MCMC and Gibbs Sampling

Sargur Srihari
srihari@cedar.buffalo.edu

Topics

1. Markov Chain Monte Carlo
2. Markov Chains
3. Gibbs Sampling
4. Basic Metropolis Algorithm
5. Metropolis-Hastings Algorithm
6. Slice Sampling

Markov Chain Monte Carlo (MCMC)

- Simple Monte Carlo methods (Rejection sampling and importance sampling) are for evaluating expectations of functions
  - They suffer from severe limitations, particularly with high dimensionality
- MCMC is a very general and powerful framework
  - Markov refers to sequence of samples rather than the model being Markovian
  - Allows sampling from large class of distributions
  - Scales well with dimensionality
  - MCMC origin is in statistical physics (Metropolis 1949)

Markov chains

- First order Markov chain is a sequence of random variables $z^{(1)},\ldots,z^{(M)}$ such that
  - conditional independence property holds:

$$\boxed{p(z^{(m+1)}|\, z^{(1)},\ldots,\, z^{(m)}) = p(z^{(m+1)}|\, z^{(m)})}$$

Each sample dependent only on previous sample

- Represented in a directed graph as a chain
- Markov chain specified by
  - Distribution of initial variable $p(z^{(0)})$
  - Conditional (transition) probabilities

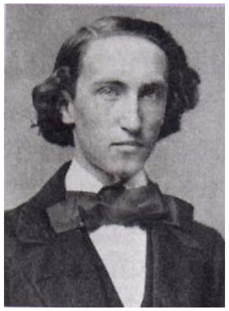
$$T_m(z^{(m)},z^{(m+1)}) = p(z^{(m+1)}|z^{(m)})$$

- Markov chain is homogeneous if all transition probabilities are the same for all m

Gibbs Sampling

- A simple and widely applicable MCMC algorithm
  - Special case of Metropolis-Hastings
- Consider distribution $p(\mathrm{z})=p(z_1,..,z_M)$ from which we wish to sample
- We have chosen an initial state for the Markov chain
- Each step involves replacing value of one variable by a value drawn from $p(z_i|\mathrm{z}_{\backslash i})$

  where symbol $\mathrm{z}_{\backslash i}$ denotes $z_1,..,z_M$ with z_i omitted
- Repeat procedure by cycling through variables in some particular order



Josiah Willard Gibbs
1839-1903
Born New Haven CT
First US PhD in Engg.
Developed
vector analysis,
crystallography and
planetary orbits

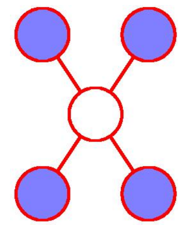
Gibbs with Three Variables

- Distribution $p(z_1,z_2,z_3)$ over three variables
- At step t selected values are $z_1^{(t)}$, $z_2^{(t)}$ and $z_3^{(t)}$
- Replace $z_1^{(t)}$ by new value $z_1^{(t+1)}$ obtained by sampling from $p(z_1|z_2^{(t)},z_3^{(t)})$
- Replace $z_2^{(t)}$ by value $z_2^{(t+1)}$ by sampling from $p(z_2|z_1^{(t+1)},z_3^{(t)})$
  - New value for z_1 is used straightaway
- Update z_3 with a sample $z_3^{(t+1)}$ drawn from $p(z_3|z_1^{(t+1)},z_2^{(t+1)})$
- Cycle through three variables in turn

Gibbs Sampling with M variables

- Initialize first sample: $\{z_i,\ i=1,..,M\}$
- For $t=1,..,T$, T = no of samples
  - Sample $z_1^{(t+1)} \sim p(z_1|z_2^{(t)},z_3^{(t)},\ldots,z_M^{(t)})$
  - Sample $z_2^{(t+1)} \sim p(z_2|z_1^{(t+1)},z_3^{(t)},\ldots,z_M^{(t)})$
  - .....
  - Sample $z_j^{(t+1)} \sim p(z_j|z_1^{(t+1)},..z_{j-1}^{(t+1)},\ z_{j+1}^{(t)}\ \ldots,\ z_M^{(t)})$
  - .....
  - Sample $z_M^{(t+1)} \sim p(z_M|z_1^{(t+1)},z_2^{(t+1)},\ldots,z_{M-1}^{(t+1)})$
- $p(z_j|z_{-j})$ is called a *full conditional* for variable j

