MCMC: Markov Chain Monte Carlo

Yuzhen Ye
School of Informatics and Computing
Indiana University, Bloomington
Spring 2013

Contents
- Review of Markov Chains
- Monte Carlo simulation
- Introduction of MCMC
  - Motivating problems
  - MCMC updating schemes
- Practical implementation issues
  - The choice of transition mechanism for the chain
  - The number of chains to be run and their length
- MCMC algorithms & applications
  - Gibbs sampler (motif finding problem)

Review: 1st order Markov chain

An integer time stochastic process, consisting of a set of m>1 states \{s_1,...,s_m\} and
1. An m-dimensional initial distribution vector \( (p(s_1),...,p(s_m)) \)
2. An \( m \times m \) transition probabilities matrix \( M = (a_{ij}) \)

For example, for DNA sequence:
the states are \{A, C, T, G\} (m=4)
p(A) the probability of A to be the 1st letter
a_{AG} the probability that G follows A in a sequence.

Motivating problems for MCMC

- The integration operation that plays a fundamental role in Bayesian statistics
  - For calculating the normalizing constant
  - Marginal distribution
  - Expectation
- MCMC, first introduced by Metropolis (1953), provides an alternative whereby we sample from the posterior directly, and obtain sample estimates of the quantities of interest

Sampling and optimization

- To maximize a function, \( f(x) \):
  - Brute force method: try all possible \( x \)
  - Sample method: sample \( x \) from probability distribution: \( p(x) \sim f(x) \)
  - Idea: suppose \( x_{max} \) is a maximum of \( f(x) \), then it is also maximum of \( p(x) \), thus we have a high probability of sampling \( x_{max} \)

Monte Carlo simulation

- The idea of Monte Carlo simulation is to draw an i.i.d set of N samples from a target density \( p(x) \) defined on a high-dimensional space \( X \).

\[
I(f) = \frac{1}{N} \sum_{i=1}^{N} f(x^{(i)}) \approx \int f(x)p(x)dx
\]

- The N samples can also be used to obtain a maximum of the objective function \( p(x) \)
Rejection sampling
- Sample from a distribution \( p(x) \), which is known up to a proportionality constant, by sampling from another easy-to-sample proposal distribution \( q(x) \) that satisfies \( p(x) \leq Mq(x) \), using an accept/reject procedure.

Ref: An introduction to MCMC for Machine Learning, 2003

MCMC: basics
- Any Markov chain which is irreducible and aperiodic will have a unique stationary distribution.
  - Irreducibility: from any state of the Markov chain, there is a positive probability of visiting all other states (i.e., the transition matrix cannot be reduced to separate smaller matrices).
  - Aperiodicity: the chain should not get trapped in cycles
- From any starting point, the chain will converge to the invariant distribution \( p(x) \), as long as \( T \) is a stochastic transition matrix that have the two properties: irreducibility & aperiodicity.
- MCMC samplers are irreducible and aperiodic Markov chains that have the target distribution as the invariant distribution.

MCMC algorithms
When samples cannot be drawn from \( p(x) \) directly but \( p(x) \) can be evaluated up to a normalizing constant, MCMC can be used, which is a strategy for generating samples \( x \) while exploring the state space \( X \) using a Markov chain mechanism.

\[
T = \begin{bmatrix}
0 & 1 & 0 \\
0 & 0.1 & 0.9 \\
0.5 & 0.4 & 0
\end{bmatrix}
\]

\( p(x|T) = (0.5, 0.2, 0.3) \)

\( p(x|T)^T \) converges to \( p(x) = (0.2, 0.4, 0.4) \)

MCMC approaches
- The Metropolis-Hastings (MH) algorithm
  - The MH algorithm is the most popular MCMC method
  - Most practical MCMC algorithms can be interpreted as special cases or extensions of this algorithm
- Simulated annealing for global optimization
- Mixtures and cycles of MCMC kernels
  - It is possible to combine several samplers into mixtures and cycles of the individual samplers
- The Gibbs sampler

The motif finding problem
- Given a set of DNA sequences:
  
  ```
  cctgatagctttcctggatccgactgttagttttctctgcgggtatgttaattggtcaacct
  agtcattgggctgttctgatagctttcctttgctagtcgaactgtgagaacgttaaagagctgtaatc
  aaagttacagcgcccttttttctctgcgggtatgttaattggtcaacct
  agtcattgggctgttctgatagctttcctttgctagtcgaactgtgagaacgttaaagagctgtaatc
  cttatcataaaccgctcctcgggtctgctctgggtatgttaattggtcaacct
  ```
- Find the motif in each of the individual sequences

