Markov Chain Monte Carlo (MCMC), Gibbs Sampling, Metropolis Algorithms, and Simulated Annealing

2001 Bioinformatics Course

Supplement

SNU Biointelligence Lab

http://bi.snu.ac.kr/

Outline

- Markov Chain Monte Carlo (MCMC)
- Metropolis-Hastings Algorithm
- Metropolis Algorithm
- Gibbs Sampling
- Simulated Annealing

Introduction

- Markov Chain Monte Carlo
  - Monte Carlo integration using Markov chains
  - *Monte Carlo* integration draw samples from the required distribution, and then forms sample averages to approximate expectations.
  - *Markov chain* Monte Carlo draws samples by running a cleverly constructed Markov chain for a long time.
  - MCMC is usually used for Bayesian inference.

Bayesian Inference (1)

- Bayesian inference
  - Most applications of MCMC are oriented.
  - D : observed data, θ : model parameters
  - $P(\theta)$: prior distribution, $P(D \mid \theta)$: likelihood
  - Full probability model

$$P(D, \theta) = P(D \mid \theta)\, P(\theta)$$

  - Posterior distribution of θ: $P(D|\theta)$
    - Having observed D, Bayes theorem is used.
    - Object of all Bayesian inference

$$P(\theta \mid D) = \frac{P(\theta)P(D \mid \theta)}{\int P(\theta)P(D \mid \theta)d\theta}$$

Bayesian Inference (2)

- Any features of the posterior distribution are legitimate for Bayesian inference.
  - The posterior expectation of a function $f(\theta)$:

$$E[f(\theta) \mid D] = \frac{\int f(\theta)P(\theta)P(D \mid \theta)d\theta}{\int P(\theta)P(D \mid \theta)d\theta}$$

- Difficulty
  - Integration, especially in high dimensions is impossible.
  - Numerical evaluation is also difficult and inaccurate.
  - Analytic approximation: Laplace approximation, Monte Carlo integration (MCMC)

Calculating Expectations (1)

- Terms
  - X
    - A vector of k random variables with distribution $\pi(.)$
    - In Bayesian applications, X will comprise model parameters θ
  - $\pi(.)$
    - Posterior distribution for Bayesians, i.e. $P(\theta|D) = P(D|\theta)P(\theta) / P(D)$
    - Likelihood for frequentists, i.e. $P(D|\theta)$
  - Task
    - Evaluate the expectation for some function $f(.)$

$$E[f(X)] = \frac{\int f(x)\pi(x)dx}{\int \pi(x)dx}$$

Calculating Expectations (2)

- Problem
  - $\int \pi(x)dx$ is unknown
  - Generality of X
    - X takes values in k-dimensional Euclidean space.
    - Discrete random variables
    - Mixture of discrete and continuous random variables
    - k can itself be variable.

Monte Carlo Integration (1)

- Drawing samples $\{X_t, t=1, \dots, n\}$ from $\pi(.)$
- Approximating

$$E[f(X)] \approx \frac{1}{n}\sum_{t=1}^{n} f(X_t)$$

- When the samples $\{X_t\}$ are independent, laws of large numbers ensure that the approximation can be made as accurate as desired by increasing the sample size n.

Monte Carlo Integration (2)

- Problem
  - Drawing $\{X_t\}$ independently from $\pi(.)$ is not feasible, since $\pi(.)$ can be quite non-standard.
- $\{X_t\}$ need not necessarily be independent.
  - $\{X_t\}$ can be generated by any process which draws samples throughout the support of $\pi(.)$ in the correct proportions.
  - MCMC
    - One way of doing this is through a Markov chain having $\pi(.)$ as its stationary distribution.

Markov Chain (1)

- Consider a generated sequence $\{X_0, X_1, X_2, \ldots\}$
- X_{t+1} is sampled from a distribution $P(X_{t+1}|X_t)$.
  - *Markov Chain* : this sequence
  - $P(.|.)$: *transition kernel* of the chain
  - Assume that the chain is time-homogenous.

