Dynamics on Landscapes

Hannu Rummukainen

18.3.2002

Definitions

Let S be a configuration space.

Let S^* denote the set $\{x \in S : f(x) = \min_{y \in S} f(y)\}$ of global minima.

Let each $x \in S$ be assigned a neighbourhood $\mathcal{N}(x) \subset S$ with $x \notin \mathcal{N}(x)$.

Algorithm: Simulated annealing

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.

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

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

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

The transition probabilities $P_{xy}(k) := \mathrm{P}[X_{k+1} = y \mid X_k = x]$ 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.

For a given 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 are given by

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

Theorem. *Let $P = (p_{xy})_{x,y\in\Omega}$ be the transition matrix associated with a finite 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 $\sum_{x\in S} \pi_x = 1$ and*

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

As a consequence of this Theorem, it can be shown 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 $(\pi_x)_{x\in S}$ given by

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

As $T \to 0$, the above stationary distribution approaches the limit distribution $(\pi^*_x)_{x\in S}$ for which $\pi^*_x = 1/|S^*|$, $x \in S^*$, and $\pi^*_x = 0$, $x \in S \setminus S^*$.

Definition. *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. *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.*

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

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

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

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

Definition. *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.*

Theorem (Hajek 1988). *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 $\sum_{k=1}^{\infty} \exp(-d/T_k) = \infty$.*
3. *(Consequence of 1 and 2.) Let D be the maximum of the depths of all states which are local but not global minima. Then*

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

if and only if

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

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

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

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

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

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 now

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

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 $\hat{x}$. □

Kern (1993) has shown that for the problem MAX CUT with the neighbourhood defined by moving single vertices from one side of the cut to the other side, computing the maximum depth D of a problem instance is NP-hard.

Also, Kern makes the following conjectures:

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

Conjecture. *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.

Definition. *The* conductance Φ_P *of a homogenous Markov process 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*

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

Let Φ be the conductance of the simulated annealing process at an infinite temperature (ie. a random walk on the neighbourhood graph).

Nolte and Schrader (1996) show the following bound on the finite time behaviour of simulated annealing using a logarithmic cooling schedule $T_k = \gamma / \ln(k)$, and assuming a symmetric neighbourhood relation.

Theorem. *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*

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

it holds that

$$\sum_{x \in S} |\mathrm{P}[X_k = x] - \pi_x^*| \leq \epsilon.$$

Descent on random landscapes

Let the configuration space S be the set of N-bit strings. Suppose that the costs $f(x)$, $x \in S$, are i.i.d. random variables.

Definition. *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, is called an* adaptive walk*.*

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 (1992) study the behaviour of adaptive walks on large random landscapes.

First, 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 $(0, 1)$.

A heuristic argument

- 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 a 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 fitness 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 somewhat more rigorously that the length L of an adaptive walk that starts from a configuration with $F = 1$ is approximately Poisson-distributed, with expectation of the form $\ln N + \text{constant} + O(1/N)$ and variance $\ln N + \text{constant} + O(1/N)$.

Further, they show that the number of configurations tested during such a walk is $1.224 \cdots N + O(1)$ with variance $1.72 \cdots N^2 + O(N)$.

In essentially the same model, Macken, Hagan and Perelson (1991) demonstrate similar results, including the following:

- The cost of a randomly chosen local minimum is $1/N + o(1/N)$ with variance $1/N^2 + o(1/N^2)$.
- The cost of the final state of an adaptive walk is $0.6243 \ldots /N + o(1/N)$ with variance $0.8534 \ldots /N^2 + o(1/N^2)$.

Definition. *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 i.i.d. 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 $N/2$.

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 $O(\ln N)$.

Weinberger (1991) 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.

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

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

2. The expected cost of a local minimum is approximately

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

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

4. The average length of an adaptive (first-descent) walk is approximately $N\frac{\ln(k+1)}{k+1}$.

3. The expected length of a gradient (steepest-descent) walk is approximately $\frac{D}{2} - \left(\frac{D}{2\pi}\right)^{1/2}$ where $D \approx N \log_2(k+1)/(k+1)$.

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

$$D = \log_2\left(\frac{2^N}{O((2\lambda)^N)}\right) = N - \log_2\left(C(2\lambda)^N + o(N)\right)$$
$$\approx N - \log_2(C) - N + N\log_2(\lambda) \approx N\log_2(\lambda).$$

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

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

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

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

□

References

[1] H. Flyvbjerg and B. Lautrup. Evolution in a rugged fitness landscape. *Phys. Rev. A*, 46:6714–6723, 1992.

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

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

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

[5] 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.

[6] Andreas Nolte and Rainer Schrader. A note on the finite time behaviour of simulated annealing. In *Operations Research Proceedings*, 1996.

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