markov chain associated with the Cooling Schedule (T'_t) as $\pi(t)'$. It remains to show that

$$\sum_{t=0}^{\infty} \|\pi'(t) - \pi'(t+1)\|_1 < \infty.$$

When $t \geq t_0$,

$$\begin{aligned} \|\pi'(t) - \pi'(t+1)\|_1 &= \sum_{i \in S^*} [\pi'(t+1) - \pi'(t)] + \sum_{i \in S \setminus S^*} [\pi'(t) - \pi'(t+1)] \\ &= 2 \sum_{i \in S^*} [\pi'(t+1) - \pi'(t)]. \end{aligned}$$

Hence

$$\begin{aligned} \sum_{t=0}^{\infty} \|\pi'(t) - \pi'(t+1)\|_1 &= \sum_{t=0}^{t_0-1} \|\pi'(t) - \pi'(t+1)\|_1 + \sum_{t=t_0}^{\infty} \|\pi'(t) - \pi'(t+1)\|_1 \\ &= \sum_{t=0}^{t_0-1} \|\pi'(t) - \pi'(t+1)\|_1 + 2 \sum_{i \in S^*} [\pi'(t+1) - \pi'(t)] \\ &\leq 2t_0 + 2 < \infty. \end{aligned}$$

In view of Theorem 3.9, we conclude that the Markov chain is strongly ergodic. By Corollary 2.16, the limiting distribution is e^*. □

We briefly summarize this chapter here.

In section 3.3 we stated that, under strong reversibility condition, the quasi-stationary distribution converges to the optimal distribution as $t \to \infty$, or $T \to 0$, and further, a Cooling Schedule of the form in Theorem 3.10 is sufficient to guarantee convergence of the solutions.

In section 3.4 we have shown that when the condition on proposal matrix is relaxed to the weak reversibility condition, the quasi-stationary distributions still converges to the optimal distribution as $t \to \infty$. Further, we have shown the existence of a Cooling Schedule that will guarantee convergence in distribution of the random variables (X_t) to the random variable whose probability measure is concentrated on the global optimal solution.

Chapter 4

Numerical Experiments

We demonstrate both convergence of quasi-stationary distributions and the finite time behavior of the algorithms. Comparisons are between the Simulated Annealing algorithm (i.e. when strong reversibility condition is satisfied, no gradient information used), and GO-POLARS with Cooling Schedules (i.e. when weak reversibility condition is satisfied and proposal based on gradient information) under the same acceptance rules and Cooling Schedules.

Greedy algorithms such as Steepest Descent are not included for comparison for two reasons. If, say, a Cooling Schedule is introduced and the modified Steepest Descent seen a limiting case of GO-POLARS with Cooling Schedule, one immediately sees that Weak Reversibility condition fails at boundaries for constrained optimization. On the other hand, if we do not modify the algorithm, a Cooling Schedule is absent for a fair comparison, and more importantly, the algorithm may lead to suboptimal solutions for unfavorable starting conditions, which is always inferior to the previous two algorithms.

4.1 Convergence of Quasi-stationary Distribution

We illustrate the convergence process of quasi stationary distributions under both algorithms with five test functions: Ackley function, Griewank function, Rosenbrock function, Goldstein-Price function, and Styblinski-Tang function.

The Styblinski-Tang function is defined on the hypercube $x_i \in [-5, 5] \quad \forall i \in \{1, \ldots, d\}$ to be

