Doktorandkurs i matematisk statistik
BAYESIAN STATISTICS AND MCMC COMPUTATION
Simulated Annealing and Optimization
ht03/vt01 2004-2005
Timo Koski
matematisk statistik, LiTH

1 Introduction

Simulated annealing is a term derived from the physical process of heating and then slowly cooling a crystalline substance and the observation that, if the structure is cooled sufficiently slowly, the molecules will line up in a rigid pattern corresponding to a state of minimum energy. This indicates a deep connection between statistical physics and optimization of functions of many variables.

The simulated annealing algorithm (SA) imitates this process by producing a sequence of samples from a series of probability distributions that move towards the point mass at the minimum of a chosen objecitvd function as 'temperature' is lowered.

We are going to interpret SA via the Metropolis-Hastings algorithm.

2 A Concept for Optimization Theory

We recognize first the following result found in (Pincus 1968).

Proposition 2.1 Consider $h(x)$ a real-valued continuous function defined on a closed and bounded set $\mathcal{X} \subset \mathcal{R}^n$. If there exists a unique x^* satisfying

$$x^* = \text{argmax}_{x \in \mathcal{X}} h(x), \tag{2.1}$$

then

$$\lim_{\lambda\to\infty} \frac{\int_{\mathcal{X}} x e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx} = x^*. \tag{2.2}$$

■

The proof of this is given in Appendix A below.

The result in the proposition shows that the for large λ the major contribution to the integral in (2.2) comes from a small neighborhood of x^*. Hence a Markov Chain Monte Carlo method that generates a Markov chain that spends, in the long run, most of the time visiting states near the maximizing point could find x^*.

Let us think of generating samples with the (target) density

$$f_\lambda(x) = \begin{cases} \frac{1}{\int_{\mathcal{X}} e^{\lambda h(x)} dx} e^{\lambda h(x)} & x \in \mathcal{X}, \\ 0 & \text{elsewhere.} \end{cases} \tag{2.3}$$

We make a discretization, as discussed in (Pincus 1970). Let us partition the region $\mathcal{X}$ into a finite nubmer of N mutually disjoint subsets $\mathcal{X}_j$. We fix points $y^j \in \mathcal{X}_j$. Then one consructs an irreducible, aperiodic finite Markov chain $\{X_n\}_{n\geq 0}$ with state space

$$S = \{y^1, y^2, \ldots, y^N\},$$

with target distribution (invariant distribution)

$$f_j = \frac{e^{\lambda h\left(y^j\right)}}{\sum_{j=1}^N e^{\lambda h(y^j)}}$$

where the subsets $\mathcal{X}_j$ have been assumed to have equal volumes. Then, the strong law of large numbers for $\{X_n\}_{n\geq 0}$ gives

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n X_i \to \frac{\int_{\mathcal{X}} x e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx} \approx x^*.$$

This is a technique of optimization that can be implemented by a simulated annealing algorithm, which is a special kind of Metropolis algorithm.

C-R. Hwang (1980) has in more general setting shown that the probability measure P_λ that corresponds to the density $f_\lambda(x)$ in (2.3) converges (weakly) to a measure that is uniform on the set $\mathbf{M}$ of maxima of $h(x)$, or

$$\mathbf{M} = \{x^* \in \mathcal{R}^n | x^* = \text{argmax}_{x\in\mathcal{X}} h(x)\}.$$

(Häggström 2002, chapter 13).

3 The Simulated Annealing Algorithm

3.1 Definition of SA as as a Metropolis algorithm

First let us recall the Metropolis Algorithm with the target distriubtion (density) f and a symmetric proposal density $q(y|x)$. In a quite general manner we can acheive symmetry by assuming that

$$q(y|x) = q\left(\|y - x\|\right)$$

The probability of acceptance in the Metropolis Algorithm depends only on the ratio $f(y)/f(x)$, or

$$\rho(x, y) = \min\left\{1, \frac{f(y)}{f(x)}\right\}. \tag{3.1}$$

Let us now consider target distribution of the following form

$$f(x) = \frac{e^{\frac{h(x)}{T}}}{Z}, \tag{3.2}$$

where $1/Z$ is the normalizing constant, and $h(x)$ is some function of x, $T > 0$, where x is notation for a vector of real numbers.

Then the probability of acceptance in (3.1) becomes

$$\rho(x, y) = \min\left\{1, e^{\frac{C(y)-h(x)}{T}}\right\}. \tag{3.3}$$

If $h(y) > h(x)$, then y is accepted with probability one. However, even if $h(y) < h(x)$, y may be accepted with probability $e^{\frac{h(y)-h(x)}{T}}$. Both of these probabilities depend on the scale T.

