On the Convergence of Genetic Algorithms

Matthias Löwe
Universität Bielefeld
Fakultät für Mathematik
Postfach 100131
33501 Bielefeld

Abstract

We analyze the convergence of a class of Genetic Algorithms by reducing them to a class of stochastic search algorithms on graphs. We show that, if the underlying random walk is irreducible and reversible, the invariant measure has a Gibbsian structure and that a logarithmic cooling schedule similar to that used in Simulated Annealing guarantees convergence to the uniform distribution on the set of minima of an energy function. Moreover, using results of Deuschel and Mazza, we bound the entrance time into typical subsets of the state space stochastically in terms of the Gibbs measure at time t.

Keywords: Genetic Algorithms, Spectral gap, Markov processes, Simulated Annealing

1 Introduction

In the sequel we are going to give a small contribution to the following well studied problem:
Let (X, E) be a finite, but very large graph with vertex set X and edge set E. Let $V : X \to \mathbb{R}$ be a real valued function of the vertices. Then the task to find the minimum value of V in reasonable time or, equally hard, to find a vertex, where this minimum is attained, is not an easy one.

The simplest deterministic search method, which definitely finds the global minimum, is the exhaustive or global search, that calculates explicitly the values of any vertex $x \in X$. For large sets X (which e.g. in the case of image processing are in the order of 10^{10^4}) this algorithm even if implemented on very fast computers obviously takes too long to be interesting. On the other hand the so called local search, an algorithm that makes use of the topological structure on X by starting in a vertex $x_0 \in X$ and then stepping to x_1, that vertex in $\{x_0\} \cup \mathcal{N}(x_0)$ with smallest V-value (where $\mathcal{N}(x)$ denotes the neighborhood of a vertex $x \in X$) and so on, soon gets trapped in a local minimum which usually is not the global one. So this algorithm is applicable only if one is satisfied with suboptimal solutions.

The class of stochastic search algorithms with the well known Simulated Annealing (SA) algorithm (introduced in 1983 by Kirkpatrick et. al. in [K83]) as its most prominent representative are based on the following idea. Run a time inhomogeneous Markov-chain or Markov-process on X such that the invariant measure at time t becomes more and more concentrated on the set of minima of V (e.g. in many cases the invariant measure at time t is a Gibbsian measure with energy V and at some inverse temperature β_t with respect to some a priori measure on X). So if one changes the invariant measure carefully enough (in t) such that the process is able to follow its invariant measure, the distribution of the Markov chain will converge to the uniform distribution on the set of global minima of V.

The problems implicit in the description of this class of stochastic search algorithms, i.e. the convergence of the algorithm and the optimal choice of the way the invariant measure changes in time (in the case of a Gibbsian measure as above this change is governed by the way β_t converges to infinity; in this case we will refer to the functional dependence of β_t from t as cooling schedule or freezing schedule) has intensively been studied for SA processes. So the convergence first has been investigated by Gidas [G85] and Mitra et al. [M85]. The problem of an asymptotically optimal cooling schedule was settled by Hajek

[Ha88] by proposing a freezing schedule β_t of the form

$$\beta_t = \frac{1}{m} \log(A + Bt)$$

where A and B are positive constants and m is, loosely speaking, the largest hill to be climbed to reach the global minimum (so m is bounded by $M := \max_x V(x)$.
By using Freidlin-Wentzell techniques on perturbed Markov chains Catoni in [C90] was able to bound the probability that the algorithm at time t is above a certain energy level by a constant times $t^{-\alpha}$ with nearly optimal α.

A new light was shed on the intrinsic mechanism of the convergence of SA algorithms by the work of Holley and Stroock [HS88], who analyzed the L^2-distance of the Radon-Nikodym derivative of the distribution of the process at time t with respect to its invariant measure to one by means of differential equations. The solution of these differential equations was worked out by assuming a Sobolev inequality, which was verified by analyzing the spectral gap of the operator associated to the Markov process. A refined analysis of these spectral gap estimates by Götze [G92] led to a more precise and hence applicable estimate of the entrance probability into typical subsets of the state space. The ideas of [G92] were generalized to a wide class of Markov processes of the Freidlin-Wentzell type by Deuschel and Mazza in [DM94].
From the rich literature concerning applications of SA we only refer to [V93] for a survey, and [GG84] and [G92] for an application to the reconstruction of disturbed images.

Another class of stochastic search algorithms, which also has found a wide range of applications (see e.g. [D88] or [Go89]) are the so called Genetic Algorithms or Adaptive Algorithms. These algorithms were already introduced in 1975 by Holland (see [Ho75]). The basic idea behind Genetic Algorithms is to regard a vertex $x \in X$ as a potential individual of a population (which accordingly is a subset of the vertex set X) and to understand $V(x)$ or even better $-V(x)$ as the fitness of the individual x. Then the population reproduces according to a reproductive plan consisting of the application of genetic operators. From an abstract point of view genetic operators are nothing but functions from a random subset (of a certain size) of X to another subset of X. More intuitively the genetic operator represents the reproduction of a group of parents (the first subset) resulting in a number of children (the second subset).
After every step of reproduction we apply the rule of "survival of the fittest", i.e. we select some of the individuals no matter how old they are to die. This selection is carried out with the help of a probability measure on the individuals that prefers individuals with low V-values. To achieve convergence of the

algorithm to the set of minima of V one has to choose this selection scheme, i.e. mainly the probability distribution mentioned above, time dependent in such a way that it finally (i.e. for $t \to \infty$) deterministically selects those individuals with minimum V-values.

In spite of their practical importance Genetic Algorithms are not very well studied from a theoretical point of view. As a matter of fact, up to our best knowledge, the only mathematically rigorous analysis is contained in the Ph.D. thesis of Cérf [Ce94]. There Cérf analyzes Genetic Algorithms by using Freidlin-Wentzell techniques.

The investigations of this paper are guided by an example of an Genetic Algorithm on the discrete unit cube $\{0,1\}^d$ for some d large enough. This corresponds to considering each individual as a string of genes. The reproduction of a population on $\{0,1\}^d$ in this example is governed by mutation and crossing-over. Crossing-over is chosen since it is the most natural way of describing the reproduction of two individuals, whereas mutation works as a sort of error-correcting operator ensuring the conditions of irreducibility and reversibility of the algorithm.
By generalizing this example we are able to analyze a wide class of Genetic Algorithms. Note that, of course, this class is not exhaustive and that especially the reversibility condition is quite restrictive. On the other hand, if this condition is not fulfilled, either the algorithm converges to suboptimal solutions or the population size depends on the difficulty of the energy landscape V in an uncontrolable way (see e.g. [Ce94], where the author shows, that for each population size there is an optimization problem such that the population size is insufficent to solve the problem).

The central aim of this paper is to investigate the resulting class of Genetic Algorithms, especially to examine the speed of convergence, and to bound the time they spend before entering sets of low V-values as has been done in [G92] for SA algorithms. This is done by making use of results by Deuschel and Mazza in [DM94] for Markov processes from the Freidlin-Wentzell class similar to SA.