$$f(\mathbf{x}) = \frac{1}{2} \sum_{i=1}^{d} (x_i^4 - 16x_i^2 + 5x_i)$$

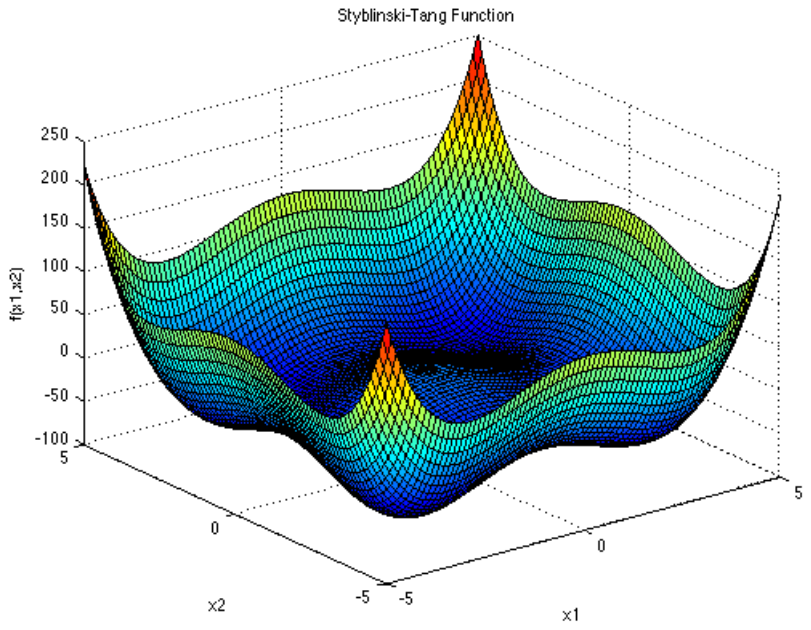

FIGURE 4.1: Styblinski-Tang function

with global minimum $f(-2.903534, \ldots, -2.903534) = -39.16599d$, and multiple local minima.

With average step size set at one tenth of the range of each variable, we compare the quasi-stationary distributions of Simulated Annealing and GO-POLARS with cooling when a moderate amount of gradient information is used for proposing candidate solutions (angle of step increment measured from $-\nabla f$ follows a Beta(2,2)).

Upper panel of Figure 4.2 shows the heat-map of estimated quasi-stationary distribution when temperature decreases from 2^6 to 2. Lower panel shows the corresponding estimation for GO-POLARS with cooling at the same temperatures. Darker regions indicates a higher probability density and lighter regions a lower probability density.

GO-POLARS with Cooling Schedules demonstrates potential ability in identifying likely regions of local and global optima compared to standard Simulated Annealing. At a fixed temperature, GO-POLARS is able to 'recognize' and move in the direction towards regions potentially containing local and global optima, while Simulated Annealing, restraint by the symmetry of strong reversibility condition, is ignorant of gradient information. In numerical examples when moderate amount of gradient information is used (moderate exploitation/exploration in the terms used in [6]), GO-POLARS gives clearer identification of optima regions. We find this feature especially salient in the case when absolute difference between regions are small (e.g. Goldstein-Price and Griewank function on -10 to 10).

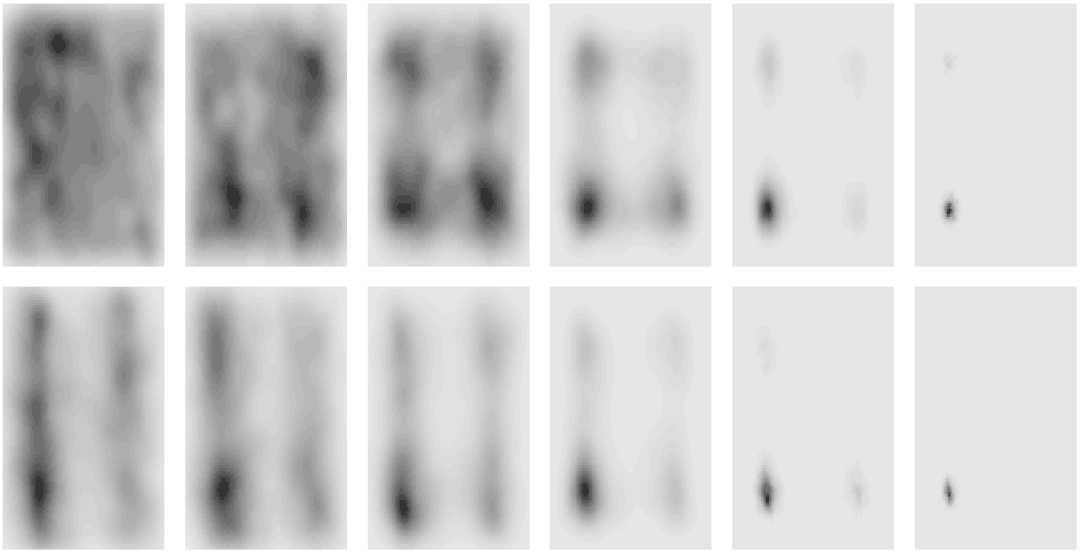

FIGURE 4.2: Heat-map of Estimated quasi-stationary distributions under Simulated Annealing (upper panels), and GO-POLARS with Cooling Schedule (lower panels) at temperatures 2^6, 2^5, 2^4, 2^3, 2^2, 2^1 (from left to right)

Convergence of quasi-stationary distribution is observed in all test cases for both Simulated Annealing and GO-POLARS with the same Cooling Schedules. We demonstrate only the output for the case with Styblinski-Tang function and refer readers to Appendix for output for the other four illustrations.

