T-79.300 Postgraduate Course in Theoretical Computer Science
Landscape theory, spring 2002

Dynamics on Landscapes

Hannu Rummukainen

1 Overview

As an aside to the more developed aspects of landscape theory, in Section 3 we survey the landscape properties that have been used to prove convergence of simulated annealing. Additionally, in Section 4 some results on the behaviour of local optimization on a few different random landscape classes are reviewed.

2 Definitions

We consider the combinatorial optimization problem of minimizing a cost function $f : S \to \mathbb{R}$ defined on a finite set S. The set S is called a *configuration space*, and its elements are *configurations*. We denote the set $\{x \in S : f(x) = \min_{y \in S} f(y)\}$ of global minima by S^*.

Let each configuration $x \in S$ be assigned a *neighbourhood* $\mathcal{N}(x) \subset S$ with $x \notin \mathcal{N}(x)$. The *neighbourhood graph* is defined on the vertex set S by adding the arcs (x, y) for each $x \in S$ and $y \in \mathcal{N}(x)$. The neighbourhood graph is assumed to be connected, and usually we also assume that $y \in \mathcal{N}(x)$ implies $x \in \mathcal{N}(y)$, so that the neighbourhood graph can be treated as undirected.

3 Simulated annealing

Simulated annealing is a generic combinatorial optimization approach based on an analogy with statistical mechanics, first introduced by Kirkpatrick et al. [6]. Simulated annealing has been popular among practitioners due to its simple implementation and among theoreticians due to its amenability to theoretical analysis.

Let us consider the generic simulated annealing algorithm defined in Figure 1. The temperature sequence $\{T_k\}_{k=0}^{\infty}$ is considered as input to the algorithm in the definition, but in practice be it would be computed step by step in the loop. The name temperature originates from the analogy with statistical mechanics, and accordingly we require that $T_k \geq 0$ for all k. The temperature sequence is usually referred to as *temperature schedule* or as *cooling schedule* when the temperature approaches 0 as $k \to \infty$. The algorithm is presented without specifying the stop condition; for now it suffices to assume that the algorithm continues indefinitely, although finite-time behaviour will be discussed later on.

The presented algorithm is essentially the same as originally proposed by Kirkpatrick et al. [6]. In the literature [12, 1], variants of the algorithm with different acceptance probabilities and with different neighbour generation probabilities have been considered. However, provided that the neighbour generation probabilities are symmetric, by a natural measure of convergence speed all these variants are inferior to the form of the algorithm presented here [12, Proposition 3.2].

Given an initial state $x_0 \in S$ and a temperature sequence $\{T_k\}_{k=0}^{\infty}$.
$k := 0$.
Repeat
    Generate a random $y \in \mathcal{N}(x)$ by uniform probabilities.
    Generate a random number $r \sim \mathrm{U}(0,1)$.
    If $r < \exp\left(-(f(y) - f(x_k))/T_k\right)$
        $x_{k+1} := y$.
    Else
        $x_{k+1} := x_k$.
    End if
    $k := k + 1$.
Until STOP.

Figure 1: Pseudocode for a simulated annealing algorithm.

3.1 Markov chains

The analysis of simulated annealing is based on the theory of Markov chains [4]. We present here the most basic definitions and a convergence theorem.

Definition 1. *A* Markov chain *on the state space Ω is a sequence of random variables $X_k \in \Omega$, $k = 0, 1, \ldots$ that satisfies*

$$\mathrm{P}[X_k = x_k \mid X_0 = x_0, \ldots, X_{k-1} = x_{k-1}] = \mathrm{P}[X_k = x_k \mid X_{k-1} = x_{k-1}] \tag{1}$$

for all $k = 1, 2, \ldots$ and all $x_0, x_1, \ldots, x_k \in \Omega$.

We shall assume that Ω is finite. Condition (1) is called the Markov property, and it essentially says that the future development of the process only depends on its current state, not on any of the previous states. The state-dependent *transition probabilities* $P_{xy}(k) := \mathrm{P}[X_{k+1} = y \,|\, X_k = x]$ form a *transition matrix* $P(k)$ for each k. A Markov chain is called *homogenous* if the transition matrix $P(k)$ does not depend on k. Otherwise, the chain is called *inhomogenous*.