To this end we introduce a class of stochastic search algorithms on graphs. The basic idea behind this class is to partition the neighborhood $\mathcal{N}(x)$ of each vertex $x \in X$ (i.e. $\mathcal{N}(x) = \cup N_i^x$) and to endow the elements of the partition with transition probabilities $p^x(N_i^x)$ such that

$$\sum_i p^x(N_i^x) = 1 \qquad \text{for each } x \in X.$$

Assume that the partition fulfills a certain symmetry condition to be formulated later on. Then the algorithm works in the following way: start in a vertex

$x_0 \in X$, choose one of the elements of the partition with probability $p^{x_0}(N_i^{x_0})$, and finally select one element of $N_i^{x_0} \cup \{x_0\}$ according to a Gibbsian distribution with respect to the energy V and at some inverse temperature β. While carrying on with this procedure as above slowly cool down the process, i.e. increase β.

Note that if we choose the partitions of the neighborhoods in such a way that each element contains exactly one vertex the algorithm becomes a slightly modified (and slightly slower) version of SA. If, on the other hand, for each $x \in X$ the partition of each $\mathcal{N}(x)$ consists of only one element (namely $\mathcal{N}(x)$), the algorithm becomes a sort of local search with Gibbsian weight factors, which, up to our best knowledge has not been investigated yet. The latter however usually does not fulfill the symmetry condition imposed below. Thus its analysis is to be postponed to a forthcoming paper.
However, to make the algorithms effective , it is desirable to choose the size of the elements of the partitions depending on how easily a new value $V(y)$ can be calculated, provided $y \in \mathcal{N}(x)$ and $V(x)$ is already known. If this can be carried out rather quickly, as e.g. in the case for many Genetic Algorithms, the elements of the partition can be chosen to contain many vertices, and our algorithms have the advantage - e.g. compared to SA - of easy parallelisation (for some progress concerning parallelisation of SA see e.g. the Ph.D. thesis of Trouvé [T93]).

We divide the rest of the paper into three parts. In Section 2 we introduce the class of Genetic Algorithms we want to analyze. We present our central example and the main results on the convergence of Genetic Algorithms. Section 3 contains a translation of the Genetic Algorithms into search algorithms on graphs. In Section 4 we prove the convergence results stated in Section2 . We apply the methods of Deuschel and Mazza [DM94].

2 Genetic Algorithms

Genetic Algorithms were introduced by Holland in his fundamental book "Adaption in Natural and Artificial Systems" [Ho75]. The basic idea behind Genetic Algorithms is vaguely described as taking the principles of evolutionary processes over to "populations" of particle systems on some set, e.g. some subset of a high dimensional lattice. These principles can generally be described in terms of reproduction and selection. The latter can be easily understood for natural processes. If an individual fails to fit to its environment properly, it probably dies. In the case of particle systems this test for fitness is performed with the help of a function V which we would like to minimize. So, the smaller the V-value of a particle is, the larger is its probability to survive. Concerning reproduction there are several processes in nature which can serve as an example, among them recombination, crossing over , inversion, and mutation. In Chapter 6 of [Ho75] Holland shows how to adapt these principles to artificial systems.

As this discussion shows we on one hand can try to model evolutionary processes by particle systems on a certain state space. On the other hand, if the model works, we also can try to minimize given functions V on a certain set by regarding it as the fitness function of Genetic Algorithms. This idea has given Genetic Algorithms a wide range of applications. However, rigorous results on Genetic Algorithms a hardly available in literature. As a matter of fact, up to our best knowledge the only convergence theorems for Genetic Algorithms are those in the Ph.D. thesis of Cérf [Ce94] proved with Freidlin-Wentzell techniques. Since we work in a slightly different model, we obtain results of a different flavor. Especially we are able to drop some bounds on the population size and replace them by a reversibility condition. Moreover we can analyze the speed of convergence to equilibrium.

We will now give the above ideas a formal framework.
To this end introduce the space $\tilde{X}$ as the set of possible states of a single individual, i.e. all $\tilde{x} \in \tilde{X}$ are possible configurations of a single individual (so every $\tilde{x}$ can be understood as sequence of genes e.g.). Now fix a number $\nu \in \mathbb{N}$ and keep ν fixed during the evolution of the process. ν will denote the number of individuals we will work with. So the state space of the whole population of ν particles is given by

$$\bar{X} := \prod_{i=1}^{\nu} \tilde{X}.$$

It will be convenient to regard two populations as identical, if one can be obtained by a permutation of the individuals of the other. So instead of $\bar{X}$ we consider the space

$$X = \bar{X}/S_\nu$$

where for the moment S_ν denotes the symmetric group on the set $\{1, \ldots, \nu\}$. However, in a slight abuse of notation we will denote an element in X by a vector $(x_1, \ldots, x_\nu)$ rather than by the equivalence class of this vector under permutations.

The dynamics of Genetic Algorithms are governed by a set of genetic operators $\{G_1, \ldots, G_\mu\}$. Formally each genetic operator G_j is a random variable

$$G_j : \tilde{X}^{l_j} \to \tilde{X}$$

(for $1 \leq l_j \leq \nu$, $j = 1, \ldots, \mu$) describing how a new particle emerges from one or several old particles. The most frequently used genetic operators are mutation (which historically has been the first example of a genetic operator), and crossing over or recombination, respectively. They are conveniently described on a subset of $\mathbb{Z}$ and we will meet them in the example following this description.

It should be emphasized here that, although all examples of genetic operators in nature work on a set of at most two individuals there is no reason for us to restrict ourselves to this case. So in general in our model children may have more than two parents. On the other hand it will be extremely important that not more than one new particle occurs at any time (no twins!).

Now we are able to describe the process we have always had in mind in the first three sections:
Suppose we are given a function $V : \tilde{X} \to \mathbb{R}$, then

a) Choose a monotone function $\beta_t \to_{t\to\infty} \infty$

b) Choose at random (i.e. with the uniform distribution on X) an a priori distribution of the ν individuals.

c) Having arrived at a certain population $(x_1^s, \dots, x_\nu^s) \in X$ at time s wait an exponential time with mean one.

d) Then, say at time t, choose a genetic operator $G_j \in \{G_1, \dots, G_\mu\}$ with respect to a given time invariant probability distribution on the set $\{1, \dots, \mu\}$.

e) Choose l_j individuals $y_1, \dots, y_{l_j}$ of $(x_1^s, \dots, x_\nu^s)$ with respect to a time invariant probability distribution on the population x (see below).

f) Apply G_j, i.e. build $x_0^s = G_j(y_1, \dots, y_{l_j})$.

g) Select an element $x_k^s \in (x_0^s, x_1^s, \dots, x_\nu^s)$ according to

$$\gamma_t(x_k^s) := \frac{e^{\beta_t V(x_k^s)}}{\sum_{i=0}^{\nu} e^{\beta_t V(x_i^s)}}.$$

h) The new population is given by

$$(x_1^t, \dots, x_\nu^t) = (x_0^s, x_1^s, \dots, x_{k-1}^s, x_{k+1}^s, \dots, x_\nu^s).$$

Here the notion "time invariant probability distribution on the population x" in e) precisely means that we are given a set of probability distributions (one for every possible population $x \in X$ and every genetic operator G_j) and that this set is time invariant. However, introducing abbreviations for such families of distributions only would lead to a mess of notation.
To ensure good ergodic properties of the Genetic Algorithm we impose the following two conditions

($\hat{I}$) The time homogeneous random walk described by a)-h) above when substituting γ_t in g) by the uniform distribution on $(x_0^s, x_1^s, \ldots, x_\nu^s)$ is irreducible, i.e. for each pair $x, y \in X$ there exists a positive integer N such that the poplulation y can be reached from the population x in N steps.