4.2 Finite-time Behavior

We examined some finite-time behaviors of the two algorithms. The first of our examples is on Goldstein-Price function:

$$\begin{aligned} f(\mathbf{x}) = &[1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + ex_2^2)] \\ &\times [30 + (2x_1 - 3x_2)^2(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)] \end{aligned}$$

which entails several local minima and attains global minimum $f(\mathbf{x})$ at $\mathbf{x} = (0, -1)$. Function values does not vary much relative to the range of the function values on a large region. Reader may refer to Appendix Figure A.3 for a graph of the function on $[-5, 5] \times [-5, 5]$.

We have run 1000 replications of both algorithms with the same Cooling Schedule as in Theorem 3.10 and the standard acceptance/rejection rule; initial solution at $(-2, 2)$.

We plotted performance of both algorithms in terms of their (1) ensemble average of objective values at each iteration, (2) ensemble average of current best objective values

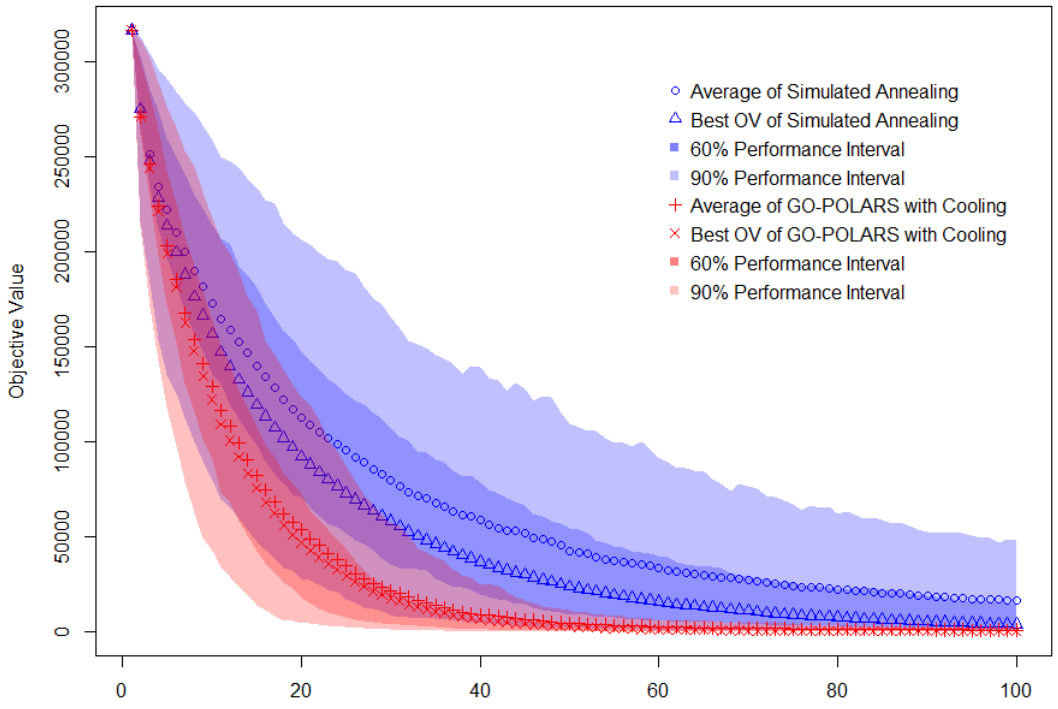

FIGURE 4.3: Ensemble average across 1000 replications of objective values, best objective values, and 60- and 90-percentile performance range in first 100 iterations when performing optimization on Goldstein-Price function; initial solution at (2,-2).

within replication at each iteration, (3) Central 60-percent of objective values across replications, and (4) Central 90-percent of objective values across replications; see Figure 4.3.

GO-POLARS with cooling out-performs standard Simulated Annealing in all measures of performances in this case, achieving a faster rate of convergence towards global minimum, and attains near optimum for most replications within 60 iterations.

Although GO-POLARS with a proper Cooling Schedule leads to convergence to the global optimal, finite-time behavior is sensitive to starting conditions, as we demonstrate with the following example on Styblinski-Tang function.

We have run 1000 replications of both algorithms with the same Cooling Schedule as in Theorem 3.10 and the standard acceptance/rejection rule. Initial solution is chosen at $(2.7, 2.7)$, which is near a local minimum of the function.

Again we plotted performance of both algorithms in terms of their (1) ensemble average of objective values at each iteration, (2) ensemble average of current best objective values within replication at each iteration, (3) Central 60-percent of objective values across replications, and (4) Central 90-percent of objective values across replications; see Figure 4.4.