Markov Chain (2)

- Effect of X_0 to X_t : $P^{(t)}(X_t|X_0)$
  - Subject to regularity conditions, the chain will gradually forget its initial state and $P^{(t)}(.|X_0)$ will eventually converge to a unique stationary distribution $\phi(.)$.
  - As t increases, the sampled points $\{X_t\}$ will look increasingly like dependent samples from $\phi(.)$.

Markov Chain (3)

- After a sufficiently long *burn-in* of, say, m iterations, points $\{X_t : t = m+1, \ldots, n\}$ will be dependent samples approximately from $\phi(.)$.
- Use the output from the Markov chain to estimate $E[f(X)]$, where X has distribution $\phi(.)$.
- ***Ergodic average***

$$\bar{f} = \frac{1}{n-m} \sum_{t=m+1}^{n} f(X_t)$$

The Metropolis-Hastings Algorithm (1)

- How to construct a Markov chain such that its stationary distribution $\phi(.)$ is precisely our distribution of interest $\pi(.)$?
  - At each time t, the next state X_{t+1} is chosen by first sampling *candidate* point Y from a *proposal* distribution $q(.|X_t)$.
  - The candidate point Y is then accepted with probability $\alpha(X_t, Y)$.

$$\alpha(X,Y) = \min\left(1, \frac{\pi(Y)q(X \mid Y)}{\pi(X)q(Y \mid X)}\right)$$

  - If the candidate point is accepted, the next state becomes $X_{t+1}=Y$.
  - If the candidate is rejected, the chain does not move, i.e. $X_{t+1}=X_t$.

The Metropolis-Hastings Algorithm (2)

Initialize X_0 ; set $t = 0$
Repeat {
  Sample a point Y **from** $q(.|X_t)$
  Sample a Uniform(0,1) random variable U
  If $U \leq \alpha(X_t, Y)$ **set** $X_{t+1} = Y$
  Otherwise set $X_{t+1} = X_t$
  Increment t
}

Implementation Issues

- Canonical forms of proposal distribution
  - Any proposal distribution will ultimately deliver samples from the target distribution $\pi(.)$.
  - However, the rate of convergence to the stationary distribution will depend crucially on the relationship between $q(.|.)$ and $\pi(.)$.
  - For computational efficiency, $q(.|.)$ should be chosen so that it can be easily sampled and evaluated.

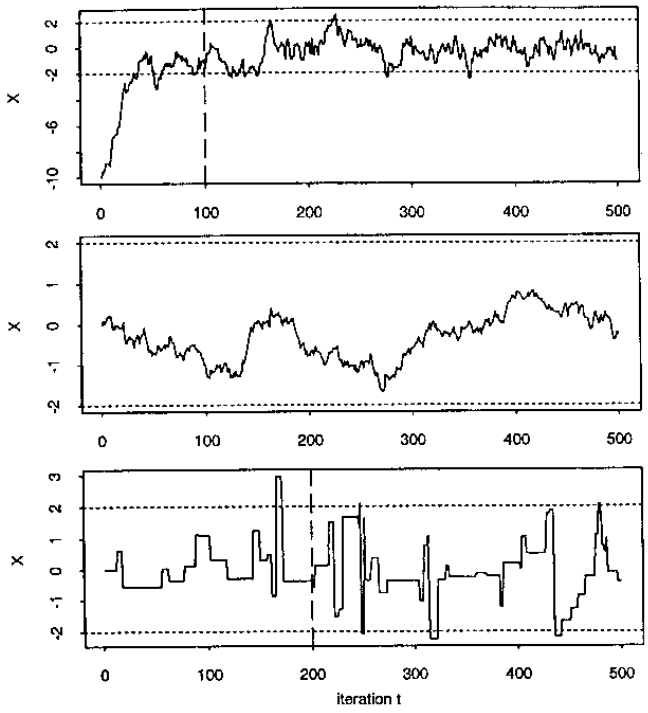
Metropolis Algorithm

- Symmetric proposals $q(Y|X) = q(X|Y)$ which generates Y conditionally independently, given X_t.

$$\alpha(X, Y) = \min\left(1, \frac{\pi(Y)}{\pi(X)}\right)$$

  - Random-walk Metropolis: $q(Y|X)=q(|X-Y|)$
- Scale of a proposal distribution may need to be chosen carefully.
  - A cautious proposal distribution generating small steps
  - A bold proposal distribution generating large steps
  - Avoid both these extremes.