($\hat{R}$) The time homogeneous random walk described by a)-h) above when substituting γ_t in g) by the uniform distribution on $(x_0^s, x_1^s, \ldots, x_\nu^s)$ is reversible with respect to a probability distribution ϱ on X. This means

$$\varrho(x)Prob(x,y) = \varrho(y)Prob(y,x) \tag{1}$$

for all $x, y \in X$. Here $Prob(x,y)$ denotes the probability to get from poluation x to population y.

Before stating our central result let us mention the example we always have in mind when talking about Genetic Algorithms.

Example 2.1 *This example describes a polpulation of ν individuals "moving forward" by mutation and crossing over . When crossing over works two individuals are randomly selected and cut into two pieces and the pieces belonging to different individuals are glued together. If mutation works one individual is randomly chosen and changed in such a way that each "gene" has a very small but positive probability to be "flipped".*
To this end consider each individual as a sequence of genes and encode this sequence by a 0-1-sequence. So $\tilde{X} = \{0,1\}^d$ for some d large enough. So X essentially is the ν-fold product space of $\{0,1\}^d$. On $\tilde{X}$ consider the following class of genetic operators: $G_1^k : \tilde{X} \to \tilde{X}$, $1 \leq k \leq d$ are mutation operators. They work in the following way: Given an individual $\tilde{x} = (\tilde{x_1}, \ldots, \tilde{x_d}) \in \tilde{X}$ (with the $\tilde{x_i} \in \{0,1\}$) the operator G_1^k flips the k'th coordinate into a 0, if it is 1 and into 1, if it is 0. So

$$G_1^k((\tilde{x_1}, \ldots, \tilde{x_k}, \ldots \tilde{x_d})) = (\tilde{x_1}, \ldots, -(\tilde{x_k} - 1), \ldots \tilde{x_d}).$$

Introduce as second class $G_2^k, k = 0, \ldots, d$ of genetic operators. $G_2^k : \tilde{X} \times \tilde{X} \to \tilde{X}$ are crossing over operators. For any two $x, y \in \tilde{X}$ with $x = (\tilde{x_1}, \ldots \tilde{x_d})$ and $y = (\tilde{y_1}, \ldots \tilde{y_d})$ and any $0 \leq k \leq d$ we have

$$G_2^k(x,y) = (\tilde{x_1}, \ldots, \tilde{x_k}, \tilde{y}_{k+1} \ldots \tilde{y_d}).$$

Then the algorithm we have in mind and we have described above is given by X, the set

$$\{G_1^{k_1} \circ \ldots \circ G_1^{k_l}, 1 \leq k_1 < \ldots < k_l \leq d, 1 \leq l \leq d\} \cup \{G_2^k, k = 0, \ldots, d\}$$

of genetic operators and a)-h) above. So if we take probability distributions on the set of genetic operators and on the individuals of each population such that the associated random walk fulfills $\hat{I}$ and $\hat{R}$ above, then this example belongs to the class of Genetic Algorithms we can treat.

Now we are able to formulate our central result concerning the convergence of Genetic Algorithms.

Theorem 2.2 *Consider a Genetic Algorithm Y_t given by a set of ν individuals on a state space $\tilde{X}$, a fitness function $V : \tilde{X} \to \mathbb{R}$, and a set of genetic operators $\{G_1, \ldots, G_\mu\}$ together with a)-h) above. Assume that $(\hat{I})$ and $(\hat{R})$ are fulfilled. Then*

1. *The invariant measure at time t is the Gibbs measure*

$$\varrho_t(x) = \varrho_t((x_1, \ldots, x_\nu)) = \frac{e^{-\beta_t U(x)}}{\sum_{y \in X} e^{-\beta_t U(x)}} \tag{2}$$

   where $U(x) = U((x_1, \ldots, x_\nu)) = \sum_{j=1}^{\nu} V(x_j)$.

2. *As $t \to \infty$ the invariant measure converges to the uniform distribution on the set of minima of U.*

3. *Choose $\beta_t = \frac{1}{m} \log(1 + \frac{2cm}{3M} t)$ for $t \geq 0$ where $M := \max_{y \in X} U(y)$ and m is a positive number (the difficulty of the energy landscape), which we will define precisely in the following section. Then for the Radon-Nikodym density f_t of the distribution of Y_t with respect to ϱ_t we have*

$$||f_t - 1||_{2,t} \leq 1 + e^{-\beta_t M} ||f_0 - 1||_{2,0}.$$

4. *Choose the same cooling schedule $\beta_t = \frac{1}{m} \log(1 + \frac{2cm}{3M} t)$ for $t \geq 0$ and any non empty, proper subset V of X. Then for the first entrance time τ_{V^c} into V^c we have*

$$P(\tau_{V^c} > t) \leq (e^{-\beta_t M} ||f_0 - 1||_{2,0}) \varrho_t(V)^{1/2} + 5 \varrho_t(V).$$

3 A Class of Search Algorithms on Graphs

In this section we translate the Genetic Algorithms into a class of search algorithms on graphs. For those the convergence results of Deuschel and Mazza [DM94] are applicable.
Recall that (X, E) was a finite graph (without loops and multiple edges) and

let $U : X \to \mathbb{R}$ be a real valued function. To simplify notation we assume that U is positive and that $\min_x U(x) = 0$. For each $x \in X$ denote by $\mathcal{N}(x)$ its neighborhood (all vertices which are connected to x by an edge) and assume that $\mathcal{N}(x)$ is partitioned into k_x subsets, i.e. there are sets $N_1^x, \ldots, N_{k_x}^x$ with

$$\cup_{i=1}^{k_x} N_i^x = \mathcal{N}(x)$$

and

$$N_i^x \cap N_j^x = \emptyset \quad \text{for all } i \neq j \in \{1, \ldots, k_x\}$$

for each $x \in X$. Moreover let us assume that the partitions fulfill the following symmetry condition.

$$\begin{aligned} (y \in N_i^x \subset \mathcal{N}(x) \text{ and } x \in N_j^y \subset \mathcal{N}(y)) \\ \text{implies} \\ \{x\} \cup N_i^x = \{y\} \cup N_j^y. \end{aligned} \tag{3}$$

This condition may look somewhat artificial in the context of search problems on graphs, but as it will turn out in Lemma 3.1 this setting is a natural generalization of the class of Genetic Algorithms introduced in Section 2. Moreover, as we will see below it is essential to make the algorithm reasonable, i.e. to equip it with an invariant measure that becomes concentrated on the set of minima of the energy function U.
Furthermore introduce a probability distribution $p_x(\cdot)$ on the sets $N_1^x, \ldots, N_{k_x}^x$ for each $x \in X$. Finally suppose that for every $x \in X$ and every $j \in \{1, \ldots, k_x\}$ we are given a probability distribution π_j^x on $\{x\} \cup N_j^x$ that charges every point. Let us assume that the π_j^x are compatible with the symmetry condition (3), i.e. we assume that

$$\pi_j^x = \pi_k^y \text{ if } \{x\} \cup N_j^x = \{y\} \cup N_k^y.$$

Usually, i.e. in all cases we have in mind, it will be sufficient to choose the π_j^x to be induced by a global probability distribution on X via conditioning on $\{x\} \cup N_j^x$, so e.g. we can think of π_j^x as the uniform distribution on $\{x\} \cup N_j^x$.