A homogenous Markov chain is called *irreducible* if for any $x, y \in \Omega$ there is a positive probability of reaching y from x in a finite number of trials. A homogenous Markov chain is called *aperiodic* if for every state $x \in \Omega$ the greatest common divisor $\gcd(\mathcal{D}_x) = 1$, where $\mathcal{D}_x$ is the set of all integers $n > 0$ with positive probability of returning to state x on step n when starting from x on step 0.

Theorem 1. *Let $P = (p_{xy})_{x,y \in \Omega}$ be the transition matrix associated with a homogenous Markov chain on state space Ω, and suppose that the Markov chain is both irreducible and aperiodic. Then there exists a unique stationary distribution $\pi_x := \lim_{k \to \infty} \mathrm{P}[X_k = x]$, $x \in \Omega$, which is uniquely determined by the equations*

$$\sum_{y \in \Omega} \pi_y p_{yx} = \pi_x \quad \textit{for all } x \in \Omega \tag{2}$$

and

$$\sum_{x \in \Omega} \pi_x = 1. \tag{3}$$

It should be clear that the stationary distribution, considered as a vector in $\mathbb{R}^{|\Omega|}$, is a left eigenvector of the transition matrix P.

3.2 Simulated annealing as a Markov chain

Let us consider the initial state x_0 of the simulated annealing algorithm as a random variable on the state space S. Observe that for a fixed initial state distribution and temperature schedule, the state sequence $x_0, x_1, \ldots$ of the simulated annealing algorithm forms a Markov chain on the configuration space S.

The transition probabilities of the process are given by

$$P_{xy}(k) = \begin{cases} 0 & \text{if } y \notin \mathcal{N}(x) \cup \{x\}, \\ \frac{1}{|\mathcal{N}(x)|} \exp\left(-\frac{f(y)-f(x)}{T_k}\right) & \text{if } y \in \mathcal{N}(x)\,, \\ 1 - \frac{1}{|\mathcal{N}(x)|} \sum_{z \in \mathcal{N}(x)} \exp\left(-\frac{f(z)-f(x)}{T_k}\right) & \text{if } y = x. \end{cases} \tag{4}$$

As a consequence of Theorem 1, it is quite easy to show that if the neighbourhood relation is symmetric and the neighbourhood graph is connected, then using a constant temperature $T > 0$ the simulated annealing process converges to a unique stationary distribution $(\pi_x(T))_{x \in S}$ given by

$$\pi_x(T) = \frac{\exp(-f(x)/T)}{\sum_{y \in S} exp(-f(y)/T)} \quad \text{for } x \in S. \tag{5}$$

As T approaches 0, the above stationary distribution approaches the limit distribution $(\pi^*_x)_{x \in S}$ defined as $\pi^*_x = 1/|S^*|$ for $x \in S^*$, and $\pi^*_x = 0$ for $x \in S \setminus S^*$. This is the basis for the convergence to S^* of the simulated annealing process. Of course, the stationary distribution of the current temperature cannot actually be reached before T is decreased. Nevertheless, if the temperature is decreased sufficiently slowly, the process will get "close enough" to the stationary distribution of each temperature step, and eventually the state distribution will converge to $(\pi^*_x)_{x \in S}$. The theory of inhomogenous Markov chains provides the necessary tools for this approach, as sketched in both [1] and [12]; unfortunately the proofs are rather technical, and moreover, the results are not as comprehensive as the theorem discussed in the following Subsection.

3.3 Hajek's convergence theorem

The convergence behaviour of the generic simulated annealing algorithm for cooling temperature schedules under weak conditions has been completely characterized by Hajek [3]. To present the convergence theorem, we first need to define a number of concepts.

Definition 2. *A state $y \in S$ is* reachable at height h *from state $x \in S$ if $x = y$ and $f(x) \leq h$, or if there is a sequence of states $x = x_0, x_1, \ldots, x_n = y$ for some n such that $x_{k+1} \in \mathcal{N}(x_k)$ for $k = 0, 1, \ldots, n-1$ and $f(x_k) \leq h$ for $k = 0, 1, \ldots, n$.*

Definition 3. *A simulated annealing process with a fixed cost function is* weakly reversible *if for any $h \in \mathbb{R}$ and any two states $x, y \in S$, x is reachable at height h from y if and only if y is reachable at height h from x.*

Note that if the neighbourhood relation is symmetric, weak reversibility follows immediately.