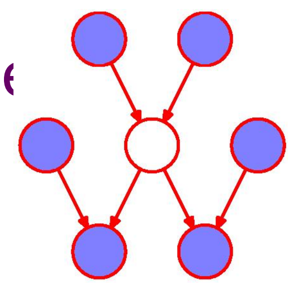
Gibbs sampling and Graphical Models



Markov blanket
for undirected graph

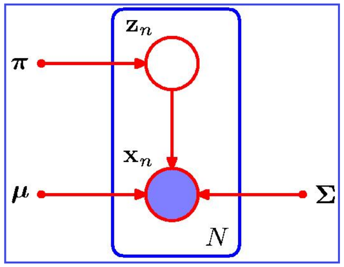
- Practical Applicability of Gibbs sampling
  - depends on ease with which samples can be drawn from $p(z_k|z_{\setminus k})$
  - In the case of PGMs
    - Conditional distributions for nodes depe only on variables in Markov blanket
      - which are its neighbors in the graph
  - Gibbs sampling is a distributed algorithm
    - It is not parallel since samples generated sequentially

Markov blanket
for directed graph



Gibbs Sampling for inferring parameters of GMM

- Full joint distribution



$$\begin{aligned} p(\mathrm{x},\mathrm{z},\pi,\mu,\Sigma) &= p(\mathrm{x}|\mathrm{z},\mu,\pi,\Sigma)p(\mathrm{z},\pi,\mu,\Sigma) \\ &= p(\mathrm{x}|\mathrm{z},\mu,\Sigma)p(\mathrm{z},\pi)p(\mu,\Sigma) \ ; \ given\ \mathrm{z}\ we\ dont\ need\ \pi \\ &= p(\mathrm{x}|\mathrm{z},\mu,\Sigma)p(\mathrm{z}\,|\,\pi)p(\pi)p(\mu)p(\Sigma) \end{aligned}$$

  - Using a semi-conjugate prior

$$\begin{aligned} p(\mathrm{x},\mathrm{z},\pi,\mu,\Sigma) &= p(\mathrm{x}|\mathrm{z},\mu,\Sigma)p(\mathrm{z}\,|\,\pi)p(\pi)\prod_{k=1}^{K} p(\mu_k)p(\Sigma_k) \\ &= \left(\prod_{n=1}^{N}\prod_{k=1}^{K}\left(\pi_k N(\mathrm{x}_n\,|\,\mu_k,\Sigma_k)\right)^{I(z_n=k)}\right)\times Dir\left(\pi\,|\,\alpha\right)\prod_{k=1}^{K} N(\mu_k\,|\,m_0,V_0)IW(\Sigma_k\,|\,S_0,\nu_0) \end{aligned}$$

  - IW is the inverse Wishart distribution

- We need full conditionals (to obtain samples) for
  - Discrete indicators, $p(z_n=1|\mathrm{x}_n,\mu,\pi,\Sigma)$
  - Mixing weights $p(\pi|\mathrm{z})$ for $\pi_1,..\pi_K$
  - Means $p(\mu_k|\Sigma_k,\mathrm{z},\mathrm{x})$
  - Covariances $p(\Sigma_k|\mu_k,\mathrm{z},\mathrm{x})$
- Note that we know the values of x_n

Full Conditionals for Gibbs GMM

- Discrete Indicators

$$\boxed{p(\mathrm{z}_n = k \mid \mathrm{x}_n, \pi, \mu, \Sigma) \propto \pi_k N\left(\mathrm{x}_n \mid \mu_k, \Sigma_k\right)}$$

- Mixing weights

$$\boxed{p(\pi \mid \mathrm{z}) = Dir\left(\left\{\alpha_k + \sum_{n=1}^{N} I\left(\mathrm{z}_n = k\right)_{k=1}^{K}\right\}\right)}$$

- Means

$$\boxed{p(\mu_k \mid \Sigma_k, \mathrm{z}, \mathrm{x}) = N(\mu_k \mid \mathrm{m}_k, V_k)}$$

Terms V_k and S_k are sample statistics

- Covariances

$$\boxed{p(\Sigma_k \mid \mu_k, z, \mathrm{x}) = IW\left(\Sigma_k \mid S_k, \nu_k\right)}$$

Proof of Sampling

- To show that procedure samples from given distribution
- First show that distribution $p(z)$ is invariant (or stationary) of each sampling step and hence of whole Markov chain
- Second requirement is ergodicity
  - Every state reachable from every other state
- The two requirements are formally defined next

Markov chain properties

- Marginal probability for a variable
  - Expressed in terms of marginal probability of previous variable in chain

$$p(z^{(m+1)}) = \sum_{z^{(m)}} p(z^{(m+1)} \mid z^{(m)})\, p(z^{(m)})$$

Probability of sample is sum of probs over all values of prev sample

- Invariant or stationary
  - A distribution is invariant wrt a Markov chain if each step leaves the marginal distribution invariant
  - $p^*(z)$ is invariant if

$$p^*(z) = \sum_{z'} T(z', z)\, p^*(z')$$

$T(z', z) = p(z|z')$

- Required distribution $p(z)$ is invariant if it satisfies property of *detailed balance*

$$p^*(z)T(z,z') = p^*(z')T(z',z)$$

Two directions between z and z' are the same

Markov chain that respects detailed balance is reversible.