Consider the random walk associated with $(p_x(\cdot))_{x \in X}$ and $(\pi_j^x(\cdot))_{j,x}$ by

$$q(x, y) = \sum_{j=1}^{k_x} \delta_y(N_j^x) p_x(N_j^x) \pi_j^x(y) \tag{4}$$

where $\delta_y(\cdot)$ is the Dirac measure concentrated in $y \in X$, and

$$q(x, x) = 1 - \sum_{y \in \mathcal{N}(x)} q(x, y) = \sum_{j=1}^{k_x} p_x(N_j^x) \pi_j^x(x).$$

If $q(x,y) \neq 0$ and $x \neq y$ we will also write

$$q(x,y) = p_x(N^x_{j(y)})\pi^x_{j(y)}(y) \tag{5}$$

where we denote by $j(y)$ the unique index in $\{1,\ldots,k_x\}$ such that $y \in N^x_{j(y)}$. Let us assume that the following condition is true.

(I) The random walk $(q(x,y))_{x,y}$ is irreducible, i.e. for each pair $x,y \in X$ there exists a positive integer N such that $q^N(x,y) > 0$.

Here we define $q^N(x,y)$ inductively by $q^1(x,y) = q(x,y)$ and

$$q^N(x,y) = \sum_{z\in X} q^{N-1}(x,z)q(z,y)$$

for $N \geq 2$.
Let us also introduce the following condition.

(R) The random walk $(q(x,y))_{x,y}$ is reversible with respect to some probability measure $\varrho(\cdot)$ on X. This means

$$\varrho(x)q(x,y) =: \alpha(x,y) = \alpha(y,x) = \varrho(y)q(y,x) \tag{6}$$

for all $x,y \in X$

One may think of $q := (q(x,y))_{x,y}$ as the infinite temperature analogue of the process $q_t := (q_t(x,y))_{x,y}$ given by

$$q_t(x,y) = p_x(N^x_{j(y)})\pi^x_{j(y),t}(y) \quad \text{for } x \neq y \tag{7}$$

and $q_t(x,x) = 1 - \sum_{z\neq x} q_t(x,z)$. Here $\pi^x_{j(y),t}(\cdot)$ denotes the Gibbs measure on $\{x\}\cup N^x_{j(y)}$ with energy U, at inverse temperature β_t and with respect to $\pi^x_{j(y)}(\cdot)$. More formally

$$\pi^x_{j(y),t}(z) := \frac{e^{-\beta_t(U(z)-U(\kappa_{x,y}))}}{Z^x_{j(y),t}}\pi^x_{j(y)}(z) \tag{8}$$

for $z \in \{x\} \cup N^x_{j(y)}$, where

$$Z^x_{j(y),t} = \sum_{z\in\{x\}\cup N^x_{j(y)}} e^{-\beta_t(U(z)-U(\kappa_{x,y}))}\pi^x_{j(y)}(z) \tag{9}$$

and $\kappa_{x,y}$ is given by

$$U(\kappa_{x,y}) = \min_{z\in\{x\}\cup N^x_{j(y)}} U(z).$$

To make the algorithm reasonable it will be sensible to choose β_t monotonely increasing in t. Of course, this cooling of the process has to be carried out quite carefully in order not to get stuck in a local minimum which is not a global one. The exact dependency of β_t from t we work with - for which we, as in SA, reserve the names cooling schedule or freezing schedule - will be determined later.

In the language of functional analysis we have arrived at the following situation. Let Q and Q_t respectively denote the operators associated with the random walks q and q_t respectively, i.e. more precisely $Q = (q(x,y))_{x,y\in X}$ and $Q_t = (q_t(x,y))_{x,y\in X}$. If q fulfills condition (I) then both q and q_t possess unique invariant measures. If moreover q fulfills (R) its invariant measure is given by ϱ. Let ϱ_t denote the invariant measure for the random walk q_t. Then Q and Q_t respectively can be regarded as self-adjoint operators on $L^2(\varrho)$ and $L^2(\varrho_t)$ respectively. More precisely they act on a function $\Phi \in L^2(\varrho_{(t)})$ via

$$[Q\Phi](x) := \sum_{y\in X} \Phi(y)q(x,y)$$

and

$$[Q_t\Phi](x) := \sum_{y\in X} \Phi(y)q_t(x,y)$$

for $x \in X$. Let us also introduce the generator $L_t := Q_t - I$ of the random walk q_t which acts on $L^2(\varrho_t)$ by

$$[L_t\Phi](x) := \sum_{y\in X} (\Phi(y) - \Phi(x))q_t(x,y)$$

for $\Phi \in L^2(\varrho_t)$ and $x \in X$. With these notations we are able to give a formal description of the search process we have in mind. This description is given by the following forward equation for the transition probabilities $P_{s,t}(x,y)$ of the process, i.e. for the probabilities to arrive in y at time t when starting in x at time s, $s \le t$:

$$\frac{\partial}{\partial t} P_{s,t}(x,y) = [L_t^* P_{s,t}(x,\cdot)](y) \text{ for } s < t \tag{10}$$

and

$$P_{s,s}(x,y) = \delta_x(y),$$

where the meaning of (10) is that

$$\frac{\partial}{\partial t}[P_{s,t}\Phi](x) = [P_{s,t}[L_t[\Phi]](x) \text{ for } s < t \tag{11}$$

and

$$[P_{s,s}\Phi](x) = \Phi(x)$$

for $\Phi : X \to \mathbb{R}$ and $x \in X$. Here the operator $[P_{s,t}]$ associated with the transition probabilities $P_{s,t}(x, y)$ is defined by

$$[P_{s,t}\Phi](x) := \sum_{y \in X} \Phi(y) P_{s,t}(x, y).$$

Let us finally reserve the letter Y_t for the process described by (11) at time t.

The intuitive description of the process is that starting in x at time s the process waits an exponential time with mean one and then at time t chooses a subset N_j^x of the neighborhood $\mathcal{N}(x)$ of x according to p_x and then chooses an element y of $N_j^x \cup \{x\}$ according to the Gibbs distribution $\pi_{j,t}^x$ on $N_j^x \cup \{x\}$ and so on. People interested in programming such kind of algorithms on a computer will remark that a model with continuous time scale is only the second best solution and that discrete time would be a more natural choice. Although we commit that this remark contains some truth, we on the other hand reply that the introduction of continuous time does not essentially change the behavior of the algorithm, whereas we gain a lot of techniques that are very well understood, instead. Especially when analyzing the time evolution of the Radon-Nikodym derivative of the distribution of Y_t with respect to ϱ_t we are able to apply techniques for the solution of differential inequalities instead of treating difference inequalities which are harder to solve.