Definition 4. *A set $C \subseteq S$ is a* cup *if there is an $h \in \mathbb{R}$ such that for every $x \in C$, it holds that $C = \{y \in S : y$ is reachable at height h from $x\}$.*

Definition 5. *The depth $d(C)$ of a cup C is defined as*

$$d(C) = \min\{f(y) : y \notin C \text{ and } y \in \mathcal{N}(x) \text{ for some } x \in C\} - \min_{x \in C} f(x) \tag{6}$$

for $C \neq S$, and as $+\infty$ for $C = S$.

Definition 6. *The* bottom *of a cup C is the set $\{x \in C : f(x) = \min_{y \in C} f(y)\}$.*

Notice that the bottom of an arbitrary cup is always a set of local minima (which may be weak, ie. have equal-cost neighbours). The concept of a cup is related to the concept of a basin of attraction of a local optimum: Generally, the basin of attraction of a local minimum contains one or more cups as subsets, and these cups share a common bottom set that contains the local minimum (as the only element, if the local minimum has no equal-cost neighbours).

Definition 7. *The* depth of a local minimum x *is the smallest* $d \in \mathbb{R}$ *such that some state* $y \in S$ *with* $f(y) < f(x)$ *can be reached from* x *at height* $f(x) + d$*, or* $+\infty$ *if no such* y *exists.*

The depth of a local minimum can be thought of as the relative height of the lowest hill that must be overcome to find a state of better cost. Intuitively it is clear that the depth of a local minimum is related to the probability that a simulated annealing process can escape the local minimum, and the relation is made exact in the following theorem.

Theorem 2 (Hajek). *Let the temperature schedule* $\{T_k\}_{k=0}^{\infty}$ *be strictly positive, nonincreasing and satisfy* $\lim_{k\to\infty} T_k = 0$*. Suppose that weak reversibility holds.*
Then

1. *For any state* x *that is not a local minimum,* $\lim_{k\to\infty} \mathrm{P}[X_k = x] = 0$*.*
2. *Suppose that the set of states* B *is the bottom of a cup of depth* d *and that the states in* B *are local minima of depth* d*. Then* $\lim_{k\to\infty} \mathrm{P}[X_k \in B] = 0$ *if and only if*

$$\sum_{k=1}^{\infty} \exp(-d/T_k) = \infty. \tag{7}$$

3. *(Consequence of 1 and 2.) Let* D *be the maximum of the depths of all states which are local but not global minima, and* S^* *the set of global minima. Then*

$$\lim_{k\to\infty} \mathrm{P}[X_k \in S^*] = 1 \tag{8}$$

if and only if

$$\sum_{k=0}^{\infty} \exp(-D/T_k) = \infty. \tag{9}$$

See [3] for the proof. The Theorem immediately yields a foundation for logarithmic temperature schedules.

Corollary 1. *Suppose that the temperature schedule is of the form*

$$T_k = \frac{c}{\log(k+2)}, \quad k = 0, 1, \ldots, \tag{10}$$

where c *is constant. Then the simulated annealing algorithm converges asymptotically to the set* S^* *of globally optimal states with probability* 1 *if and only if* $c \geq D$*.*

Proof. Suppose $c \geq D$. Then

$$\sum_{k=0}^{\infty} \exp(-D/T_k) \geq \sum_{k=0}^{\infty} \exp(-c/T_k) = \sum_{k=0}^{\infty} \exp(-\log(k+2)) = \sum_{k=2}^{\infty} \frac{1}{k} = \infty \tag{11}$$

and convergence to S^* follows.