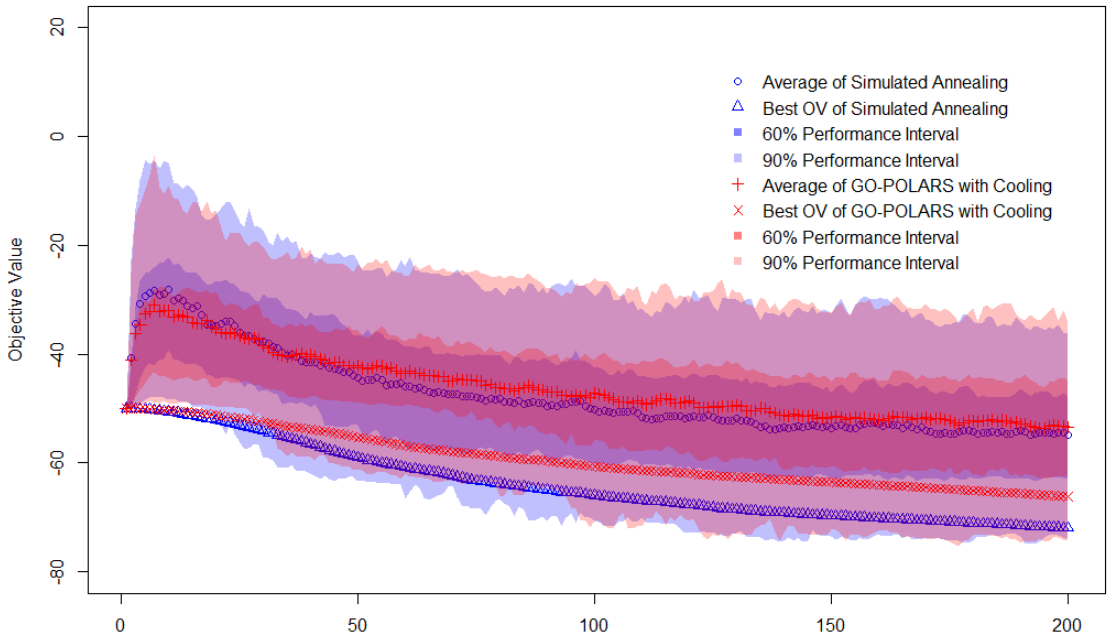

FIGURE 4.4: Ensemble average across 1000 replications of objective values, best objective values, and 60- and 90-percentile performance range in first 200 iterations when performing optimization on Styblinski-Tang function; initial solution at (-2.7,-2.7).

Although GO-POLARS with cooling starts to escape the local minimum, the escape is on average slower than that of the standard Simulated Annealing. Quantiles of performance does not seem distinctively different; ensemble average of current best objective values under GO-POLARS with cooling is dominated by Simulated Annealing.

The two numerical examples demonstrates some characteristics of GO-POLARS with cooling in finite time compared to Simulated Annealing and shows that finite-time performance is sensitive to starting conditions.

With favorable starting conditions, GO-POLARS with cooling out-performs standard Simulated Annealing in all measures of performances, achieving a faster rate of convergence towards global minimum. We may speak heuristically of such favorable in terms of short distance to the global optimum relative to the step-size, 'smoothness' (in contrast to 'wiggliness'), and convexity of objective functions.

Escaping of local minima is sometimes more difficult than with standard Simulated Annealing as the directed generation of candidate solutions makes 'hill-climbing' difficult in addition to the acceptance/rejection step in Simulated Annealing. The GO-POLARS's 'gravitating' proposal of candidate solutions in effect draws the chain towards nearest local minima, and the chain becomes similar to that generated by a greedy algorithm in the very short run, although it eventually escapes and converges to the global optimum.

Chapter 5

Conclusions and Future Studies

One of the primary rationales for injecting randomness into gradient based algorithms is in doing so enabling escapes from local optima. Gradient-based algorithms sometimes fail in finding global optimum and get stuck in the local optima for their greedy nature in search.

On the other hand, introducing gradient information to simulation optimization algorithms may help with discovering favorable regions at a faster speed, which could surpass that of standard Simulated Annealing.

We investigated GO-POLARS with Cooling Schedules which incorporates both gradient information and randomized search methods. The modified algorithm performs better than purely gradient-based algorithms in terms of global optimization.

We summarize our work, and point out some limitations and directions for further investigation.

5.1 Project Achievements

Some theoretical results that supports and justifies the use of the modified algorithm in practice were established. In particular, we have shown under suitable conditions the existence of Cooling Schedules which guarantee convergence in probability to the global optimum.