In order to justify the intoduction of the algorithms above, we should check that the Genetic Algorithms of the previous section are covered by them.

Lemma 3.1 *Assume the Genetic Algorithm described by a)-h) in Section 2 fulfills conditions $(\hat{I})$ and $(\hat{R})$, then the process is a random walk on X of the form described above with energy function*

$$U(x) = U((x_1, \ldots, x_\nu)) = \sum_{i=1}^{\nu} V(x_i).$$

Proof: Consider the graph (X, E) associated with the algorithm given by a)-h) in Section 2. This means that $e = (x, y) \in E$ if and only if there exist a genetic operator $G_j \in \{G_1, \ldots, G_\mu\}$ and a subset $(z_1, \ldots, z_{l_j})$ of size l_j of $x = (x_1, \ldots, x_\nu)$ with $G_j(z_1, \ldots, z_{l_j}) = x_0$ such that there exists $0 \le k \le \nu$ with

$$(x_0, x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_\nu) = y = ((y_1, \ldots, y_\nu).$$

Note that the choice of a genetic operator G_j and a subset $(z_1, \ldots, z_{l_j})$ of size l_j of $x = (x_1, \ldots, x_\nu)$ already generates a partition of the neighborhood of x in (X, E). Indeed if $G_j(z_1, \ldots, z_{l_j}) = x_0$, an element of the partition consists of all

$\nu+1$ vectors of size ν that can be build from $(x_0, x_1, \ldots, x_\nu)$. These partitions are symmetric in the sense of definition (3) since two populations of an element of a partition determine this element uniquely.
Moreover the random walk on X described by a)-h) can be split into one part consisting of the choice of a genetic operator G_j and l_j individuals (on which it acts) and a second part in which the next generation is selected, i.e. we have a part of first choosing a component of the neighborhood of x and then choosing a new vertex in this component with probability induced by γ_t. So we have to check that this probability is a Gibbs distribution on the component with energy U. But this follows immediately from the fact that the probability of choosing a population $(x_0, x_1, \ldots, x_{k-1}, x_{k+1}, \ldots, x_\nu)$ from the component $(x_0, x_1, \ldots, x_\nu)$ is given by $\gamma_t(x_k)$ and that

$$\begin{aligned}\gamma_t(x_k) &= \frac{e^{\beta_t V(x_k)}}{\sum_{i=0}^{\nu} e^{\beta_t V(x_i)}} \\ &= \frac{e^{-\beta_t \sum_{\substack{i=0\\ i\neq k}}^{\nu} V(x_i)}}{\sum_{j=0}^{\nu} e^{-\beta_t \sum_{\substack{i=0\\ i\neq j}}^{\nu} V(x_i)}}.\end{aligned}$$

Finally observe that conditions (I) and (R) follow directly from conditions $(\hat{I})$ and $(\hat{R})$. □

In the following section we will study under which conditions the Markov chain Y_t follows its invariant measure and in case it does, how fast the convergence is. As already mentioned we can apply the results of Deuschel and Mazza [DM94] and the main tools there are differential inequalities for the Radon-Nikodym density of the distribution of Y_t with respect to ϱ_t. The derivation of these differential inequalities is essentially based on the following relation between the second smallest eigenvalue λ_t of $-L_t$ (the smallest eigenvalue being 0) and the Dirichlet form associated with L_t. This relation is given by

$$\lambda_t = \inf\{\frac{\mathcal{E}_t(\Phi,\Phi)}{Var_t(\Phi)} : \Phi : X \to \mathbb{R}, \gneqq \text{non constant}\}. \tag{12}$$

Here

$$\mathcal{E}_t(\Phi,\Phi) := - < \Phi, [L_t\Phi] >_t = \frac{1}{2}\sum_{x,y\in X}(\Phi(x)-\Phi(y))^2\alpha_t(x,y)$$

is the discrete Dirichlet form associated with $-L_t$, $< \cdot,\cdot >_t$ denotes the scalar product in $L^2(\varrho_t)$, $Var_t(\Phi)$ is the variance of a function Φ with respect to ϱ_t and α_t is given by

$$\alpha_t(x,y) := \varrho_t(x)q_t(x,y) = \alpha_t(y,x).$$

Note that $\mathcal{E}_t$ is related to the Dirichlet form $\mathcal{E}_0$ given by

$$\mathcal{E}_0(\Phi, \Phi) := \frac{1}{2} \sum_{x,y \in X} (\Phi(x) - \Phi(y))^2 \alpha(x,y)$$

via the identity

$$\mathcal{E}_t(\Phi, \Phi) = \frac{1}{2} \sum_{x,y \in X} (\Phi(x) - \Phi(y))^2 \alpha(x,y) \frac{\exp(-\beta_t(U(x) + U(y) - U(\kappa_{x,y}))}{Z_t \, Z^x_{j(y),t}}. \tag{13}$$

The relation (12) is derived by the well known variational characterization of eigenvalues (see e.g. [HJ85]) using the fact that the eigenspace of the eigenvalue 0 is one-dimensional with the constants as the only eigenfunctions.

4 Proofs

In this section we are going to give the proofs of Theorem 2.1. Due to Lemma 3.1 above it suffices to prove the corresponding facts for the search algorithm described in Section 3. So throughout this section we will assume that the process we have in mind is given by (11) and fulfills (I) and (R).
Proceeding step by step we first of all check that the invariant measure is a Gibbs measure.

Lemma 4.1 *Let conditions (I) and (R) be fulfilled. Then the invariant measure ϱ_t of the random walk q_t is the Gibbs measure with energy U and at inverse temperature β_t with respect to ϱ, i.e.*

$$\varrho_t(x) = \frac{e^{-\beta_t U(x)}}{Z_t} \varrho(x)$$

for $x \in X$ where

$$Z_t := \sum_{y \in X} e^{-\beta_t U(y)} \varrho(y).$$

Proof: We have to check that

$$\sum_{x \in X} \varrho_t(x) q_t(x,y) = \varrho_t(y).$$

Since $q_t(y,y) = 1 - \sum_{y \neq x} q_t(y,x)$ it suffices to check that

$$\varrho_t(y) q_t(y,x) = \varrho_t(x) q_t(x,y)$$

i.e. the ϱ_t-reversibility of q_t. But this fact follows from the definitions of ϱ_t and q_t together with ϱ-reversibility of q and the symmetry condition (3) for the partition $(N_j^x)_{j,x}$ (note that (3) especially implies $Z^x_{j(y),t} = Z^y_{j(x),t}$ for each pair $x, y \in X$ and each t). $\square$

Remark 4.2 *It should be emphasized that property (3) is not only essential to equip the Markov process q_t with an invariant measure which is Gibbs with energy U, but even more important to guarantee that the invariant measure of q_t becomes concentrated on the right set, which is the set of minima of U, at all. As a matter of fact using the Matrix-Tree-Theorem by Bott and Mayberry (see e.g. [FW84]) to describe the invariant measure one is able to give an example of a graph on four vertices where the absence of condition (3) leads to an invariant measure that becomes concentrated in a suboptimal point.*

Now we are going to give a cooling schedule for which the associated process (as described in the previous section) converges. Motivated by the fact that for SA a good cooling schedule is a function of the spectral gap of the associated operator we first of all discuss the right order of magnitude for the spectral gap λ_t (see (12)) of the operator $-L_t$. In the case of SA the size of the spectral gap is governed by the largest hill to be climbed to reach the global minimum (a more precise definition will be given below).