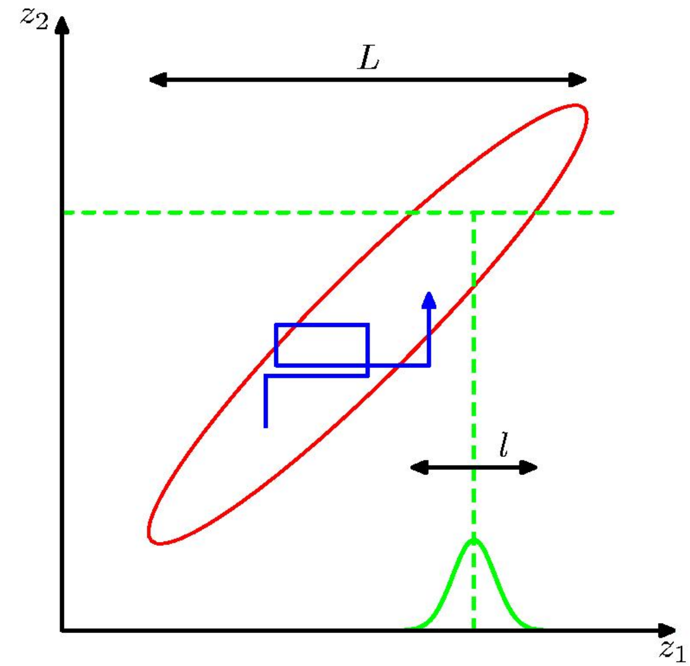
Ergodicity

- Our goal is to use Markov chains to sample from a given distribution
- We need to set up a Markov chain such that the desired distribution is invariant
  - Also, irrespective of choice of initial distribution $p(\mathrm{z}^{(0)})$,
  - As $m \to \infty$ the distribution $p(\mathrm{z}^{(m)})$ converges to the required invariant distribution $p^*(\mathrm{z})$
- This property is called *ergodicity*
  - And the invariant distribution is called the *equilibrium distribution*

Ergodic also means no state has a zero probability of exit from it
And every state is reachable from every other state

Gibbs sampling of two variables

- Two correlated Gaussian variables
- Step size is governed by standard deviation of conditional distribution
  - Is $O(l)$
  - Leads to slow progress in direction of elongation of the joint distribution
- No of steps needed to obtain an independent sample is $O((L/l)^2)$



Conditional distribution of width l
Marginal distribution of width L

Basic Metropolis Algorithm

- As with rejection and importance sampling use a *proposal* distribution (simpler distribution)
- Maintain a record of current state $z^{(t)}$
- Proposal distribution $q(z|z^{(t)})$ depends on current state (next sample depends on previous one)
  - E.g., $q(z|z^{(t)})$ is a symmetric Gaussian with mean $z^{(t)}$ and a small variance
- Thus sequence of samples $z^{(1)}, z^{(2)}\ldots$ forms a Markov chain
- Write $p(z) = \frac{1}{Z_p}\tilde{p}(z)$ where $\tilde{p}(z)$ is readily evaluated
- At each cycle generate candidate z^* and test for acceptance