- If starting positions \( \mathbf{s} = (s_1, s_2, \ldots, s_t) \) are given, finding consensus is easy because we can simply construct (and evaluate) the profile to find the motif.
- But... the starting positions \( \mathbf{s} \) are usually not given. How can we find the "best" profile matrix?
  - Gibbs sampling
  - Expectation-Maximization algorithm
### Notations
- Set of symbols: $\Sigma$
- Sequences: $S = (S_1, S_2, ..., S_n)$
- Starting positions of motifs: $A = \{a_1, a_2, ..., a_n\}$
- Motif model ($\theta$): $q_j = P(\text{symbol at the } i\text{-th position} = j)$
- Background model ($\theta_0$): $p_j = P(\text{symbol} = j)$
- Count of symbols in each column: $c_j^i = \text{count of symbol } j \text{ in the } i\text{-th column in the aligned motif instances}$

### Equivalent scoring function
- Maximize the log-odds ratio:
  $$F = \log \frac{P(S | A, \theta)}{P(S | A, \theta_0)} = \sum_{i=1}^{n} \sum_{j=1}^{m} c_j^i \log \frac{q_j}{p_j}$$

### Motif finding problem
- Problem: find starting positions and model parameters simultaneously to maximize the posterior probability:
  $$\max_{\theta, A} P(\theta, A | S)$$
- This is equivalent to maximizing the likelihood by Bayes’ Theorem, assuming a uniform prior distribution over different models:
  $$\max_{\theta, A} P(S | A, \theta)$$

### Gibbs sampling
- Idea: a joint distribution in a high dimension may be hard to sample from, but it may be easy to sample from the conditional distributions where all variables are fixed except one
- To sample from $p(x_1, x_2, ..., x_n)$, let each state of the Markov chain represent $(x_1, x_2, ..., x_n)$, the probability of moving to a state $(x_1, x_2, ..., x_n)$ is: $p(x_1, x_2, ..., x_n)$. It is an algorithm in a class of sampling techniques called Markov Chain Monte Carlo (MCMC) method.

### Gibbs sampling
- Initialization, t=0: sample $(x_1^{(0)}, x_2^{(0)}, x_3^{(0)}, ..., x_n^{(0)})$
- $x_1^{(t+1)} \sim P(1 | x_1^{(t)}, x_2^{(t)}, ..., x_n^{(t)})$
- $x_2^{(t+1)} \sim P(2 | x_1^{(t+1)}, x_2^{(t)}, ..., x_n^{(t)})$
- $x_3^{(t+1)} \sim P(3 | x_1^{(t+1)}, x_2^{(t+1)}, x_3^{(t)}, ..., x_n^{(t)})$
- $\ldots$
- $x_n^{(t+1)} \sim P(n | x_1^{(t+1)}, x_2^{(t+1)}, ..., x_n^{(t+1)})$
Estimator of $\theta$

- Given an alignment $A$, i.e. the starting positions of motifs, $\theta$ can be estimated by its MLE with prior probabilities (e.g. Dirichlet prior with parameter $b_j$):

$$q_{ij} = \frac{c_{ij} + b_j}{N - 1 + B}$$

where $B = \Sigma b_j$

Finding TF binding sites

Sampling motifs on trees

Using both the overpresentation property and the evolutionary conservation property of motifs

Implementation: Initialization

**Initialization**

- **Parameters** are sampled using prior distributions;
- Motif instances in current species are sampled from sequences directly for each current species;
- Motif instances in ancestral species are randomly assigned with one of its immediate child motif instances.

Motif instance updating

- Updating motif instances in ancestral species
- Updating motif instances in current species
Motif instance updating

Updating motif instances for current species

Parameter sampling step

- Metropolis-Hasting algorithm is used to increase or decrease the width $w$ of the motif by 1 from the left or right side

Other applications of Gibbs sampling

- Biclustering microarray data by Gibbs sampling
  - Microarray data is discretized
  - Bioinformatics, 2003
- Assignment of ambiguously mapped reads
  - Bioinformatics, 2010

Practical implementation issues

- How many iterations?
- One run or many?