Using similar arguments in our case leads to the following definitions. For any two vertices $x, y \in X$ let us define a path leading from x to y to be any sequence of points $x = x_1, \dots, x_k = y, k \in \mathbb{N}$ with $q_t(x_{i-1}, x_i) > 0, i = 2, \dots, k$. Let $\Gamma_{x,y}$ be the set of all paths leading from x to y. Define for a path $\gamma \in \Gamma_{x,y}$ $(x, y \in X)$ the function $e(\gamma)$ by

$$e(\gamma) := \max\{U(x_i) + U(x_{i+1}) - U(\kappa_{x_i,x_{i+1}}), x_i, x_{i+1} \in \gamma\}$$

where the enumeration of the x_i follows the natural way of the path from x to y. Denote by $h(x, y)$ the minimum $e(\cdot)$ we have to face to get from x to y, i.e.

$$h(x, y) := \min_{\gamma \in \Gamma_{x,y}} e(\gamma) \qquad \text{for } x \neq y$$

and

$$h(x, x) = 0.$$

Finally put

$$m = \max_{x,y \in X} (h(x, y) - U(x) - U(y)).$$

Note that this definition of m differs from that usually given in the context of SA (see e.g. [Ha88] or [HS88]) where the interesting quantity m^* is the largest hill to be climbed to reach to global minimum. More exactly one puts for every path

$$e^*(\gamma) := \max\{U(x_i), x_i \in \gamma\}$$

and chooses $h^*(x, y)$ to be the maximum elevation to be climbed to get from x to y, i.e.

$$h^*(x, y) := \min_{\gamma \in \Gamma_{x,y}} e^*(\gamma).$$

m^* is then defined by

$$m^* = \max_{x,y \in X} (h^*(x, y) - U(x) - U(y))$$

and we have the relation

$$m^* \leq m \leq m^* + M,$$

(where M was the global maximum of U). Note that this especially implies that $m \geq 0$, since $m^* \geq 0$.
With these definitions the order of magnitude of λ_t is given by the following theorem the proof of which follows the ideas of the proof of Theorem 2.1 in [HS88] and Theorem 2.1 in [L94].

Theorem 4.3 *There exist constants* $0 < c \leq C < \infty$ *such that for all* $t \geq 0$

$$ce^{-\beta_t m} \leq \lambda_t \leq Ce^{-\beta_t m} \tag{14}$$

As in [HS88] we split the proof into two lemmas.

Lemma 4.4 *There exists a constant* $C < \infty$ *such that for all* $t \geq 0$

$$\lambda_t \leq Ce^{-\beta_t m}.$$

Proof: First we analyze the case $m > 0$.
In view of (12) all we have to prove is the existence of a function $\phi : X \to \mathbb{R}$ and a constant $C < \infty$ such that for all $t \geq 0$

$$\frac{-\sum_{x \in X} \phi(x)[L_t\phi](x)\varrho_t(x)}{Var_t(\phi)} \leq e^{-\beta_t m}.$$

By definition of m there exist two vertices, x_0 and y_0 such that

$$m = h(x_0, y_0) - U(x_0) - U(y_0).$$

Define

$$A = \{z : h(y_0, z) < h(y_0, x_0)\}$$

and

$$\phi(x) := \delta_A(x)$$

the indicator function of the set A. Observe that clearly $x_0 \notin A$. Moreover h is symmetric ($h(x,y) = h(y,x)$) which follows by reversing the paths from x to y (which is possible due to (R)) and using the symmetry condition of the partitions (1). Hence due to $m > 0$ we have that $y_0 \in A$.

Next notice that if $z \in A$ and $z' \notin A$ and $q(z,z') > 0$, then $U(z) + U(z') - U(\kappa_{z,z'}) \geq h(x_0, y_0)$ which can be seen by considering a path γ from y_0 to z with $e(\gamma) = h(y_0, z)$ and extending it by an edge connecting z with z'. Therefore for any two x, y with $x \in A$ and $y \notin A$

$$\begin{aligned} \varrho_t(x) q_t(x,y) &= \varrho_t(y) q_t(y,x) = \frac{e^{-\beta_t(U(x)+U(y)-U(\kappa_{x,y}))}}{Z_t Z^x_{j(y),t}} \alpha(x,y) \\ &\leq \frac{e^{-\beta_t h(x_0,y_0)}}{Z_t Z^x_{j(y),t}} \alpha(x,y). \end{aligned}$$

Thus for ϕ defined as above

$$\begin{aligned} -\sum_{x \in X} \phi(x)[L_t \phi](x) \varrho_t(x) &= \frac{1}{2} \sum_{x,y \in X} q_t(x,y)(\phi(x) - \phi(y))^2 \varrho_t(x) \\ &\leq \sum_{x \in A} \sum_{y \notin A} \frac{e^{-\beta_t h(x_0,y_0)}}{Z_t Z^x_{j(y),t}} \alpha(x,y). \end{aligned}$$

On the other hand

$$Var_t(\phi) = \varrho_t(A) \varrho_t(A^c) \geq \varrho(x_0) \varrho(y_0) \frac{e^{-\beta_t(U(x_0)+U(y_0))}}{(Z_t)^2}.$$

Since $Z_t \leq 1$, and $\sum_{x \in A} \sum_{y \notin A} \alpha(x,y) \leq 1$ for all $x, y \in X$, and moreover ϱ is bounded away from 0 as well as $Z^x_{j(y),t}$ is bounded away from 0 uniformly in $x, y \in X$ (the latter by $\min_x \min_{j=1,\ldots,k_x} \min_{z \in N^x_j} \pi^x_j(z)$) we obtain that there exists a constant $C < \infty$ such that

$$\lambda_t \leq C e^{-\beta_t m}.$$

If $m = 0$ then also $m^* = 0$. In this case, similar to the work of Holley and Stroock [HS88], Proof of Lemma 2.3, the same proof as above can be performed with the set A replaced by

$$A' := \{x : U(x) = M\}$$

where $M = \max_{y\in X} U(y)$. □

Lemma 4.5 *There exists a constant $c > 0$ such that for all $t \geq 0$*

$$\lambda_t \geq ce^{-\beta_t m}$$

Proof: The proof of the lower bound follows the lines of the proof of Lemma 2.7 in [HS88]. By (12) it suffices to show the existence of a constant c, such that for all t, and for all $f : X \to \mathbb{R}$, f non constant

$$\frac{\mathcal{E}_t(f,f)}{Var_t(f)} = \frac{-\sum_{x\in X} f(x)[L_t f](x)\varrho_t(x)}{Var_t(f)} \geq ce^{-\beta_t m}. \tag{15}$$