We have conducted numerical experiments for illustrations of the convergence process, as well as numerical experiments that evaluate the performance of the algorithm for test functions. Convergence of quasi-stationary distribution to the optimum is observed

in numerical examples. We find finite-time behavior of the algorithm sensitive to initial conditions. Comparisons are drawn with standard Simulated Annealing algorithm as benchmark. Examples also illustrated situations where the modified algorithm are superior or inferior to the benchmark. Favorable starting points can speed up the the process in attaining global optimal, while starting near a local but not global would lead to some 'stickiness' in escaping local optimal for GO-POLARS with cooling.

5.2 Limitations and Further Study

We proved in Section 3.4 the existence of a Cooling Schedule that guarantees convergence to global optimal; further, the proof is constructive. Heuristically speaking, the construction requires the temperature to 'cool' at a slow rate such that a the Markov Chain is weakly ergodic, and at the same time fast enough such that the chain satisfies a sufficient condition for strong ergodicity.

Such construction, however, is dependent on all values of the objective function at the finite set of solution space, which is similar to the determination of γ in Theorem 3.10. It is of interest to improve upon the construction to bound the gaps in the sequence in the sieving process to arrive at a more practically attractive result.

Numerical experiments may be performed to examine the trade-off between exploration and exploitation in the modified algorithm.

Appendix A

Illustrations for Convergence of Quasi-stationary Distribution

A.1 Ackley Function

$$f(\mathbf{x}) = -a\exp\Big(-b\sqrt{\frac{1}{d}\sum_{i=1}^{d}x_i^2}\Big) - \exp\Big(\frac{1}{d}\sum_{i=1}^{d}x_i^2\Big) + a + \exp(1)$$

where $a = 20, \quad b = 0.2, \quad c = 2\pi$

A.2 Goldstein-Price Function

$$\begin{aligned} f(\mathbf{x}) = &[1 + (x_1 + x_2 + 1)^2(19 - 14x_1 + 3x_1^2 - 14x_2 + 6x_1x_2 + ex_2^2)] \\ &\times [30 + (2x_1 - 3x_2)^2(18 - 32x_1 + 12x_1^2 + 48x_2 - 36x_1x_2 + 27x_2^2)] \end{aligned}$$

A.3 Griewank Function

$$f(\mathbf{x} = \sum_{i=1}^{d}\frac{x_i^2}{4000} - \prod_{i=1}^{d}\cos\Big(\frac{x_i}{\sqrt{i}}\Big) + 1$$