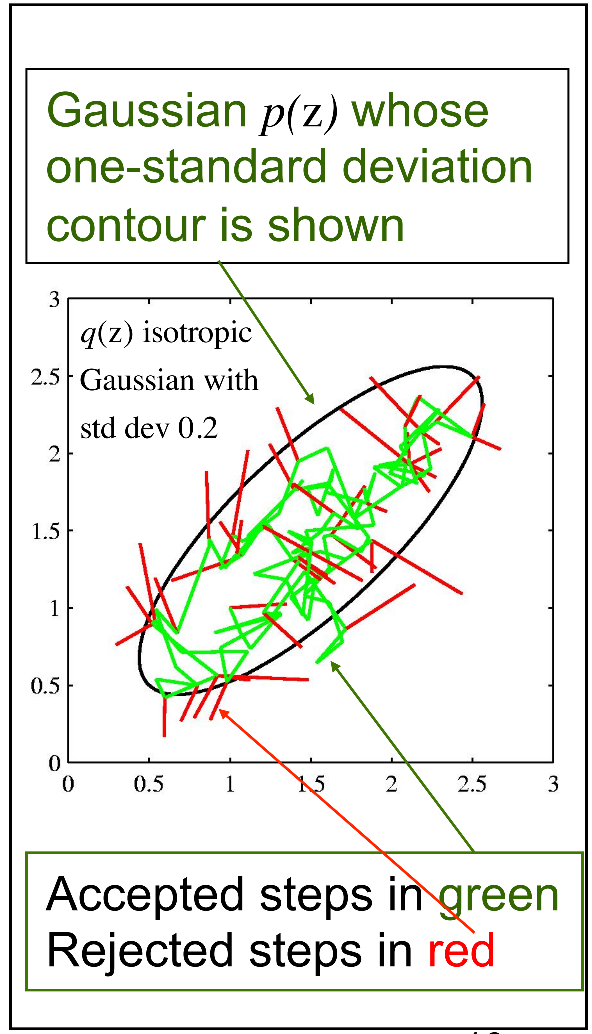
Metropolis Algorithm

- Assumes simple proposal distribution, that is symmetric
  - e.g., an isotropic Gaussian $q(z)$,

    $q(z_A|z_B) = q(z_B|z_A)$ for all z_A, z_B
- New sample z^* from $q(z)$ is accepted with probability

$$A(z^*, z^{(\tau)}) = \min\left(1, \frac{\widetilde{p}(z^*)}{\widetilde{p}(z^{(\tau)})}\right)$$

  - Done by choosing $u \sim U(0,1)$ and accepting if $A(z^*,z^{(t)}) > u$
- If accepted then $z^{(t+1)} = z^*$
- Otherwise:
  - z^* is discarded,
  - $z^{(t+1)}$ is set to $z^{(t)}$ and
  - another candidate drawn from $q(z|z^{(t+1)})$



Inefficiency of Random Walk

- Consider simple random walk
- State space z consisting of integers with probabilities

$p(z^{(t+1)} = z^{(t)}) = 0.5$	Stay in same state
$p(z^{(t+1)} = z^{(t)}+1) = 0.25I$	Increase state by 1
$p(z^{(t+1)} = z^{(t)} - 1) = 0.25$	Decrease state by 1

- If initial state is $z^{(1)}=0$
  - the expected state at time t will be zero, $\mathrm{E}[z^{(t)}] = 0$
    - Similarly $\mathrm{E}[(z^{(t)})^2] = t/2$
    - Thus after t steps distance traveled is proportional to sqrt (t)
- Random walks are inefficient in exploring state-space
  - MCMC algorithms try to avoid random walk behavior

Metropolis-Hastings Algorithm

- Generalizes Metropolis algorithm
  - Proposal distribution is no longer a symmetric function of arguments, i.e., $q(z_A|z_B)$ *.ne.* $q(z_B|z_A)$ for all z_A, z_B

1. At step t , in which current state is $z^{(t)}$ we draw a sample z^* from distribution $q_k(z|z^{(t)})$
2. Accept with probability $A_k(z^*,z^{(t)})$ where

$$A_k(z^*, z^{(\tau)}) = \min\left(1, \frac{\tilde{p}(z^*)q_k(z^{(\tau)} \mid z^*)}{\tilde{p}(z^{(\tau)})q_k(z^* \mid z^{(\tau)})}\right)$$

