Statistical Bioinformatics, Makerere
MCMC and Simulated Annealing

Timo Koski

TK

02.08.2018

KTH Matematik

Contents

KTH Matematik

Markov Chain Monte Carlo

Markov chain Monte Carlo (MCMC) is an important computational tool in Bayesian statistics, since it allows inferences to be drawn from complex posterior distributions, where analytical or numerical integration techniques cannot be applied.

KTH Matematik

Markov Chain Monte Carlo

The idea is is to generate a Markov chain via iterative Monte Carlo simulation that has, at least in the asymptotic sense, the desired posterior distribution as its invariant or stationary distribution.

KTH Matematik

Markov Chain Monte Carlo

Since direct sampling from a posterior distribution may not be possible, the Metropolis-Hastings algorithm starts by generating candidate draws from a so-called proposal distribution. These draws are the corrected so that they behave asymptotically as random observations from the desired invariant or target distribution.

KTH VETENSKAP OCH KONST
KTH Matematik

Markov Chain Monte Carlo

The MC constructed by the algorithm at each stage is thus built in two steps: a **proposal** step and an **acceptance** step. These two steps are associated with the proposal and acceptance distributions, respectively.

KTH Matematik

Time Reversible Markov Chains

$\{X_n\}_{n\geq 0} \sim \text{Markov}\,(\mathbf{P}, \phi(0))$ with the state space S. A probability distribution **a** on S is said to be **reversible** for the chain (or for the matrix **P**), if for all $i, j \in S \times S$ we have

$$a_i p_{i|j} = a_j p_{j|i}$$

The MC is said to be **reversible**, if there exists a reversible distribution for it.

KTH VETENSKAP OCH KONST
KTH Matematik

Time Reversible Markov Chains

$\{X_n\}_{n\geq 0} \sim \text{Markov}\left(\mathbf{P}, \phi(0)\right)$ with the state space S. Assume that $\mathbf{a}$ on S is reversible for the chain. Then $\mathbf{a}$ is an invariant distribution for the chain.

KTH Matematik

Time Reversible Markov Chains

Proof:

$$a_i = a_i \sum_{j \in S} p_{i|j} = \sum_{j \in S} a_i p_{i|j},$$

and by above

$$= \sum_{j \in S} a_j p_{j|i},$$

$$\Leftrightarrow$$

$$\mathbf{a} = \mathbf{a}P.$$

Here we regard $\mathbf{a}$ as a $1 \times J$ row vector.

KTH VETENSKAP OCH KONST
KTH Matematik

Metropolis-Hastings Problem

Let $\mathbf{f} = \{f_j\}_{j \in S}$ be an arbitrary probability mass function, **target distribution**, on *a discrete subset S of $\mathcal{R}$*, i.e.,

- $f_j \geq 0$.
- $\sum_{j=0}^{J} f_j = 1$.

The *Metropolis problem* is to give a Markov chain such that $\mathbf{f}$ is its invariant distribution. We shall next show that it is always possible to solve the stated problem by constructing an appropriate transition probability matrix. In fact there are infinitely many solutions to the stated problem.

KTH Matematik

Metropolis-Hastings Solution

$\mathbf{Q} = \left(q_{i|j}\right)_{i=0,j=0}^{J,J}$ is a transition probability matrix. Given that $X_n = i$

1. Generate $Y_{n+1} \sim \{q_{i|j}\}_{j=0}^{J}$.
2. Take

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{with probability } \alpha_{i,Y_{n+1}}^{H} \\ i & \text{with probability } 1 - \alpha_{i,Y_{n+1}}^{H} \end{cases},$$

where

$$\alpha_{i,j}^{H} = \min\left\{1, \frac{f_j q_{j|i}}{f_i q_{i|j}}\right\}.$$

Metropolis-Hastings Solution

This gives a reversible Markov chain with $\mathbf{f}$ as stationary distribution.