Now suppose that $c < D$. Then there is a non-local minimum $\hat{x}$ such that its depth equals D. Since

$$\sum_{k=0}^{\infty} \exp(-D/T_k) = \sum_{k=0}^{\infty} (\exp(-c/T_k))^{D/c} = \sum_{k=2}^{\infty} \frac{1}{k^{D/c}} < \infty, \tag{12}$$

by part 2 of the Theorem $\lim_{k\to\infty} \mathrm{P}[X_k \in B] > 0$ for the bottom B of the cup associated with $\hat{x}$. □

3.4 On the depth of a problem instance

As regards computation of the maximum depth D, Kern [5] has shown that computing the maximum depth of a problem instance is NP-hard for the problem MAX CUT (partitioning the vertices of a graph into two parts such that the weight of the inter-part edges is maximized), with the local search neighbourhood defined by moving single vertices from a vertex set to the other.

Also, Kern makes the following conjectures:

Conjecture 1. *Computing the maximum depth is at least as hard as solving the optimization problem.*

Conjecture 2. *Computing the maximum depth is at most as hard as solving the optimization problem.*

Nevertheless, it is often easy to construct more or less tight upper bounds on D. As a generic example, we have the following bound.

Proposition 1. *Let D be the maximum depth of local minima that are not global minima. Then*

$$D \leq r\Delta \tag{13}$$

where

$$\Delta = \max_{x \in S} \max_{y \in \mathcal{N}(x)} |f(y) - f(x)| \tag{14}$$

and

$$r = \min_{x \in S \setminus \hat{S}} \max_{y \in S} d(x, y), \tag{15}$$

$\hat{S}$ denotes the set of all local minima and $d(x, y)$ is the distance of x and y in the neighbourhood graph.

Proof. Let $\hat{y}$ be a local, non-global minimum of maximum depth D, let $\tilde{x} \in S \setminus \hat{S}$ reach the minimum in (15). Fix an arbitrary global minimum x^*. Consider the shortest path in the neighbourhood graph from $\hat{y}$ via $\tilde{x}$ to x^*. By the definition of $\tilde{x}$, the path is at most $2r$ steps long. Let w be the first configuration on the path with $f(w) < f(\hat{y})$, and let z be the highest-cost configuration between $\hat{y}$ and w on the path.

Suppose that $D > r\Delta$. Then by the definition of the depth of a local minimum

$$f(z) - f(\hat{y}) \geq D > r\Delta \tag{16}$$

and z must be at least $r + 1$ steps away from $\hat{y}$ on the path. Now all the explicitly mentioned configurations must lie on the path in the order $\hat{y}, \tilde{x}, z, w, x^*$. Thus the path from z to x^* is less than r steps long, and we get

$$f(z) - f(x^*) < r\Delta. \tag{17}$$

Combining these inequalities yields $f(x^*) - f(\hat{y}) > 0$, implying that x^* is not globally optimal. This is a contradiction, and we conclude that necessarily $D \leq r\Delta$. □

For the cooling schedule (10) with the constant c given by the above upper bound $r\Delta$, the proof approach by inhomogenous Markov chain theory mentioned at the end of Subsection 3.2 is sufficient to show that the simulated annealing process converges to the set S^* of global minima. Indeed according to [1] this was one of the first cooling schedules shown to result convergence to S^*.

3.5 Conductance and finite-time simulated annealing

The property of conductance [8], defined as follows, allows more sophisticated analysis of the rate at which a Markov chain approaches its stationary distribution.

Definition 8. *The* conductance Φ_P *of a homogenous Markov chain with state space* Ω*, transition matrix* $P = (p_{xy})_{x,y\in\Omega}$ *and stationary balance probabilities* π_x*,* $x \in \Omega$*, is defined as*

$$\Phi_P = \min_{\substack{A\subset\Omega\,:\\ \sum_{x\in A}\pi_x \le 1/2}} \frac{\sum_{x\in A}\sum_{y\in S\setminus A}\pi_x p_{xy}}{\sum_{x\in A}\pi_x} \tag{18}$$

Roughly speaking, the fractional expression in (18) measures the stationary probability of escaping a subset of states, weighted by the stationary probability of being in the subset in the first place. Thus conductance can be thought of as a measure of how difficult state sets there are in the process, from the point of view of escaping said state sets.