A.4 Rosenbrock Function

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

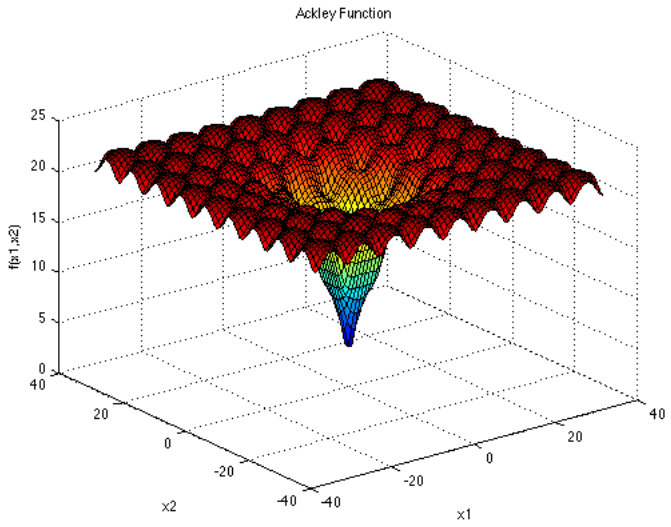

FIGURE A.1: Ackley function

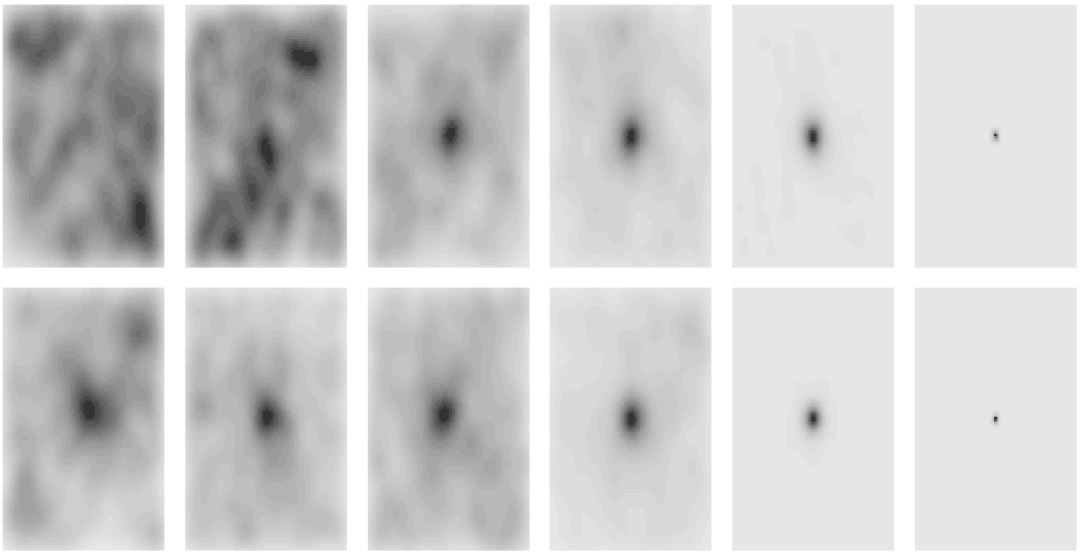

FIGURE A.2: Heat-map of Estimated quasi-stationary distributions under Simulated Annealing (upper panels), and GO-POLARS with Cooling Schedule (lower panels) at temperatures 2^6, 2^5, 2^4, 2^3, 2^2, 2^1 (from left to right)

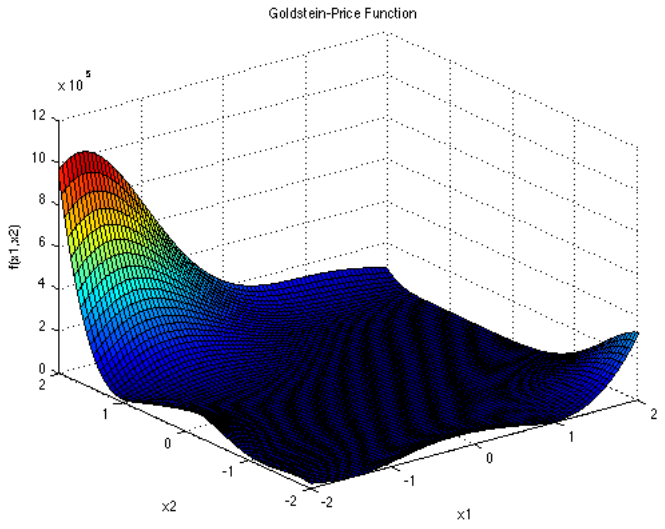

FIGURE A.3: Goldstein-Price function

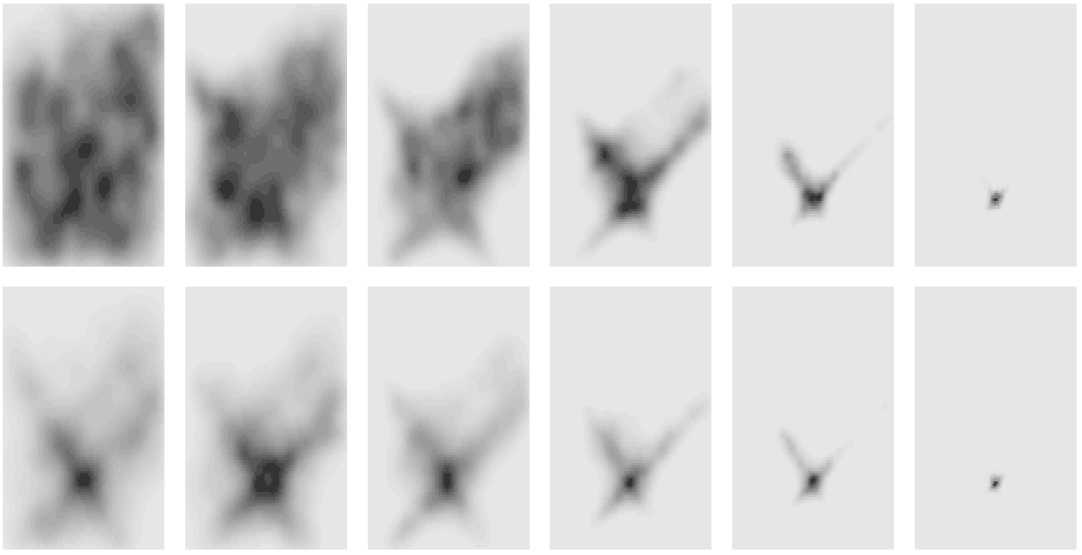

FIGURE A.4: Heat-map of Estimated quasi-stationary distributions under Simulated Annealing (upper panels), and GO-POLARS with Cooling Schedule (lower panels) at temperatures 2^18, 2^15, 2^12, 2^9, 2^6, 2^3 (from left to right)