We take T first as a constant independent of n. Then we get

Definition 3.1 /**Homogeneous Simulated Annealing Algorithm**/ Given that $X_n = x_n$

1. Generate $Y_{n+1} \sim q\left(y|x_n\right)$.
2. Take

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{with probability } \rho\left(x_n, Y_{n+1}\right) \\ i & \text{with probability } 1 - \rho\left(x_n, Y_{n+1}\right), \end{cases}$$

where

$$\rho(x, y) = \min\left\{1, e^{\frac{h(y)-h(x)}{T}}\right\}. \tag{3.4}$$

3. $X_{n+1} \mapsto x_n$ and return to 1.

■

Let us allow T as a decreasing function of n, denoted by T_n,

$$T_1 > T_2 > \ldots > T_n > \ldots$$

is called a **cooling schedule**. The cooling will have to be slow enough, and a usual choice is

$$T_n = \frac{c}{\log(n+1)}, \quad n \geq 1. \tag{3.5}$$

where c is a positive constant. Then we have

Definition 3.2 *[***Simulated Annealing Algorithm***]* Given that $X_n = x_n$

1. Generate $Y_{n+1} \sim q\,(y|x_n)$.

2. Take

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{with probability } \rho_n\,(x_n, Y_{n+1}) \\ i & \text{with probability } 1 - \rho_n\,(x_n, Y_{n+1}), \end{cases}$$

   where

$$\rho_n(x, y) = \min\left\{1, e^{\frac{h(y)-h(x)}{T_n}}\right\}. \tag{3.6}$$

3. Update T_n to T_{n+1}.

4. $X_{n+1} \mapsto x_n$ and return to 1.

■

The definition (3.2) constructs a Markov chain, which is not homogeneous. The sequence T_n is decreasing to zero,

3.2 Optimization

The function $h(x)$ can be thought of as a criterion, which we want to optimize as a function of x. In statistical physics this is the energy function to be minimized. We shall, however, think of maximization, minimization can be handled by maximizing $-h(x)$.

We suppose that $h(x)$ is a function that cannot be optimized by straightforward means, or that contains a very large number of variables, and that may have many local maxima.

But from the aforementioned first properties of SA we see that proposals with $h(y) < h(x)$, y can be accepted with probability $e^{\frac{h(y)-h(x)}{T_n}}$, and therefore the algorithm is allowed to escape local maxima. The probability of this escape depends on the scaling T_n. As T_n decreases to zero, the values simulated become more and more concentrated around the local maxima.

4 Exercises

Corollary 4.1 Let $\pi(\theta)$ be a positive density on on a closed and bounded set Θ. If there exists a unique maximum likelihood estimator θ^*, it satisfies then

$$\lim_{\lambda\to\infty} \frac{\int x e^{\lambda l(\theta|x)} \pi(\theta)\, d\theta}{\int e^{\lambda l(\theta|x)} \pi(\theta)\, d\theta} = \theta^*. \tag{4.1}$$

■

5 Appendix A: Proof of proposition 2.1

We follow Pincus (1968) and give first a lemma.

Lemma 5.1 For any $\epsilon > 0$, set

$$N_\epsilon(x^*) = \{x \in \mathcal{X} \mid \|x - x^*\| < \epsilon\} \tag{A.1}$$

Then it there exists $\delta > 0$ such that

$$\max_{x\in\mathcal{X}\setminus N_\epsilon(x^*)} \mid h(x) - h(x^*) \mid < -\delta. \tag{A.2}$$

Proof pf proposition 2.1: We prove the asserion by contradiction. Assume that

$$\max_{x\in\mathcal{X}\setminus N_\epsilon(x^*)} \mid h(x) - h(x^*) \mid = 0.$$

Since $\mathcal{X} \setminus N_\epsilon(x^*)$ is a bounded set in $\mathcal{R}^n$, and $h(x)$ is continuous, there exists then a sequence $\{x_i\}$ of points in $\mathcal{X} \setminus N_\epsilon(x^*)$ such that

$$\lim_{i\to\infty} \mid h(x_i) - h(x^*) \mid \to 0,$$

Since $\mathcal{X} \setminus N_\epsilon(x^*)$ is compact ($N_\epsilon(x^*)$ is open), there exists a subsequence $\{x_{i_j}\}$ that converges. Let

$$\lim_{j\to\infty} x_{i_j} = y.$$

Then $y \in \mathcal{X} \setminus N_\epsilon(x^*)$, and by continuity we have

$$\lim_{j\to\infty} h\left(x_{i_j}\right) = h(y) = h(x^*).$$

Since $h(x)$ attains by assumtion a unique global maximum at x^*, it follows that $x^* = y$. This is a contradiction, since $y \in\in \mathcal{X} \setminus N_\epsilon(x^*)$. ∎