Gibbs sampler

Gibbs sampling from lecture *I* is a special case of the Metopolis-Hastings algorithm.

KTH Matematik

Gibbs sampler

Gibbs sampling is an MCMC scheme for generating samples from a distribution of interest when the full distribution is not available but when generating a sample from a full conditional distribution

$$P\left(X_i \mid X_1, \ldots, X_{i-1}, X_{i+1}, \ldots, X_d\right)$$

is possible.

KTH VETENSKAP OCH KONST
KTH Matematik

Gibbs sampler

Sample $X_1^{(j)}, \ldots, X_d^{(j)}$ at iteration j is given, then

$$X_1^{(j+1)} \sim P\left(X_1 \mid X_2^{(j)}, \ldots, X_d^{(j)}\right)$$

$$\vdots$$

$$X_i^{(j+1)} \sim P\left(X_i \mid X_1^{(j)}, \ldots, X_{i-1}^{(j)}, X_{i+1}^{(j)}, \ldots, X_d^{(j)}\right)$$

$$\vdots$$

$$X_d^{(j+1)} \sim P\left(X_d \mid X_1^{(j)}, \ldots, X_{d-1}^{(j)}\right)$$

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

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.

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

A 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 objective function as 'temperature' is lowered.
The aforementioned indicates a deep connection between statistical physics and optimization of functions of many variables.

KTH Matematik

Simulated Annealing

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

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

then

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

KTH Matematik

Simulated Annealing

The preceding says in fact that for large λ the major contribution to the integral above comes from a small neighbourhood of x^*. Hence a Monte Carlo method that generates samples that spend, in the long run, most of the time visiting states near the maximizing point could find x^*.

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

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

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

for the purpose of calculating the expectation

$$I_\lambda = \int_{\mathcal{X}} x f_\lambda(x) dx$$

by a Monte Carlo method.

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

We make a digitalized variant of this. We us partition the region $\mathcal{X}$ into a finite number of N mutually disjoint subsets $\mathcal{X}_j$. We fix points $y^j \in \mathcal{X}_j$. Then we construct 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)

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

where the subsets $\mathcal{X}_j$ have been assumed to have equal volumes.

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

Then, the strong law of large numbers for the Markov chain $\{X_n\}_{n\geq 0}$ gives

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} X_i = l_\lambda$$

and for large values of λ we have that

$$l_\lambda \approx x^*.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

The question is, how to choose λ in the implementation. There are a couple of opportunities:

1) Run the Metropolis MC for several values of λ and choose the best value of optimizer found.
2) Develop a schedule for changing λ as a function of n.
3) Hybridize the two preceding by running with a fixed λ for a time and then change. This is the strategy most commonly used in practice, but is difficult to couch in other than pragmatic terms.

KTH Matematik

Simulated Annealing

Statistical mechanics studies physical systems consisting of a large number of identical components, such as molecules in a gas. This can be justified in two ways. Either chance plays a fundamental part in laws of physics, or the laws are deterministic, but it is impossible to collect enough data to be able to apply them deterministically, whereby probability covers up ignorance.
If a classical system is in thermal equilibrium with its surroundings, and is in state x with energy $h(x)$, then the probability density in the phase space of th point representing x is proportional to

$$e^{-\frac{1}{kT}h(x)}$$

where T is the absolute temperature, and k is Boltzmann's constant. According to ergodic theory the proportion of time that the system spends in state x is also proportional to $e^{-\frac{1}{kT}h(x)}$.

KTH Matematik

Simulated Annealing

Then the probability of acceptance becomes

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

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 $e^{\frac{h(y)-h(x)}{T}}$. The acceptance probabilities depend on the scale T.