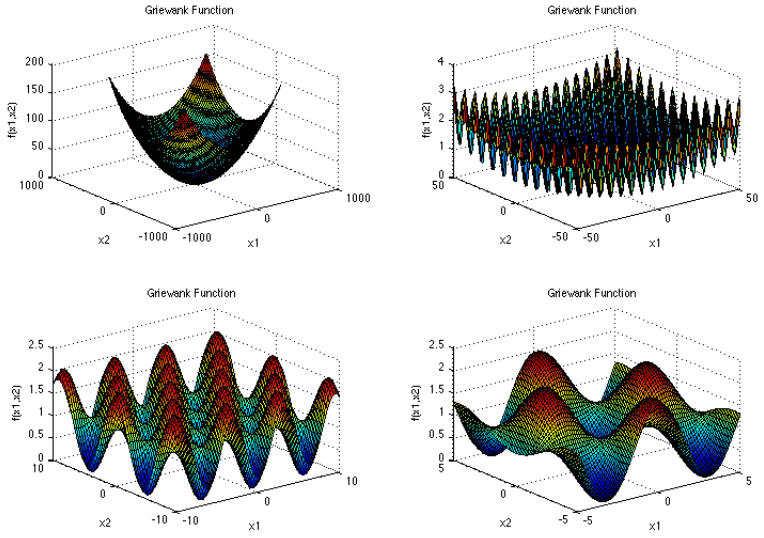

FIGURE A.5: Griewank function

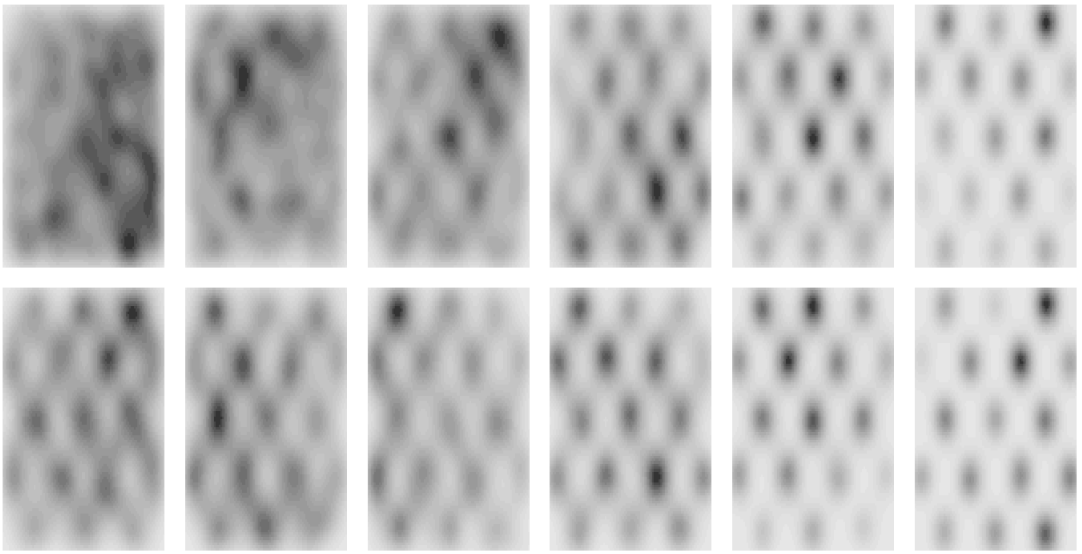

FIGURE A.6: Heat-map of Estimated quasi-stationary distributions under Simulated Annealing (upper panels), and GO-POLARS with Cooling Schedule (lower panels) at temperatures 2^3, 2^2, 2^1, 2^0, 2^{-1}, 2^{-2} (from left to right)