Then we prove the proposition, i.e., the limit in (2.2). It sufficse to prove

$$\lim_{\lambda\to\infty} \frac{\int_{\mathcal{X}} |x_k - x_k^* \mid e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx} = 0, \qquad k = 1, \ldots, n. \tag{A.3}$$

The quantity in the left hand side can be written as

$$\frac{\int_{N_\epsilon(x^*)} \mid x_k - x_k^* \mid e^{\lambda h(x)} dx + \int_{\mathcal{X}\setminus N_\epsilon(x^*)} |x_k - x_k^* \mid e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx}$$

By definition of $N_\epsilon(x^*)$ we have

$$\frac{\int_{N_\epsilon(x^*)} \mid x_k - x_k^* \mid e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx}$$

$$< \frac{\int_{N_\epsilon(x^*)} \epsilon e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx}$$

$$< \epsilon \frac{\int_{\mathcal{X}} \epsilon e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx} = \epsilon.$$

We let

$$M = \max_{\mathcal{X}} \mid x_k - x_k^* \mid,$$

and let $V(B)$ designate the n-dimensional volume of B. Then

$$\frac{\int_{\mathcal{X}\setminus N_\epsilon(x^*)} |x_k - x_k^* \mid e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx}$$

$$< \frac{M \int_{\mathcal{X}\setminus N_\epsilon(x^*)} e^{\lambda(h(x)-h(x^*))} dx}{\int_{\mathcal{X}} e^{\lambda(h(x)-h(x^*))} dx}$$

$$< \frac{MV\left(\mathcal{X} \setminus N_\epsilon\left(x^*\right)\right) e^{-\lambda\delta}}{\int_{\mathcal{X}} e^{\lambda(h(x)-h(x^*))} dx}, \tag{A.4}$$

where we invoked the lemma above.

Since $h(x)$ is continuous we can choose $\eta > 0$ so that if $N_\eta\left(x^*\right) = \{x \in \mathcal{X} \mid \|x - x^*\| < \eta\}$,

$$e^{-\lambda\delta/2} \leq e^{\lambda(h(x)-h(x^*))}, \quad x \in N_\eta\left(x^*\right).$$

Therefore we have in the denominator

$$\int_{\mathcal{X}} e^{\lambda(h(x)-h(x^*))} dx \geq \int_{N_\eta(x^*)} e^{\lambda(h(x)-h(x^*))} dx$$

$$\geq V\left(N_\eta\left(x^*\right)\right) e^{-\lambda\delta/2}.$$

Combining the last inequality with (A.4) we get

$$\frac{\int_{\mathcal{X}\setminus N_\epsilon(x^*)} |x_k - x_k^* \mid e^{\lambda h(x)} dx}{\int_{\mathcal{X}} e^{\lambda h(x)} dx}$$

$$\leq MV\left(\mathcal{X} \setminus N_\epsilon\left(x^*\right)\right) e^{-\lambda\delta/2} / V\left(N_\eta\left(x^*\right)\right).$$

Hence we obtained the claim in the proposition, as asserted. ∎

6 References and further reading:

1 Journal articles and technical reports:

- S.P. Brooks, N. Friel, and R. King (2003): Classical model selection via simulated annealing. *Journal of the Royal Statistical Society, B*, 65, pp. 503−520.
- B. Hájek (1988): Cooling Schedules for Optimal Annealing. *Mathematics of Operations Research*, 13, pp. 311−329.
- S. Kirkpatrick, C.D. Gelatt J.r., and M.P. Vecchi (1983): Optimization by Simulated Annealing. *Science*, 220, pp. 671−680.
- M. Pincus (1968): A Closed Form Solution of Certain Programming Problems. *Operations Research*, 16, pp. 690−694.
- M. Pincus (1970): A Monte Carlo Method for the Approximate Solution of Certain Types of Constained Optimization Problems. *Operations Research*, 18, pp. 1225−1228.

- C-R. Hwang (1980): Laplace's method revisited: Weak Convergence of Probability Measures. *The Annals of Probability*, 8, pp. 1177−1182.

2 Books:

- J.M. Hammersley & Handscomb (1964): *Monte Carlo Methods*. Methuen & Co. LTD, London.
- O. Häggström (2002): *Finite Markov Chains and Algorithmic Applications*. London Mathematical Society, Student Texts, Cambridge University Press, Cambridge.
- P.J.M. van Laarhoven and E.H.L. Aarts (1987): *Simulated Annealing: Theory and Applications.*, Kluwer.
- C.P. Robert and G. Casella (1999): *Monte Carlo Statistical Methods*. Springer Verlag, New York.