KTH Matematik

Simulated Annealing

Let us assume that we want to optimize a function $h(i)$ on a finite set of states, now also called 'configurations', $i \in \mathcal{S}$. The word configuration is an idiosyncratic phrase from statistical physics. We assume that we have a symmetric transition probability matrix, the **proposal matrix**,

$$\mathbf{Q} = \{q(i|j)\}_{i,j \in S \times S}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

Given that $X_n = x_n$

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

$$\rho_n(i,j) = \min\left\{1, e^{\frac{h(i)-h(j)}{T_n}}\right\}.$$

3. Update T_n to T_{n+1}.
4. $X_{n+1} \mapsto x_n$ and return to 1.

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

The proposals $Y_{n+1} = j$ with $h(j) < h(i)$, can be accepted with probability $e^{\frac{h(i)-h(j)}{T_n}}$, and therefore the algorithm is allowed to escape local maxima. The probability of this escape depends on the scaling T_n. If T_n decreases to zero with increasing n, the values simulated become more and more concentrated around the maxima of h.

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

By the preceding lectures we know that the one-step transition probabilities of the Markov chain constructed by the algorithm above are, for $i \neq j$,

$$p_{i|j}(n) = q_{i|j} \cdot \rho_n(i,j).$$

Hence the transition probabilities depend on n through the scales T_n , and therefore we have a non-homogeneous Markov chain. (Note that $p_{i|j}(n)$ is no longer the n-step transition probability from the previous lectures.)

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

The matrix

$$\mathbf{æ}(n) = \{\rho_n(i,j)\}_{i,j \in S \times S}$$

is called the **acceptance matrix**.
We need a symbol for the set of globally maximal configurations of h, and set

$$S_{\text{opt}} = \{i \in \mathcal{S} | h(i) = \max_{x \in \mathcal{S}} h(x)\}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

We say that such **a simulated annealing algorithm obtains asymptotically a global maximum**, if

$$\lim_{k \to \infty} P\left(X_k \in S_{\text{opt}}\right) = 1.$$

The necessary and sufficient conditions on the acceptance and proposal matrices for this convergence to be true are surveyed in (van Laarhoven and Aarts 1987).

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing

T as a decreasing function of n, denoted by T_n,

$$T_1 \geq T_2 \geq \ldots > T_n \geq \ldots.$$

An idiosyncratic name for this is **cooling schedule**.

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

where c is a positive constant. It turns out that c will have to satisfy an additional condition.

KTH VETENSKAP OCH KONST
KTH Matematik

Non-homogeneous Markov Chains

We start by giving a notation for the inhomogeneous multi step transition probabilities for a Markov chain $\{X_n\}_{n\geq 0}$ with the finite state space S,

$$p_{i|j}(m,k) = P\left(X_k = j | X_m = i\right).$$

We designate the corresponding transition matrix by

$$\mathbf{P}(m,k) = \{p_{i|j}(m,k)\}_{i,j\in S\times S}.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Non-homogeneous Markov Chains

An inhomogeneous Markov chain $\{X_n\}_{n\geq 0}$ is **weakly ergodic**, if for all $m \geq 1$, and all $i, j, l \in S \times S \times S$

$$\lim_{k\to\infty} \left(p_{i|l}(m,k) - p_{j|l}(m,k)\right) = 0.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Non-homogeneous Markov Chains

An inhomogeneous Markov chain $\{X_n\}_{n\geq 0}$ is **strongly ergodic**, if there exists a vector $\pi = (\pi_i)_{i\in S}$ satisfying

$$\sum_{i\in S} \pi_i = 1, \pi_i \geq 0,$$

such that for all $m \geq 1$, and all $i, j, \in S \times S$

$$\lim_{k\to\infty} p_{i|j}(m,k) = \pi_j.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Non-homogeneous Markov Chains

Strong ergodicity implies convergence in distribution, and

$$\lim_{k \to \infty} P\left(X_k = j\right) = \pi_j.$$

KTH Matematik

Non-homogeneous Markov Chains

An inhomogeneous Markov chain is strongly ergodic, if it is weakly ergodic and if for all k there exists a vector $\pi(k)$ such that $\pi(k)$ is an eigenvector corresponding to the eigenvalue 1 of $\mathbf{P}(k-1,k)$, $\sum_{i\in S}\pi_i(k)=1$, and

$$\sum_{k=0}^{\infty}\sum_{i\in S} \mid \pi_i(k)-\pi_i(k+1) \mid < \infty.$$

If

$$\pi = \lim_{k\to\infty} \pi(k),$$

then π satisfies

$$\lim_{k\to\infty} p_{i|j}(m,k) = \pi_j.$$

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing for HMM

Bayesian learning for HMM can be done by simulated annealing.

