Lect 3: A Variety of Tricks for MCMC Design

Lect 2 introduces 2 general designs: Gibbs sampler and Metropolis-Hastings. This lecture discusses a variety of tricks for designs using Gibbs and Metropolis.

1. Hit-and-run, random ray,
2. Generalized Gibbs sampler.
3. Simulated tempering
4. Data augmentation
5. Slice sampling
6. Cluster sampling
7. Metropolized Gibbs Sampler

Thank Jun Liu (Harvard) for sharing some slides in this lecture.

Gibbs sampler revisit

Goal: sampling a joint probability

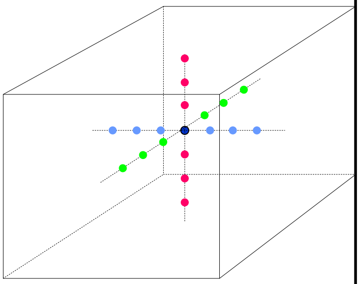
$$X = (x_1, x_2, ..., x_d) \sim \pi(x_1, x_2, ..., x_d)$$

Sampling in each dimension according to a conditional probability

$$x_i \sim \pi(x_i | x_{-i}), \quad \forall i$$

Principle: Each move keeps $\pi(\boldsymbol{x})$ invariant.

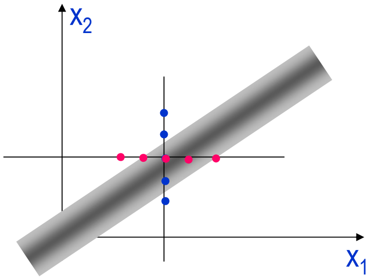
This is ensured by the conditional distribution of the possible moves along dimension i at current x.



A problem with Gibbs sampler

For a probability $p(x_1,x_2)$ whose probability mass is focused on a 1D line segment, sampling the two dimensional iteratively is obviously inefficient. i.e. the chain is "jagging".

This is because the two variables are tightly coupled. It is best if we move along the direction of the line.



A problem with Gibbs sampler (cont.)

Sampling in low dimensional manifolds

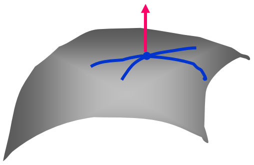
In general, problems arise when the probability is defined in a much lower dimensional manifold in a d-dimensional space. The Markov chain is not allowed to move in the normal directions (off the manifold) but the tangent directions.

As we know, Gibbs distributions are derived from constraints on the variables X, and thus they are defined in some implicit manifold (like the Julesz ensemble in lecture 1)

$$\Omega(H_o) = \{X \ ,:\ H_i(X) = h_i,\ i = 1, 2, ..., K\},$$

$$H_o = (h_1, h_2, ..., h_K)$$

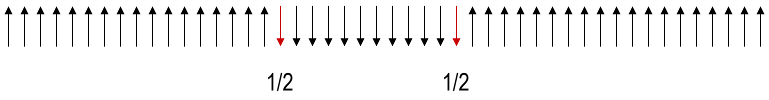
These two examples show that the best way towards fast computation is to find the intrinsic dimensions and hidden (auxiliary) variables.



A problem with Gibbs sampler (cont.)

Coupling in Markov random fields, e.g. the Ising / Potts model

$$p(X;\beta) = \frac{1}{Z}\exp\{\beta \sum_{<s,t>} 1(X_s = X_t)\},\ \beta > 0$$



For single site Gibbs sampler, the boundary spins are flipped with a p=½ probability. Flipping a string of length n will need on average

$t >= 1/p^n = 2^n$ steps!

This is exponential waiting time.

Design ex. 1: Hit-and-Run

It selects a direction at random and shoot like a sniper.

Suppose the current state is X_t

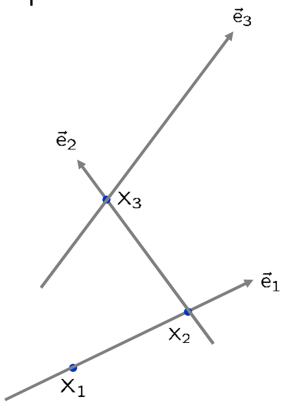
(1) Select a direction or axis $\vec{e}_t$

(2) Sample along the axis.

$$r \sim \pi(X_t + r \cdot \vec{e}_t)$$

(3) Update

$$X_{t+1} = X_t + r \cdot \vec{e}_t$$



The problem is: how do we select the direction?
The sampling along the axis will be a continuous Gibbs and implemented by Multi-Try Metropolis.