k labels set of possible transitions considered

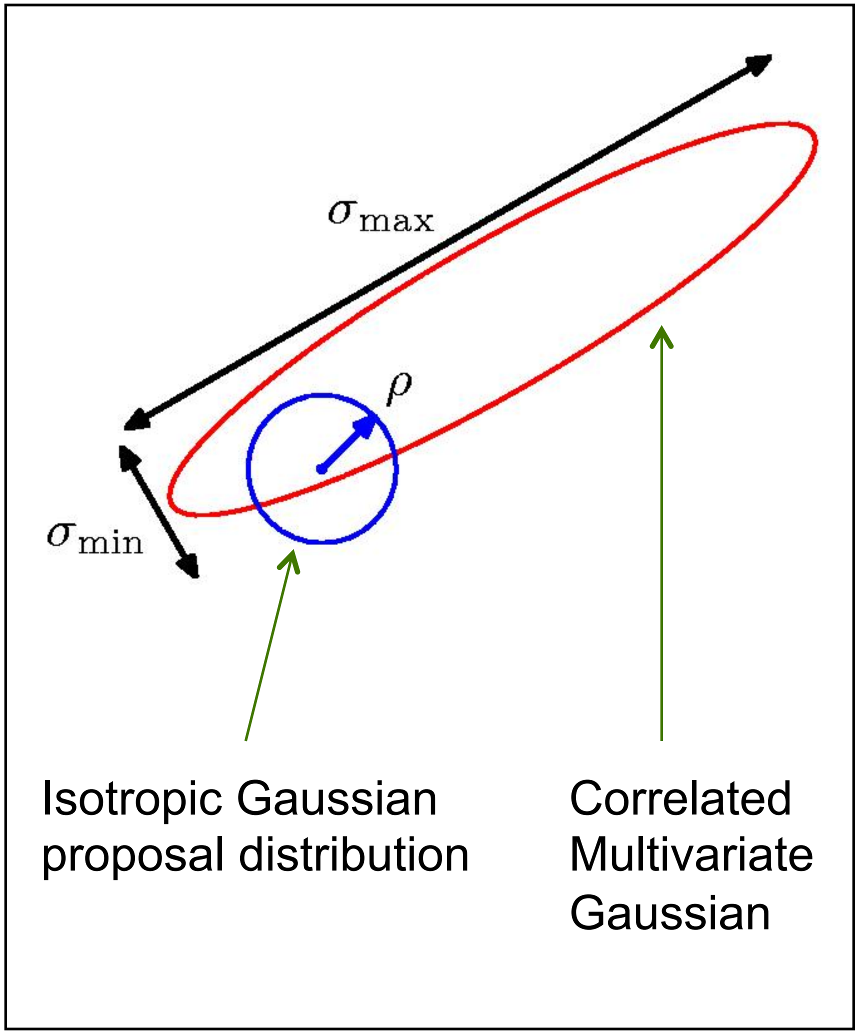
- Can show that $p(z)$ is an invariant distribution of the Markov chain defined by Metropolis-Hastings Algorithm
  - since detailed balance is satisfied

Choice of Proposal Distribution for Metropolis-Hastings

- Gaussian centered on current state
- Keeping rejection rate low

  Scale ρ of proposal distribution should be of order $\sigma_{\min}$
- Independent sample

  No. of steps needed to get independent sample is of order $(\sigma_{\max} / \sigma_{\min})$

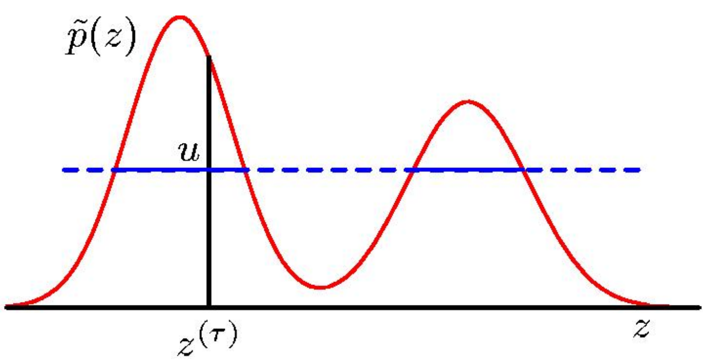


Slice Sampling

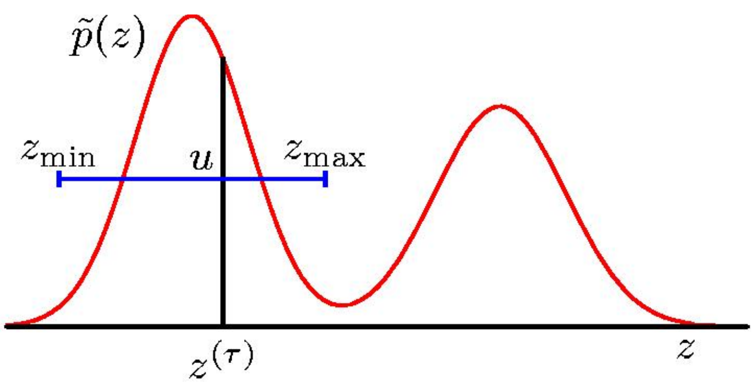
- Metropolis algorithm is sensitive to step size
  - Too small: slow decorrelation due to random walk behavior
  - Too large: inefficiency due to high rejection rate
- Slice sampling provides an adaptive step size to match the distribution

Illustration of Slice

- For given $z^{(t)}$, a value of u is chosen uniformly in region $0 \leq u \leq \tilde{p}(z^{(t)})$
  - which is a slice through the distribution
- Since infeasible to sample from slice, a new sample is drawn from
  $z_{min} \leq z \leq z_{max}$
  which contains the previous value $z^{(t)}$



(a)



(b)