Simulated Annealing

Terminology and Definitions

The annealing algorithm is defined on a **state space** S, which is a collection of individual **states**. Each of these states represents a **configuration** of the problem under investigation. The states are related by a neighborhood system, and the set of neighbor pairs in S defines a substructure M in $S \times S$. The elements in M are called **moves**. Two states s, s' are called **adjacent**, if they can be reached by a single move (i. e. $(s, s') \in M$). Similarly, $(s, s') \in M^k$ are said to be connected via a set of k moves. We require the state space to be finite. The size of the state space is fixed, but can be arbitrarily large, therefore this assumption does not result in a loss of generality - for our purposes a computer must be able to distinguish all states anyways. The following functions govern the search through the state space.

Definition 1 *The* **score function**

$$\epsilon : S \to I\!R_+ \tag{1}$$

assigns a positive real number (score) to each state.

The score is understood as a measure of the quality of the state. In the following we always assume that lower scores are associated with states that represent better quality configurations. Since the state space is finite, there exists at least one state with a minimal score. This score is denoted by ϵ_0.

Definition 2 *The* **selection probability** *is a function*

$$\beta : S \times S \to [0, 1] \tag{2}$$

such that

$$\forall_{(s,s') \notin M} \beta(s, s') = 0, \tag{3}$$

$$\forall_{(s,s') \in M} \beta(s, s') \neq 0, \tag{4}$$

$$\forall_{s \in S} \sum_{s' \in S} \beta(s, s') = 1. \tag{5}$$

The selection probability therefore is the probability that state s' is proposed as new state, given that the current state is s. Therefore the move set can be defined as

$$M := \{(s, s') \in S \times S : \beta(s, s') > 0\} \tag{6}$$

We call the move set M **symmetric** if

$$\forall_{s \in S} \forall_{s' \in S} \left[(s, s') \in M \Rightarrow (s', s) \in M\right] \tag{7}$$

Definition 3 *The* **acceptance function**

$$\alpha : I\!R_+^3 \to (0, 1] \tag{8}$$

assigns a positive probability to a pair of scores and a positive real number, called the **temperature**.

The acceptance function decides whether or not the proposed state will be accepted as the new state. Note that for any fixed temperature, this probability only depends on the scores of the current and proposed state, but not on those states themselves.

Definition 4 *The* **transition probability** *is a function*

$$\tau : I\!R_+^3 \to [0, 1] \tag{9}$$

defined as

$$\tau(s, s', t) := \begin{cases} \alpha(\epsilon(s), \epsilon(s'), t) \times \beta(s, s') & s \neq s' \\ 1 - \sum_{s'' \in S} \alpha(\epsilon(s), \epsilon(s''), t) \times \beta(s, s'') & s = s'. \end{cases} \tag{10}$$

Therefore, the transition probability $\tau(s, s', t)$ can be understood as the probability that the next step is a move to state s', given that the current state is s and the temperature is t. The probability that the state after n moves is s', given the current state s and temperature the t, will be denoted $\tau_n(s, s', t)$.

A process that possesses the above property is called a **Markov process**. A sequence of events as a special case of such a Markov process is called a **Markov chain**. A Markov chain in which the transition probabilities between the pairs of states are constant throughout the process is called **homogeneous**. A Markov chain is called **irreducible** if any state in the chain is connected to any other state by only a finite number of moves, i. e. if

$$\bigcup_k M^k = S \times S. \tag{11}$$

A Markov chain is called **aperiodic** if for every state s the greatest common divisor of all integers $n \geq 1$ with $\tau_n(s, s, \cdot) > 0$ is equal to 1.

Properties of Markov Chains

Theorem 1 below is usually referred to as the Chain Limit Theorem. It states that an irreducible and aperiodic (homogeneous) Markov chain has a limiting distribution.

Theorem 1 *For each irreducible and aperiodic chain there exists a density function*

$$\pi : S \times I\!R_+ \to (0, 1], \tag{12}$$

in s for any given $t > 0$, with

$$\pi(s, t) = \lim_{n\to\infty} \tau_n(s', s, t), \tag{13}$$

(independent of s') and satisfying the following equations:

$$\sum_{s'\in S} \pi(s', t)\tau(s', s, t) = \pi(s, t), \tag{14}$$

$$\sum_{s\in S} \pi(s, t) = 1. \tag{15}$$

Hence, if we constructed an irreducible and aperiodic (homogeneous) Markov chain for the annealing algorithm (i. e. run the chain at a fixed temperature), the distribution of states we sample from approaches a limit. However, the search through the state space should yield low scoring states. Some simple additional requirements will guarantee this.

Theorem 2 *An irreducible and aperiodic chain with a symmetric move set has the property*

$$\forall_{s\in S}[\epsilon(s) \neq \epsilon_0 \Rightarrow \lim_{t\downarrow 0} \pi(s, t) = 0] \tag{16}$$

if it has an acceptance α function satisfying

$$\epsilon \geq \epsilon' \Rightarrow \alpha(\epsilon, \epsilon', t) = 1, \tag{17}$$

$$\epsilon > \epsilon' > \epsilon'' \Rightarrow \alpha(\epsilon, \epsilon', t) \times \alpha(\epsilon', \epsilon'', t) = \alpha(\epsilon, \epsilon'', t), \tag{18}$$

$$\epsilon < \epsilon' \Rightarrow \lim_{t\downarrow 0} \alpha(\epsilon, \epsilon', t) = 0. \tag{19}$$

Hence, if the requirements (17), (18) and (19) are satisfied, the likelihood of a non-optimal scoring state in the limiting distributions goes to zero as the temperature goes to zero. Therefore, if the annealing is run as a sequence of homogeneous Markov chains with decreasing temperatures, the search is guided towards optimal scoring states. The above mentioned requirements only affect the acceptance function and do not pose any constraints on β or M. In general it is quite easy to construct a state space with a symmetric move set that guarantees irreducibility and aperiodicity for the chain in the search algorithm. The desirable properties of the chains as stated in Theorem 2 can be achieved by choosing the right acceptance function.

Otten and van Ginneken make the point that "it seems reasonable that smaller score increases are accepted with higher probability than bigger ones, and that this probability varies smoothly with the score difference". Sufficient but not necessary for this would be the requirement that the acceptance only depends on the score difference.

Theorem 3 *The only acceptance functions* $\alpha(\epsilon, \epsilon', t)$

- *that are differentiable in* ϵ'*,*
- *whose values depend on* t *and the difference of* ϵ *and* ϵ'
- *that satisfy the conditions of Theorem 2*

have the form

$$\alpha(\epsilon, \epsilon', t) = \min\{1, e^{(\epsilon' - \epsilon)c(t)}\}, \tag{20}$$

where $c(t)$ *is a negative, monotonic and continuous function satisfying*

$$\lim_{t \downarrow 0} c(t) = -\infty. \tag{21}$$

The standard choice is $c(t) = -1/t$, yielding the acceptance function

$$\alpha(\epsilon, \epsilon', t) = \min\{1, e^{-(\epsilon' - \epsilon)/t}\}. \tag{22}$$

This acceptance function has been used by far the most in the literature. This is presumably the case because condensed matter physics is the origin of simulated annealing and the above acceptance function has a striking similarity to the Boltzmann distribution, which characterizes a system of particles in thermal equilibrium. However, it also has been established that this acceptance function has many desirable properties, as described above.

Great references are:

P J van Laarhoven and E H Aarts (1987), "Simulated Annealing: Theory and Applications", Kluwer Academic Publishers.

R H Otten and L P Ginneken (1989), "The Annealing Algoritm", Kluwer Academic Publishers.