Markov Chain Monte Carlo.
Simulated Annealing.

Anatoli Iambartsev
IME-USP

[RC] Stochastic search.

General iterative formula for optimizing a function h is

$$\theta_{t+1} = \theta_t + \varepsilon_t,$$

which makes the sequence (θ_n) a Markov chain. The simulated annealing generate ε's in the following way. "Rather than aiming to follow the slopes of the function h, simulated annealing defines a sequence $\{\pi_t\}$ of densities whose maximum arguments are confounded with the arguments of $\max h$ and with higher and higher concentrations around this argument. Each θ_t in the sequence is then simulated from the density π_t according to a specific update mechanism."

[RC] Simulated annealing.

"The construction of the sequence of densities $\{\pi_t\}$ is obviously the central issue when designing a simulated annealing algorithm. The most standard choice is based on the Boltzman-Gibbs transforms of h,

$$\pi_t(\theta) \propto \exp\left(\frac{h(\theta)}{T_t}\right), \tag{1}$$

where the sequence of temperatures, $\{T_t\}$, is decreasing (under the assumption that the right-hand side is integrable). It is indeed clear that, as T_t decreases toward 0, values simulated from π_t become concentrated in a narrower and narrower neighborhood of the maximum (or maxima) of h."

[RC] Simulated annealing.

"The choice (1) is a generic solution to concentrate (in t) the distribution π_t around the maxima of an arbitrary function h, but other possibilities are available in specific settings. For instance, when maximizing a likelihood $\ell(\theta \mid x)$, the pseudo-posterior distributions

$$\pi_t(\theta \mid x) \propto \ell(\theta \mid x)^{m_t} \pi_0(\theta),$$

associated with a nondecreasing integer sequence $\{m_t\}$ and an arbitrary prior π_0, enjoy the same property."

[RC] Simulated annealing.

"Two practical issues that hinder the implementation of this otherwise attractive algorithm are

(a) the simulation from π_t and

(b) the selection of the temperature sequence (or schedule) $\{T_t\}$.

While the second item is very problem-dependent, the first item allows a generic solution, related to the Metropolis-Hastings algorithm."

[RC] Simulated annealing.

The update from θ_t to θ_{t+1} is indeed based on the Metropolis-Hastings step: ζ is generated from a distribution with symmetric density g, and the new value θ_{t+1} is generated as

$$\theta_{t+1} = \begin{cases} \theta_t + \zeta & \text{with probability } \rho = \exp(\Delta h / T_t) \wedge 1, \\ \theta_t & \text{with probability } 1 - \rho, \end{cases}$$

where $\Delta h = h(\theta_t + \zeta) - h(\theta_t)$.

By allowing random moves that may see h decrease, the simulated annealing method can explore multimodal functions and escape the attraction of local modes as opposed to deterministic (and to some extent stochastic) gradient methods.

Algorithm 2 Simulated Annealing

At iteration t,

1. Simulate $\zeta \sim g(\zeta)$;
2. Accept $\theta_{t+1} = \theta_t + \zeta$ with probability $\rho_t = \exp\{\Delta h_t / T_t\} \wedge 1$; take $\theta_{t+1} = \theta_t$ otherwise.

the density g being symmetric (around 0) but otherwise almost arbitrary.

An R version of this algorithm is associated with a random generator from g, randg, as in Algorithm 1,

```
> theta=rep(theta0,Nsim)
> hcur=h(theta0)
> xis=randg(Nsim)
> for (t in 2:Nsim){
+   prop=theta[t-1]+xis[t]
+   hprop=h(prop)
+   if (Temp[t]*log(runif(1))<hprop-hcur){
+   theta[t]=prop
+   hcur=hprop
+   }else{
+      theta[t]=theta[t-1]}}
```

where the temperature sequence Temp needs to be defined by the user.

[LA] Simulated annealing.

“As early as 1953, Metropolis et al. [MET53] proposed an algorithm for the efficient simulation of the evolution of a solid to thermal equilibrium. It took almost thirty years before Kirkpatrick et al. [KIR82] and, independently, Cerny [CER85] realized that there exists a profound analogy between minimizing the cost function of a combinatorial optimization problem and the slow cooling of a solid until it reaches its low energy ground state and that the optimization process can be realized by applying the Metropolis criterion. By substituting cost for energy and by executing the Metropolis algorithm at a sequence of slowly decreasing temperature values Kirkpatrick and his co-workers obtained a combinatorial optimization algorithm, which they called *simulated annealing*. Since then, the research into this algorithm and its applications has evolved into a field of study in its own.”