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing for HMM

Bayesian learning for HMM can be done by simulated annealing.

$$P\left(\lambda \mid Y_0 = o_0, \ldots, Y_n = o_n; \right) =$$

$$\frac{P\left(Y_0 = o_0, \ldots, Y_n = o_n; \lambda\right) p\left(\lambda\right)}{\int P\left(Y_0 = o_0, \ldots, Y_n = o_n; \lambda\right) p\left(\lambda\right) d\lambda}$$

KTH VETENSKAP OCH KONST
KTH Matematik

Simulated Annealing for HMM

We do not need to compute

$$\int P\left(Y_0 = o_0, \ldots, Y_n = o_n; \lambda\right) p\left(\lambda\right) d\lambda$$

KTH VETENSKAP OCH KONST
KTH Matematik

SA for learning HMM

C. Andrieu & A.Doucet (2000): Simulated Annealing for Maximum *A Posteriori* Parameter Estimation of Hidden Markov Models, *IEEE Transactions on Information Theory* , 46, pp. 994−1002.

KTH Matematik

SA for learning HMM

Andrieu & Doucet construct an inhomogeneous MC whose sequence of invariant distributions concentrates
on the set of MAP parameters. The algorithm is done so that the structure of HMM is taken into account.

KTH Matematik

SA with Data Augmenting

$$\mathbf{x} = j_0 \dots j_n, \mathbf{y} = o_0 \dots o_n$$

1) Initialization. $\left(\lambda^{(0)}, \mathbf{x}^{(0)}\right)$ randomly.
2) Iteration n $n \geq 1$.
   - sample $\mathbf{x}^{(n)}$ from $P\left(\mathbf{x}^{(*)} \mid \mathbf{y}; \lambda^{(n-1)}\right)$
   - sample λ^{pr} from $P\left(\lambda \mid \mathbf{y}; \mathbf{x}^{(*)}\right)$

KTH VETENSKAP OCH KONST
KTH Matematik

SA with Data Augmenting

- sample $\mathbf{x}^{(n)}$ from $P\left(\mathbf{x}^{(*)} \mid \mathbf{y}; \lambda^{(n-1)}\right)$
- sample λ^{pr} from $P\left(\lambda \mid \mathbf{y}; \mathbf{x}^{(*)}\right)$

$P\left(\mathbf{x}^{(*)} \mid \mathbf{y}; \lambda^{(n-1)}\right)$ can be calculated by a 'forward filtering-backward recursion'

KTH VETENSKAP OCH KONST
KTH Matematik

SA with Data Augmenting

Evaluate the acceptance probability

$$\alpha_n\left(\lambda^{pr}, \lambda^{(n-1)}\right) = \min\left\{1, \left[\frac{p\left(\lambda^{pr} \mid \mathbf{y}\right)}{p\left(\lambda^{(n-1)} \mid \mathbf{y}\right)}\right]^{1/(T_n+1)}\right\}$$

KTH Matematik

SA with Data Augmenting

Take a uniform random number U,if

$$U \leq \alpha_n \left(\lambda^{pr}, \lambda^{(n-1)} \right)$$

then $\lambda^{(n)} = \lambda^{pr}$, otherwise $\lambda^{(n)} = \lambda^{(n-1)}$

3) Set $n = n + 1$ and go to step 2).

KTH VETENSKAP OCH KONST
KTH Matematik