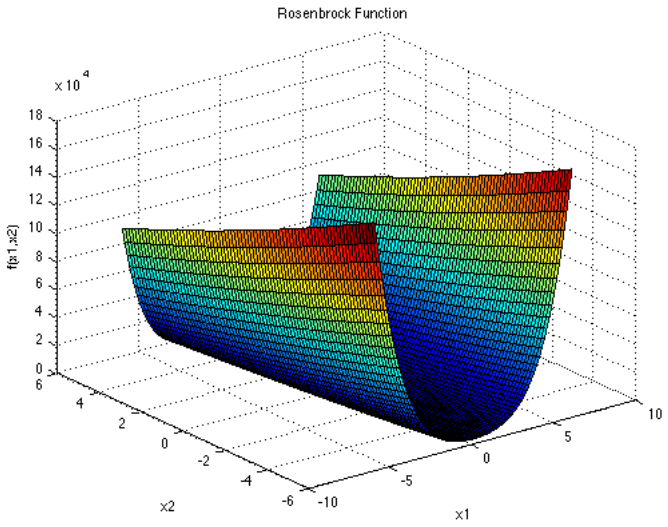

FIGURE A.7: Rosenbrock function

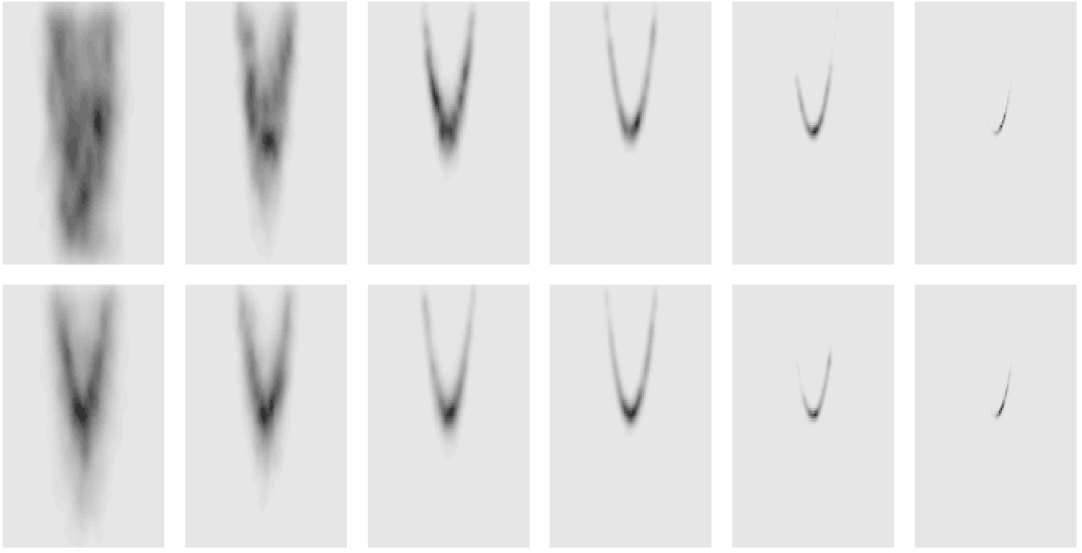

FIGURE A.8: Heat-map of Estimated quasi-stationary distributions under Simulated Annealing (upper panels), and GO-POLARS with Cooling Schedule (lower panels) at temperatures 2^15, 2^12, 2^9, 2^6, 2^3, 2^0 (from left to right)

Bibliography

[1] H. Li, L.H. Lee, and E.P. Chew. The steerable stochastic search - on the strength of hyper-spherical coordinates. 2013.

[2] K. C. Kiwiel and K. Murty. Convergence of the steepest descent method for minimizing quasiconvex functions. *Journal of Optimization Theory and Applications*, 89(1):221–226, 1996.

[3] A. Gelman et al. *Bayesian Data Analysis*. Chapman Hall, 2013. ISBN 9781439840955.

[4] F. Romeo D. Mitra and A. Sangiovanni-Vincentelli. Convergence and finite-time behavior of simulated annealing. *Advances in Applied Probability*, 18(3):747–771, 1986.

[5] D. Henderson, S. H. Jacobson, and A. W. Johnson. The theory and practice of simulated annealing. *Handbook of Metaheuristics, International Series in Operations Research and Management Science*, 57:287–319, 2003.

[6] R. Zhao. Evaluation of gradient oriented polar random search on exploration vs exploitation. 2014.

[7] D.L. Isaacson and R.W. Madsen. *Markov Chains: Theory and Applications*. Wiley, New York, 1976.

[8] M. Iosifescu. *Finite Markov Processes and Their Applications*. Wiley, New York, 1980.

[9] B. Hajek. Cooling schedule for optimal annealing. *Mathematics of Operations Research*, 13(2):311–329, 1988.

[10] U. Faigle and W. Kern. Note on the convergence of simulated annealing algorithms. *SIAM Journal on Control and Optimization*, 29(1):153–159, 1991.