This is done by bounding the variance as follows :
For any two $x, y \in X$ take a path $\gamma^{x,y} \in \Gamma_{x,y}$ from x to y with vertices $\gamma^{x,y}(i), i \geq 1$ such that $e(\gamma^{x,y}) = h(x,y)$. Let $n(x,y)$ be the length of $\gamma^{x,y}$ and put

$$N := \max_{x,y\in X} n(x,y).$$

We introduce the indicator

$$\chi_{z,w}(x,y) = \begin{cases} 1 & \text{if } \exists 1 \leq i \leq n(x,y)-1 : \gamma^{x,y}(i) = z \text{ and } \gamma^{x,y}(i+1) = w \\ 0 & \text{otherwise.} \end{cases}$$

In the following we put $\frac{\chi_{z,w}(x,y)}{\alpha(x,y)} = 0$ if $\chi_{z,w}(x,y) = 0$. Then for any $f : X \to \mathbb{R}$

$$\begin{aligned} 2Var_t(f) &= \sum_{x,y\in X} (f(x) - f(y))^2 \varrho_t(x)\varrho_t(y) \\ &= \sum_{x,y\in X} \left(\sum_{i=2}^{n(x,y)} f(\gamma^{x,y}(i)) - f(\gamma^{x,y}(i-1)) \right)^2 \varrho_t(x)\varrho_t(y) \\ &\leq \sum_{x,y\in X} n(x,y) \sum_{i=2}^{n(x,y)} \left(f(\gamma^{x,y}(i)) - f(\gamma^{x,y}(i-1))\right)^2 \varrho_t(x)\varrho_t(y) \end{aligned}$$

by Jensen's inequality. Therefore

$$\begin{aligned} 2Var_t(f) &\leq N \sum_{\substack{x,y,z,w\\ \in X}} \chi_{z,w}(x,y)(f(z)-f(w))^2 e^{-\beta_t(U(z)+U(w)-U(\kappa_{z,w}))}\alpha(z,w) \\ &\qquad \frac{\varrho_t(x)\varrho_t(y)}{e^{-\beta_t(U(z)+U(w)-U(\kappa_{z,w}))}\alpha(z,w)} \\ &\leq N[\max_{z,w\in X} \sum_{x,y\in X} \chi_{z,w}(x,y) \frac{\varrho_t(x)\varrho_t(y) Z_t Z^z_{j(w),t}}{e^{-\beta_t(U(z)+U(w)-U(\kappa_{z,w}))}\alpha(z,w)}] \\ &\qquad [\sum_{z,w\in X} (f(z)-f(w))^2 \frac{e^{-\beta_t(U(z)+U(w)-U(\kappa_{z,w}))}\alpha(z,w)}{Z_t Z^z_{j(w),t}}]. \end{aligned} \tag{16}$$

Since the second term in brackets in (16) is twice the Dirichlet form $\mathcal{E}_t(f,f)$ (due to (13)) it suffices to calculate that

$$\begin{aligned} & \chi_{z,w}(x,y)\frac{\varrho_t(x)\varrho_t(y)Z_tZ^z_{j(w),t}}{e^{-\beta_t(U(z)+U(w)-U(\kappa_{z,w}))}\alpha(z,w)} \\ = \quad & \frac{\chi_{z,w}(x,y)}{\alpha(z,w)}\frac{\varrho(x)\varrho(y)}{Z_t}e^{\beta_t(U(z)+U(w)-U(\kappa_{z,w})-U(x)-U(y))}Z^z_{j(w),t} \\ \leq \quad & e^{\beta_t m}[\frac{\chi_{z,w}(x,y)}{\alpha(z,w)}\frac{\varrho(x)\varrho(y)}{\sum_{v:U(v)=0}\varrho(v)}], \end{aligned}$$

since $Z^z_{j(w),t} \leq 1$ for all z,w,t and by definition of m and the special choice of the paths $\gamma^{x,y}$ we have $m \geq (U(z)+U(w)-U(\kappa_{z,w})-U(x)-U(y))$ whenever $\chi_{z,w}(x,y)=1$ for all $x,y,z,w \in X$. This proves the lower bound. □

With these results on the order of magnitude of the spectral gap we are able to give a "good" cooling schedule and to analyze the speed of convergence for the process associated with this cooling schedule.

Theorem 4.6 *With the definitions introduced above, choose a logarithmic cooling schedule of the form*

$$\beta_t = \frac{1}{m}\log(1+\frac{2cm}{3M}t) \tag{17}$$

where $M = \max_{x\in X} U(x)$. Let Y_t be the process starting in $x \in X$ defined by (11) and assume that if fulfills conditions (I) and (R). Let f_t denote the Radon-Nikodym density of the distribution of Y_t with respect to its invariant measure ϱ_t. Then

$$||f_t - 1||_{2,t} \leq 1 + e^{-M\beta_t}||f_0 - 1||_{2,0}, \tag{18}$$

where for each $t \geq 0$ $||\cdot||_{2,t}$ denotes the norm in $L^2(\varrho_t)$.

Proof: Check that we are in the setting of [DM94] and apply Lemma 2.2.3 formula (2.2.5).

□

Corollary 4.7 *Let $\emptyset \neq V \subset X$ be a proper subset of X. With the settings of the preceding theorem denote by τ_{V^c} the first entrance time of the process Y_t into V^c. Then for any starting point $x \in X$*

$$P(\tau_{V^c} > t) \leq \varrho_t(V) + \varrho_t(V)^{1/2}(e^{-M\beta_t}(\frac{1-\varrho(x)}{\varrho(x)})^{\frac{1}{2}} + 1).$$

Proof: We have

$$P(\tau_{V^c} > t) \leq P(Y_t \in V).$$

Moreover

$$\begin{aligned} (P(Y_t \in V) - \varrho_t(V))^2 &= (\sum_{y \in X} \delta_V(y)(f_t(y) - 1)\varrho_t(y))^2 \\ &\leq ||f_t - 1||_{2,t}^2 \varrho_t(V) \end{aligned}$$

by the Cauchy-Schwartz inequality. Hence

$$\begin{aligned} P(Y_t \in V) &\leq \varrho_t(V) + \varrho_t(V)^{1/2}||f_t - 1||_{2,t} \\ &\leq \varrho_t(V) + \varrho_t(V)^{1/2}(||f_0 - 1||_{2,0}e^{-M\beta_t} + 1) \\ &= \varrho_t(V) + \varrho_t(V)^{1/2}((\frac{1 - \varrho(x)}{\varrho(x)})^{1/2}e^{-M\beta_t} + 1) \end{aligned}$$

by calculation of $||f_0 - 1||_{2,0}$. □

It is noteworthy that the bound obtained in the corollary above is still improvable for sets of high energy, i.e. for sets consisting of vertices with large U-values.