Let us now apply this to the random walk process (ie. simulated annealing "at infinite temperature") on the neighbourhood graph of the combinatorial optimization problem. Assuming the neighbourhood graph is k-regular so that the stationary distribution is the uniform distribution, the conductance can be simplified to

$$\Phi = \min_{\substack{A\subset S:\\ |A|\le|S|/2}} \frac{\sum_{x\in A}|\mathcal{N}(x)\setminus A|}{|A|}. \tag{19}$$

Using this conductance parameter, Nolte and Schrader [9] claim the following bound on the finite time behaviour of simulated annealing using a logarithmic cooling schedule of the form (10), and assuming a symmetric neighbourhood relation.

Proposition 2. *Let* δ *be the difference between the minimal cost and the next to least cost value, and let* ρ *be the difference between the maximal and the minimal values of the cost function. Then there exist constants* $c_1, c_2 \in \mathbb{N}$ *such that for an arbitrary* $\epsilon > 0$ *and*

$$k \ge \frac{1}{\Phi^{c_2}}\left(\frac{|S|}{\epsilon}\right)^{-c_1\,\rho/\delta} \tag{20}$$

it holds that

$$\sum_{x\in S}|\mathrm{P}[X_k = x] - \pi^*_x| \le \epsilon. \tag{21}$$

The proposition gives directly the required number of iterations to reach a global optimum with predetermined probability. The bound (20) is polynomial in ϵ^{-1}, the inverse of the probability distribution accuracy, and in $|S|$, the number of configurations. Of course, the usefulness of the proposition is limited by the fact that in hard combinatorial optimization problems $|S|$ is exponential in the problem size.

4 Descent on random landscapes

In the sequel we identify the configuration space S with the set of N-bit strings. The neighbourhoods are defined by single-bit flip operations, so that every configuration has exactly N neighbours. As the present subject has been mostly studied in the context of evolution theory, some of the cited sources develop their results for the more general case of length-N strings over an arbitrary alphabet.

The connection to evolution theory also means that there is much more related material on the behaviour of genetic algorithms (sexual evolution) on the same landscape models (see eg. [11] for references), but that is beyond the scope of the present article.

4.1 Adaptive walks on uncorrelated landscapes

Let us first consider the completely random landscape, where the costs $f(x)$, $x \in S$, are independent, identically distributed random variables. An *adaptive walk* is defined as a random walk on S that proceeds by uniformly choosing a random neighbour, but then accepting the neighbour only if its cost is less than the cost of the current configuration. If the neighbour is not accepted, a new random neighbour is chosen and the acceptance test repeated. In evolution theory, the random neighbour choice represents mutation; however the results are applicable not only to randomized combinatorial optimization, but generally also to deterministic first-descent local optimization.

Note that qualitatively one may consider even a correlated landscape as uncorrelated, if one observes walks on the landscape only at intervals longer than the landscape correlation length.

Flyvbjerg and Lautrup [2] study the behaviour of adaptive walks on large random landscapes. To begin with, they observe that from the point of view of descent methods, instead of the actual costs it suffices to consider the shared cumulative distribution function of the random costs, effectively reducing the costs to uniform random variables on $(0, 1)$. We denote this transformed cost by F.

The *length* of an adaptive walk is defined as the number of configurations accepted during the walk. The *duration* of an adaptive walk is defined as the number of configurations generated, including those that are immediately rejected, during the walk.

Heuristically, the length of a walk can be characterized as follows. Consider the state of an adaptive walk at a particular configuration. Assume that adaptive walks are generally much shorter than N steps, so that the random step directions chosen during a walk are essentially all different, and on each step the current configuration has only one neighbour that has been seen before. On each step of the walk, a new cost value F' is encountered that is otherwise uncorrelated with the current cost F, except that it is smaller than the current one. Thus, on average F is halved on each step. Starting the walk with $F = 1$, after l steps the expected cost is 2^{-l}. An adaptive walk stops when all neighbour configurations have higher cost than the current configuration. On the average, this occurs when $F \sim 1/N$. Given that F decreases as 2^{-l} and the walk stops at a final cost value $F \sim 1/N$, we have an estimate for the average length L of an adaptive walk: $L \approx \log_2 N$.

Flyvbjerg and Lautrup do show more rigorously the following propositions. However, it should be noted that all their results are derived under the assumption that the search does not visit configurations that have been previously seen (either accepted, or generated and rejected), but they do give qualitative arguments for the negligibility of the error caused by this approximation.