Design ex. 2: Generalized Gibbs Sampler

In fact, one may not have to move in straight lines. In more general cases, one may use a group of transformations for the possible moves, as long as the moves preserve the invariant probability.

A Theorem (Liu and Wu, 1999, see ref. Jun Liu Ch. 6)

Let Γ={ γ } be a locally compact group, each element is a possible move

$$x_t \rightarrow x_{t+1} = \gamma \cdot x_t$$

If the element is chosen by

$$\gamma | x \sim \pi(\gamma \cdot x) |J_\gamma(x)| H(d\gamma)$$

Where *H*(dγ) be its left-invariant *Haar measure* $H(\gamma \cdot B) = H(B), \quad \forall \gamma, B$

Then the new state follows the invariant probability

$$x_{t+1} \sim \pi(x)$$

Design ex. 3: Generalized Hit-and-Run

Conceptually it helps to generalize the hit-and-run idea to an arbitrary partition of the space, especially in finite state space. This is a concept by Persi Diaconis 2000.

Suppose a Markov chain consists of many sub-chains, and the transition probability is a linear sum

$$K(x,y) = \sum_{i=1}^{N} \omega_i K_i(x,y), \quad \omega_i = p(i), \ \sum_{i=1}^{N} \omega_i = 1.$$

If each sub-kernel has the same invariant probability,

$$\sum_x \pi(x) K_i(x,y) = \pi(y), \ \forall y$$

Then the whole Markov chain follows $\pi(x)$

Design ex. 3: Generalized Hit-and-Run

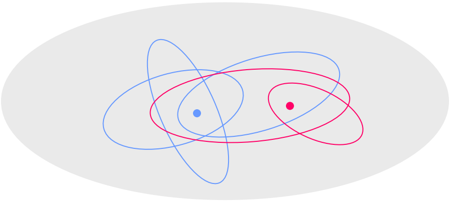
We denote the set of states connected to x by the i-th type moves by

$$\Omega_i(x) = \{y \,:\, K_i(x,y) > 0\}$$

x is connected to a set

$$\Omega(x) = \cup_{i=1}^{N} \Omega_i(x)$$

e.g. K_i be a probability within set Ω_i proportional to $\pi(x)$.



Key problems

$$K(x,y) = \sum_{i=1}^{N} \omega_i K_i(x,y), \quad \omega_i = p(i), \quad \sum_{i=1}^{N} \omega_i = 1.$$

1. How do we decide the sampling dimensions, directions, group transforms, and sets $\Omega_i(x)$ in a systematic and principled way?

2. How do we schedule the visiting order governed by p(i)? i.e. choosing the moving directions, groups, and sets

Sampling with auxiliary variables

A systematic way is to introduce auxiliary random variables:

$$x \sim \pi(x) \;\rightarrow\; (x, y) \sim \pi^{+}(x, y)$$

Examples fpr auxiliary variables y:

T --- temperature : Simulated tempering,
(Narinari and Parisi, 92, Geyer and Thompson, 95)

s --- scale: Multi-grid sampling,
(Goodman and Sokal 88, Liu et al 94)

w --- weight: Dynamic weighting,
(Liang and Wong, 1996)

b --- bond Clustering sampling, Swendsen-Wang
(Swendsen-Wang, 87, Edward and Sokal, 1988)

u --- energy level Slice sampling
(Edwards and Sokal, 88 …)

Design ex. 4: Simulated Tempering

Let the target probabilitv be

$$\pi(x) = \frac{1}{Z}\exp\{-U(x)\}$$

Augment a variable I in {1,2,…,L} for L levels of temperature

$$1 = T_1 < T_2 < \cdots < T_L$$

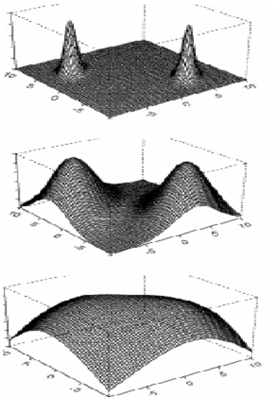
Sampling a joint probability, and keep the X's with I=1

$$(x, I) \sim \pi^{+}(x, I) = \frac{1}{Z+}\exp\{-\frac{1}{T_I}U(x)\}$$

Intuition: the sampler moves more freely in high temperature.
But it is very difficult to cross between different temperature levels.