Theorem 4.8 *In the situation and with the notations of Corollary 4.7 for any proper subset V of X such that $V \neq \emptyset$ and any starting point $x \in X$ of the process Y_t*

$$P(\tau_{V^c} > t) \leq \varrho_t(V)^{1/2}e^{-M\beta_t}(\frac{1 - \varrho(x)}{\varrho(x)})^{\frac{1}{2}} + 5\varrho_t(V). \tag{19}$$

Proof: The proof is immediate from Lemma 2.2.3 formula (2.2.7) and Corollary 2.2.8 in [DM94], if one modifies the setting as in [G92] in the following way: Let us replace V^c by a single x_0. So we define a modified Markov chain q_t^* on the state space $X^* := V \cup \{x_0\}$ by

$$\begin{aligned} q_t^*(x, y) &= q_t(x, y) \qquad \text{if } x, y \in V \\ q_t^*(x, x_0) &= \sum_{y \in V^c} q_t(x, y) \quad \text{if } x \in V \\ q_t^*(x_0, x) &= \frac{\varrho_t(x)}{\varrho_t(V^c)} q_t^*(x, x_0) \quad \text{if } x \in V \\ q_t^*(x_0, x_0) &= 1 - \sum_{x \in V} q_t^*(x_0, x). \end{aligned}$$

This yields that the distribution given by $\varrho_t^*(x) := \varrho_t(x)$ for $x \in V$ and $\varrho_t^*(x_0) = \varrho_t(V^c)$ is the stationary distribution for the random walk q_t^*.

In a similar way we may define the initial random walk q^* as a modification of q. q^* then is reversible with respect to a probability distribution ϱ^*.
If Y_t^* denotes the process described by the Fokker-Planck equation (8) when substituting q_t by q_t^* (and the operators associated with q_t accordingly), and $\tau^*_{V^c}$ denotes the first hitting time of $x_0 \in X^*$ of the process Y_t^*, we have $\tau_{V^c} = \tau^*_{V^c}$ for any starting point $x \in X$ and therefore

$$P(\tau_{V^c} > t) = P(\tau^*_{V^c} > t)$$

for any starting point $x \in X$ of the original process.
By considering the operators induced by q^* we obtain that the Dirichlet form $\mathcal{E}_t^*$ associated with q_t^* is nothing but the original Dirichlet form $\mathcal{E}_t$ restricted to those functions which are constant on V^c. So the spectral gap estimate of Lemma 3.3 implies the same estimate for $\mathcal{E}_t^*$:

$$\mathcal{E}_t^*(f,f) \geq ce^{-\beta_t m} Var_t^*(f).$$

□

Let us conclude this section with a few remarks on the results we have obtained.

Remarks 4.9 *1) Although the convergence properties of the algorithm are formally similar to those of a corresponding SA algorithm, i.e. a SA algorithm run on the same graph, our estimates suggest that the convergence to the global minimum should be somewhat slower than the convergence of SA to the global minimum is. This can be seen from the fact that the spectral gap of our process is smaller than the spectral gap of the corresponding SA process. In the convergence results Corollary 4.7 and Theorem 4.8 this property is reflected in the form of the cooling schedule: Our process cools down at a slower rate than SA does, or, equivalently, the stationary Gibbs measure for our process puts less mass on the set of global minima of U than the stationary Gibbs measure for the SA process at the same time does.*

2) Although Theorem 4.2 gives the right order of magnitude of the spectral gap of the operator L_t, this knowledge is not very applicable basically due to two reasons. First of all the central convergence results Theorem 4.8 only would need a bound of the form

$$\inf\{\frac{\mathcal{E}_t(f,f)}{Var_t(f)}, f \textit{ non constant, } f|_{V^c} \textit{ constant}\} \geq \tilde{c}e^{-\beta_t \tilde{m}} \tag{20}$$

for some modified constants $\tilde{m}$ and $\tilde{c}$. Estimates of this form are of course weaker than Theorem 4.2 and hence our bounds are not sharp for the estimates we would need.

Even more important is that the way to compute m is not applicable at all. To this end we would have to evaluate the function $e(\cdot)$ over the set of all paths in (X, E). Since this set is much larger than the set X itself, the problem to find the precise value of m is even harder than the original one.
Other methods to bound the spectral gap of the operator L_t lead to the so called Cheeger type and Poincaré type bounds as introduced in [DS91], [G92] and [DM94].

References

[C90] O. Cantoni, Asymptotics of simulated annealing algorithms, Thèse de Doctorat, Université Paris-Sud, Orsey (1990)

[Ce94] R. Cérf, Une Théorie Asymptotique des Algorithmes Génétique, Thèse de Doctorat, Université Montpellier II (1994)

[D88] L. Davis, Genetic Algorithms and Simulated Annealing, Kaufmann, Los Altos, Calif. (1988)

[DM94] J.D. Deuschel, C. Mazza, L^2 Convergence of Time Non-Homogeneous Markov Processes I: Spectral Estimates, Ann. Appl. Prob. 4, 1012-1056 (1994)

[DS91] P. Diaconis, D. Stroock, Geometric Bounds for Eigenvalues of Markov chains, Ann. Appl. Prob. 1, 36-61 (1991)

[F91] J.A. Fill, Eigenvalue bounds on Convergence to stationarity for non-reversible Markov chains, with an application to exclusion processes, Ann. Appl. Prob. 1, 62-87 (1991)

[FW84] M.I. Freidlin, A.D. Vencel, Random Pertubations of Dynamical Systems, Springer, Grundlehren d. math. Wissenschaften 260 (1984)

[GG84] S. Geman, D. Geman , Stochastic Relaxation, Gibbs Distribution, and the Bayesian Restoration of images, IEEE Trans. on Pattern Anal. and Machine Intelligence, PAMI-6, 721-741 (1984)

[Gi85] B. Gidas Nonstationary chains and the convergence of the annealing algorithm, J. Stat. Phys. 39, 73-131 (1985)

[Go89] D. Goldberg, Genetic Algorithms in search, optimization and machine learning, Reading, Mass., Addison-Wesley (1989)

[G92] F. Götze, Rate of convergence for simulated annealing processes, Preprint 92-008, SFB 343, Bielefeld (1992)

[Ha88] B. Hajek, Cooling Schedules for optimal Annealing, Math. of Operations Research 13, 311-329 (1988)

[H75] J. Holland, Adaption in Natural and Artificial Systems, Ann Arbor, University of Michigan Press (1975)

[HS88] R. Holley, D. Stroock, Simulated Annealing via Sobolev inequalities , Comm. Math. Phys. 115, 553-569 (1988)

[H89] R. Holley, S. Kusuoka, D. Stroock, Asymptotics of the Spectral Gap with Applications to the theory of Simulated Annealing, J. Funct. Anal. 83, 333-347 (1989)

[HJ85] R. Horn, C. Johnson, Matrix analysis, Camb. Univ. Press (1985)

[K83] S. Kirkpatrick, C. Gelatt, M.P. Vecchi, Optimization by Simulated Annealing, Science 220, 621-680 (1983)

[L94] M. Löwe, Simulated Annealing with Time dependent Energy Function via Sobolev Inequalities, Preprint 94-069, SFB 343 (1994), submitted to Stochastic Processes and Applications

[M85] D. Mitra, F. Romeo, A. Sangiovanni-Vincentelli, Convergence and finite time behavior of simulated annealing, Proc. 24th Conf. Dec. a. Control, 761-767 (1985)

[T93] A. Trouvé, Parallélisation massive du recuit simulé, Thèse de Doctorat, Université Paris XI (1993)

[V93] R.V.V. Vidal (Ed.), Applied Simulated Annealing, Lecture Notes in Economics and Math. Syst. 369, Springer, Berlin (1993)