[LA] Simulated annealing.

It is generally known as *simulated annealing*, due to the analogy with the simulation of the annealing of solids it is based upon, but it is also known as

Monte Carlo annealing, *statistical cooling*, *probabilistic hill climbing*, *stochastic relaxation* or *probabilistic exchange algorithm*.

[H] Cooling schedule.

MATHEMATICS OF OPERATIONS RESEARCH
Vol. 13, No. 2, May 1988
Printed in U.S.A.

COOLING SCHEDULES FOR OPTIMAL ANNEALING*†

BRUCE HAJEK

University of Illinois at Champaign-Urbana

A Monte Carlo optimization technique called "simulated annealing" is a descent algorithm modified by random ascent moves in order to escape local minima which are not global minima. The level of randomization is determined by a control parameter T, called temperature, which tends to zero according to a deterministic "cooling schedule". We give a simple necessary and sufficient condition on the cooling schedule for the algorithm state to converge in probability to the set of globally minimum cost states. In the special case that the cooling schedule has parametric form $T(t) = c/\log(1+t)$, the condition for convergence is that c be greater than or equal to the depth, suitably defined, of the deepest local minimum which is not a global minimum state.

[H] Cooling schedule.

According this paper, instead of θ_t we will use here X_k as a state of a discrete Markov chain with a state space $\mathcal{S}$. The optimize problem is to minimize a function V. Let $\mathcal{S}^*$ be the set of state in $\mathcal{S}$. We are interested in determining whether

$$\lim_{k\to\infty} \mathbb{P}(X_k \in \mathcal{S}^*) = 1.$$

[H] Cooling schedule.

Let $\pi_T(x)$ be stationary distribution for Markov chain (X_k) and let as before $\pi_T(x) \propto \exp\left(-\frac{V(x)}{T}\right)$. The fact that the chain is aperiodic and irreducible means that

$$\lim_{k\to\infty} \mathbb{P}(X_k \in \mathcal{S}^*) = \sum_{x\in\mathcal{S}^*} \pi_T(x).$$

Examination of π_T soon yields that the right-hand side can be made arbitrary close to one by choosing T small. Thus

$$\lim_{T\to 0}\left(\lim_{k\to\infty, T_k\equiv T} \mathbb{P}(X_k \in \mathcal{S}^*)\right) = 1.$$

[H] Cooling schedule.

State y is reachable at height E from state x if $x = y$ and $V(x) \leq E$, or if there is a sequence of states $x = x_0, x_1, ..., x_p = y$ for some $p \geq 1$ such that $x_{k+1} \in N(x_k)$ for $0 \leq k < p$ and $V(x_k) \leq E$ for $0 \leq k \leq p$.

Property WR (Weak reversibility): For any real number E and any two states x and y, x is reachable at height E from y if and only if y is reachable at height E from x.

We define a *cup* for $(\mathscr{S}, V, N)$ to be a set C of states such that for some number E, the following is true: For every $x \in C$, $C = \{y\colon\ y$ can be reached at height E from $x\}$. For example, by Property *WR*, if $E \geqslant V(x)$ then the set of states reachable from x at height E is a cup. Given a cup C, define $\underline{V}(C) = \min\{V(x)\colon\ x \in C\}$ and $\overline{V}(C) = \min\{V(y)\colon\ y \notin C$ and $y \in N(x)$ for some x in $C\}$. The set defining $\overline{V}(C)$ is empty if and only if $C = \mathscr{S}$, and we set $\overline{V}(\mathscr{S}) = +\infty$. We call the subset B of C defined by $B = \{x \in C\colon V(x) = \underline{V}(C)\}$ the *bottom* of the cup, and we call the number $d(C)$ defined by $d(C) = \overline{V}(C) - \underline{V}(C)$ the *depth* of the cup. These definitions are

[H] Cooling schedule.