Stationary distribution N(X,1)



$q(.|X) = \mathrm{N}(X, 0.5)$

$q(.|X) = \mathrm{N}(X, 0.1)$

$q(.|X) = \mathrm{N}(X, 10.0)$

Gibbs Sampling (1)

- *Single-Component Metropolis-Hastings*
  - Dividing X into components $\{X_{.1}, X_{.2}, \ldots, X_{.h}\}$ of possibly differing dimension, and then updating component one by one
    - An iteration of the single-component Metropolis-Hastings algorithm comprises h updating steps.
    - The i^{th} proposal distribution $q_i(.|.,.)$ generates a candidate only for the i^{th} component of X, and may depend on the *current* values of any of the components of X.

$$\alpha(X_{.-i}, X_{.i}, Y_{.i}) = \min\left(1, \frac{\pi(Y_{.i} \mid X_{.i}) q_i(X_{.i} \mid Y_{.i}, X_{.-i})}{\pi(X_{.i} \mid X_{.-i}) q_i(Y_{.i} \mid X_{.i}, X_{.-i})}\right)$$

Gibbs Sampling (2)

- Gibbs Sampling
  - A special case of single-component Metropolis-Hastings
  - Most statistical applications of MCMC have used Gibbs sampling.

$$q_i(Y_{.i} \mid X_{.i}, X_{.-i}) = \pi(Y_{.i} \mid X_{.-i})$$

  - Acceptance probability is 1; that is, Gibbs sampler candidates are always accepted.

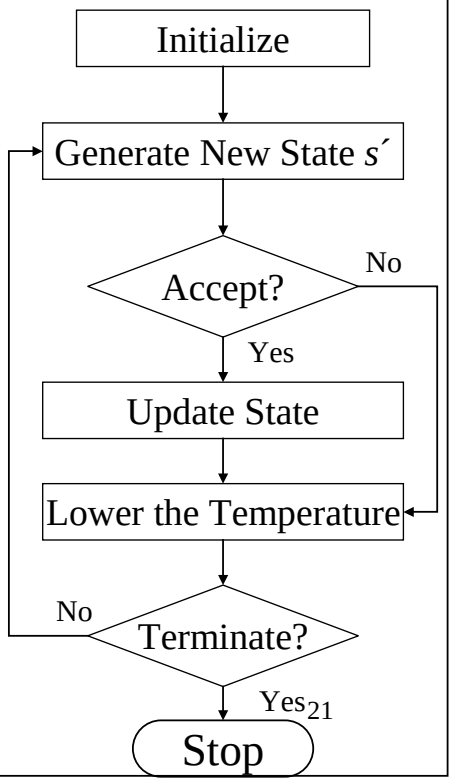
Simulated Annealing (1)

- Statistical Mechanics
  - Simulated annealing (SA) exploits an analogy between the way in which a metal cools and freezes into a minimum energy crystalline structure (the annealing process) and the search for a minimum in a more general system.
- Boltzmann-Gibbs Distribution
  - The probability of being in state *s* at temperature *T*

$$P(s) = P(x_1, \ldots, x_n) = \frac{e^{-f(s)/kT}}{Z}$$

Simulated Annealing (2)

- The implementation of the SA algorithm
  - ♦ Representation of possible solutions
  - ♦ Generator of random changes in solutions
  - ♦ Means of evaluating the problem functions
  - ♦ *Annealing schedule*: an initial temperature and rules for lowering it as the search progresses.



Simulated Annealing (3)

- Advantages
  - ♦ SA can reach the Boltzmann-Gibbs equilibrium distribution in a reasonable time, while any MCMC method fails in general.
  - ♦ SA's another advantage over other methods is an ability to avoid becoming trapped at local optimum.
  - ♦ While simulated annealing is usually used in combination with the Metropolis algorithm, it is in fact applicable to any MCMC method, and in particular Gibbs sampling.