Proposition 3. *The length L of an adaptive walk that starts from a configuration x_0 with $F(x_0) = 1$ is approximately Poisson-distributed, with expectation of the form $\ln N + C_1 + O(1/N)$ and variance $\ln N + C_2 + O(1/N)$, where $C_1 \approx 0.0991$ and $C_2 \approx 0.266$ are constants.*

Intuitively, it is clear that at least on the order of N configurations must be tested before one can know that a local optimum has been reached. The next proposition confirms this intuition.

Proposition 4. *The duration of an adaptive walk that starts from a configuration x_0 with $F(x_0) = 1$ is $C_3 N + O(1)$ with variance $C_4 N^2 + O(N)$ where $C_3 \approx 1.22$ and $C_4 \approx 1.72$.*

These results also indicate that the standard deviation of the length of an adaptive walk is relatively small, whereas the duration varies considerably more.

In essentially the same model, Macken, Hagan and Perelson [7] demonstrate similar results, notably without even discussing the possibility of the search returning in the neighbourhoods of previously seen configurations. In contrast to Flyvbjerg and Lautrup, they do not require that the initial configuration x_0 satisfy $F(x_0) = 1$, but are satisfied with $NF(x_0) \gg 1$. They derive for the expectation and variance of the length of an adaptive walk a result otherwise analogous to Proposition 3, but they get $C_1 \approx 1.10 + \log(F(x_0))$ and $C_2 \approx 0.266 + \log(F(x_0))$. Note that Macken et al. get a constant part for C_1 that is approximately 1 unit greater than the one derived by Flyvbjerg and Lautrup. For the duration of an adaptive walk Macken et al. only derive the expectation, and get

exactly the same result as Flyvbjerg and Lautrup, with no additional dependence on the initial cost value.

Additionally, Macken et al. provide the following results.

Proposition 5. *The cost of a randomly chosen local minimum is* $1/N + o(1/N)$ *with variance* $1/N^2 + o(1/N^2)$.

Proposition 6. *The cost of the final configuration of an adaptive walk that starts from a configuration* x_0 *with* $NF(x_0) \gg 1$ *is* $0.6243\ldots/N + o(1/N)$ *with variance* $0.8534\ldots/N^2 + o(1/N^2)$.

Together, these results indicate that on average an adaptive walk results in a local optimum 38% better than a randomly chosen local optimum.

In evaluating these results one should keep in mind that the only ways the neighbourhood structure affects an adaptive walk on an uncorrelated landscape are 1) by limiting the *size* of the set from which the next configuration is chosen, and 2) through the probability of returning to the neighbourhood of a previously seen configuration. Given that the second effect was ignored for simplicity in both of the cited analyses, the first effect is the only remaining difference to unrestricted random sampling of the configuration space. From this point of view, it may be of general interest to consider the results of Macken et al. [7] in their original form on N-strings on arbitrary alphabets, with the alphabet size as a parameter.

4.2 Adaptive walks on block model landscapes

Perelson and Macken [10] present a simple extension of the uncorrelated random landscape model, in order to study the effect of correlations on adaptive walks.

The *block model* is defined as a collection of B independent completely random landscapes of the type considered in Subsection 4.1, called *blocks*. The configuration spaces of the individual blocks may be of different sizes, and the length of the bit strings representing the configurations of block i is denoted by n_i. The configuration space of the whole model is the set of N-bit strings, where $N = \sum_{i=1}^{B} n_i$, considered as the Cartesian product of the configuration spaces of the individual blocks. The cost of a configuration in the block model is defined as the sum of the costs of the individual blocks.

When all the blocks are of the same size, the block model reduces to a special case of the N-k landscape model, but in general neither model subsumes the other. In a similar fashion to the N-k model, the ruggedness of the block model landscape can be tuned by adjusting the parameter B. When $B = N$, the block model has a unique local minimum, and when $B = 1$, the block model degenerates to the random landscape model.

Since the individual blocks are independent, the following result follows immediately from Proposition 3.