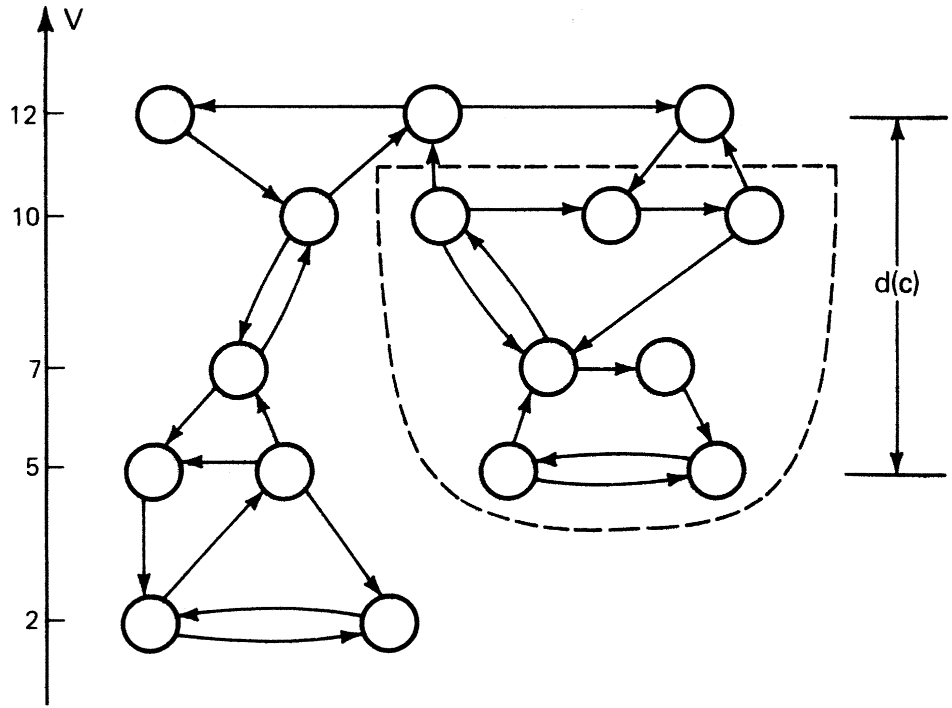

FIGURE 1.2. A cup C is enclosed with dashed lines. $\underline{V}(C) = 5$, $\overline{V}(C) = 12$, $d(C) = 7$ and the bottom B of C contains two states.

[H] Cooling schedule. Main theorem.

Assume (X_k) is irreducible and satisfies WR property, and let (T_k) be a sequence of strictly positive numbers such that $T_1 \geq T_2 \geq \ldots$ and $\lim_{k\to\infty} T_k = 0$.

(a) For any state x that is not a local minimum,

$$\lim_{k\to\infty} \mathbb{P}(X_k = x) = 0.$$

(b) Let B be a bottom of a cup of depth d (states in B are local minima of depth d). Then

$$\lim_{k\to\infty} \mathbb{P}(X_k \in B) = 0 \text{ iff } \sum_{k=1}^{\infty} \exp(-d/T_k) = \infty.$$

[H] Cooling schedule. Main theorem.

Consequence of (a) and (b): Let d^* be the maximum of the depths of all states which are local but not global minima. Then

$$\lim_{k\to\infty} \mathbb{P}(X_k \in \mathcal{S}^*) = 1 \text{ iff } \sum_{k=1}^{\infty} \exp(-d^*/T_k) = \infty. \tag{2}$$

Remark: If T_k assumes the parametric form

$$T_k = \frac{c}{\log(k+1)}$$

then (2) is true if and only if $c \geq d^*$.

References.

[RC] Cristian P. Robert and George Casella. *Introducing Monte Carlo Methods with R*. Series "Use R!". Springer

[LA] P.J.M. van Laarhoven and E.H.L. Aarts. *Simulated Annealing: Theory and Applications*. Series "Mathematics and its applications", Springer, 1987.

[KIR82] Kirkpatrick, S., C.D. Gelatt Jr. and M.P. Vecchi, *Optimization by Simulated Annealing*, IBM Research Report RC 9955, 1982.

[CER85] Černy, V., *Thermodynamical Approach to the Traveling Salesman Problem: An Efficient Simulation Algorithm*, J. Opt. Theory Appl., 45(1985) 41-51.

[H] Bruce Hajek. *Cooling schedules for optimal annealing*, Mathematics of Operational Research, Vol. 13, No. 2, May 1988.