Proposition 7. *The expected length of an adaptive walk in the block model, starting from a configuration* x_0 *with* $F(x_0) = 1$, *is approximately* $\sum_{i=1}^{B} \ln(n_i) + BC_1$ *and the variance is* $\sum_{i=1}^{B} \ln(n_i) + BC^2$, *where* $C_1 \approx 0.0991$ *and* $C_2 \approx 0.266$ *are constants.*

An analogous result can be derived for the duration of an adaptive walk from Proposition 4.

4.3 Descent walks on N-k landscapes

Definition 9. *An N-k landscape is defined on the configuration space* S *of the set of* N*-bit strings as follows. The moves comprise single-bit flip operations. The cost function is of the form* $f(x) = \sum_{i=1}^{N} f_i(x_{s_i(0)}, x_{s_i(1)}, x_{s_i(2)}, \ldots, x_{s_i(k)})$, *where* $s_i(j)$, $i = 1, \ldots, N$, $j = 0, \ldots, k$, *determines the interactions between bit* i *and* k *other bits. The values of the cost component functions* f_i *are independent identically distributed random variables.*

We shall assume here that $s_i(0) = i$ for all $i = 1, \ldots, N$.

When $k = 0$, the landscape has a unique local (and thus also global) minimum and the expected length of a downhill walk is $N/2$. At the other extreme, when $k = N - 1$ the landscape is totally random and has $O(2^N/N)$ local optima, and walks to optima are of expected length $O(\ln N)$, as seen in the previous Subsection. Generally, the larger k grows, the less correlated the landscape becomes.

A *gradient walk* is a deterministic walk on the configuration space that always chooses the neighbour with the least cost. This is also known as the *steepest-descent* method. This is compared here with the first-descent method, ie. adaptive walks.

Weinberger [13] derives several qualitative results on local optimization on $N - k$ landscapes with $1 \ll k \ll N$. Because the results are based on the Central Limit Theorem, the random values of the cost component functions must be assumed to have finite mean and variance. To give a taste of the kind of heuristic reasoning applied, for one of the results we present the justification given by Weinberger, slightly expanded.

Proposition 8. *The expected number of local minima is $O((2\lambda)^N)$ where*

$$\lambda \approx \left(\frac{1}{k+1}\right)^{1/(k+1)}. \tag{22}$$

As k grows, λ as defined by (22) approaches 1, and the number of local minima approaches $O(2^N)$. On the other hand, k is bounded by the assumption $k \ll N$ and as noted above, in the case $k = N - 1$ the expected number of local optima is only $O(2^N/N)$.

Proposition 9. *The expected cost of a local minimum is approximately*

$$\mu - \sigma \left(\frac{2\ln(k+1)}{k+1}\right)^{1/2} \tag{23}$$

where μ and σ are respectively the mean and standard deviation of the cost components.

The expected cost of a local minmum is a slowly increasing function of k. Hence the more correlated the fitness landscape is, the steeper valleys one can expect to find in the landscape. Here we have a clear qualitative difference with the uncorrelated case $k = N - 1$ as characterized in Proposition 5.

Proposition 10. *The expected length of an adaptive walk is approximately $N\frac{\ln(k+1)}{k+1}$.*

As expected, the adaptive walks become shorter when the landscape correlation decreases. Again, the contrast with the uncorrelated case, described in Proposition 3, is marked.

Proposition 11. *The expected length of a gradient walk is approximately $\frac{D}{2} - \left(\frac{D}{2\pi}\right)^{1/2}$ where*

$$D \approx N \log_2(k+1)/(k+1). \tag{24}$$

Justification. From the expected number of local minima, it is concluded that the expected Hamming distance D between local minima (which equals the expected diameter of a basin of attraction) is asymptotically

$$\begin{aligned} D = \log_2\left(\frac{2^N}{O((2\lambda)^N)}\right) &= N - \log_2\left(C(2\lambda)^N + o(N)\right) \\ &\approx N - \log_2(C) - N + N\log_2(\lambda) \approx N\log_2(\lambda). \end{aligned} \tag{25}$$

The gradient walk is expected to almost always end up in the nearest local minimum. The probability that the random initial state is at Hamming distance d from the nearest local minimum, considering both of the two closest local minima, is approximated as

$$2 \cdot \binom{D}{d} \Big/ 2^D = 2\binom{D}{d} \left(\frac{1}{2}\right)^{D-d} \left(\frac{1}{2}\right)^{d}$$
$$\approx 2\frac{1}{\sqrt{2\pi D/4}} \exp\left(-\frac{(d-D/2)^2}{2 \cdot D/4}\right) = \left(\frac{8}{\pi D}\right)^{1/2} \exp\left(-2(d-D/2)^2/D\right) \quad (26)$$

.

The mean Hamming distance from the random start to the chosen optimum can then be approximated as

$$D - \int_{D/2}^{\infty} r \left(\frac{8}{\pi D}\right)^{1/2} \exp\left(-2(r-D/2)^2/D\right)\, dr$$
$$= D + 2\frac{D}{4}\int_{D/2}^{\infty} \left(-\frac{4}{D}r + 2\right) \left(\frac{2}{\pi D}\right)^{1/2} \exp\left(-2(r-D/2)^2/D\right)\, dr$$
$$-2\frac{D}{4}2\int_{D/2}^{\infty} \left(\frac{2}{\pi D}\right)^{1/2} \exp\left(-2(r-D/2)^2/D\right)\, dr$$
$$= D + 2\frac{D}{4} \cdot \Big|_{D/2}^{\infty} \left(\frac{2}{\pi D}\right)^{1/2} \exp\left(-2(r-D/2)^2/D\right) - D \cdot \frac{1}{2}$$
$$= \frac{D}{2} + \frac{D}{2}\left(0 - \left(\frac{2}{\pi D}\right)^{1/2}\right) = \frac{D}{2} - \left(\frac{D}{2\pi}\right)^{1/2}. \quad (27)$$

□

Weinberger also presents numerical results from computer simulations of adaptive and gradient walks on landscapes with different values of N and k. The analytical estimates fit the numerical results relatively well, and the qualitative behaviour of the analytical estimates is similar to that of the numerical results of the presented test runs.

References

[1] Emile H. L. Aarts, Jan H. M. Korst, and Peter J. M. van Laarhoven. Simulated annealing. In *Local Search in Combinatorial Optimization*, chapter 4, pages 91–120. John Wiley & Sons, 1997.

[2] Henrik Flyvbjerg and Benny Lautrup. Evolution in a rugged fitness landscape. *Phys. Rev. A*, 46:6714–6723, 1992.

[3] B. Hajek. Cooling schedules for optimal annealing. *Math. Operations Res.*, 13:311–329, 1988.

[4] Daniel P. Heyman and Matthew J. Sobel. *Stochastic Models in Operations Research Volume I: Stochastic Processes and Operating Characteristics*. McGraw-Hill, 1982.

[5] W. Kern. On the depth of combinatorial optimization problems. *Discrete Applied Mathematics*, 43:115–129, 1993.

[6] S. Kirkpatrick, C. D. Gelatt, and M. P. Vecchi. Optimization by simulated annealing. *Science*, 220:671–680, 1983.

[7] C. A. Macken, P. S. Hagan, and A. S. Perelson. Evolutionary walks on rugged landscapes. *SIAM J. Appl. Math.*, 51:799–827, 1991.

[8] Milena Mihail. Conductance and convergence of Markov chains – a combinatorial treatment of expanders. In *Proceedings of the 30th annual Symposium on Foundations of Computer Science*, Washington, DC, USA, 1989. IEEE.

[9] Andreas Nolte and Rainer Schrader. A note on the finite time behaviour of simulated annealing. In *Operations Research Proceedings*, pages 175–180. Springer Verlag, 1996.

[10] A. S. Perelson and C. A. Macken. Protein evolution on partially correlated landscapes. *Proc. Natl. Acad. Sci. USA*, 92:9657–9661, 1995.

[11] Christian M. Reidys and Peter F. Stadler. Combinatorial landscapes. *SIAM Review*, 44:3–54, 2002.

[12] Fabio Romeo and Alberto Sangiovanni-Vincentelli. A theoretical framework for simulated annealing. *Algorithmica*, 6:302–345, 1991.

[13] E. D. Weinberger. Local properties of kauffman's n-k model: A tunably rugged energy landscape. *Phys. Rev. A*, 44:6399–6413, 1991.