Design ex. 5: Parallel Simulated Tempering

Suppose we run Markov chains at L levels in parallel



Define a joint probability for all chains

$$\pi^{+}(x_1, ..., x_L) \propto \prod_{i=1}^{L} \exp\{-\frac{1}{T_i}U(x_i)\}$$

Propose to permute two chains:

$$(..., x_i, ..., x_j, ...) \to (..., x_j, ..., x_i, ...)$$

Accept with Metropolis-Hastings

$$\alpha = \min(1, \exp\{(\frac{1}{T_j} - \frac{1}{T_i})(U(x_j) - U(x_i)\}).$$

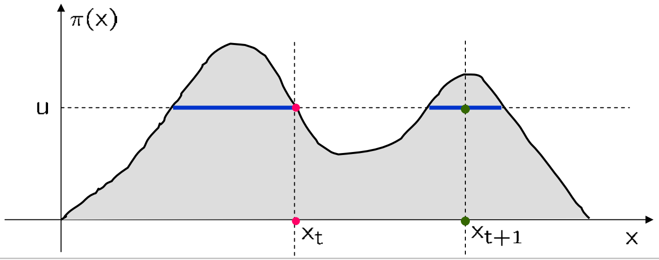
Design ex. 6: Slice Sampling

We use a 1D probability $\pi(x)$ as an example. We introduce an auxiliary variable u in [0,1] for the level of probability. Thus sampling $\pi(x)$ is equivalent to sampling uniformly from the shading area in the [x,u] space.

It proceeds as a 2-step Gibbs sampler

$$\text{(1) Given } x, \text{ sample } y|x \sim \text{unif}[0, \pi(x)]$$

$$\text{(2) Given } y, \text{ sample } x|y \sim \text{unif}\{x : \pi(x) \geq y\}$$

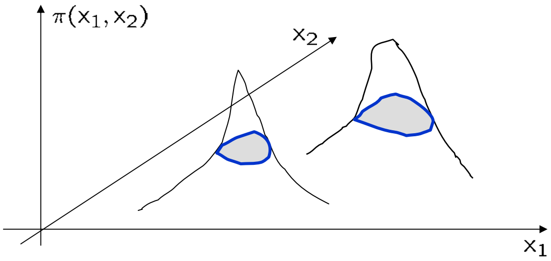


Design ex. 6: Slice Sampling

The first step is easy, but the second step is often intractable in high dimensional space.

$$x|y \sim \mathrm{unif}\{x : \pi(x) \geq y\}$$

$\{x : \pi(x) \geq y\}$ is an area (multiple connected components) bounded by the level set $\pi(x) = y$



Design ex. 7: Data Augmentation

The slice sampling suggests two general conditions for auxiliary variables

$$x \sim \pi(x) \;\rightarrow\; (x, y) \sim \pi^{+}(x, y)$$

1. The marginal probability on x is the invariant probability

$$\sum_{y} \pi^{+}(x, y) = \pi(x)$$

2. The two conditional probabilities have simple forms and are easy to sample

$$y|x \sim \pi^{+}(y|x)$$

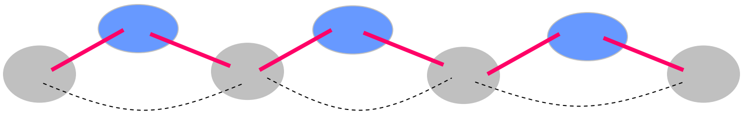
$$x|y \sim \pi^{+}(x|y)$$

Design ex. 7: Data Augmentation

Intuitions:

Very often the probability is focused on separated modes (areas), hopping between these modes are hard, for Markov chains usually move locally.

Good auxiliary variables will be like step stones ...



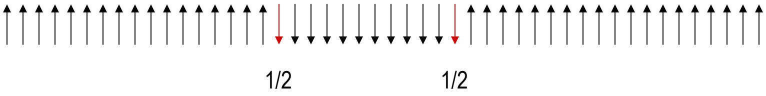
(1) It helps selecting moving directions/groups/sets (in generalized hit-and-run).

(2) It enlarges the search scopes (from a flashlight to a RADAR).

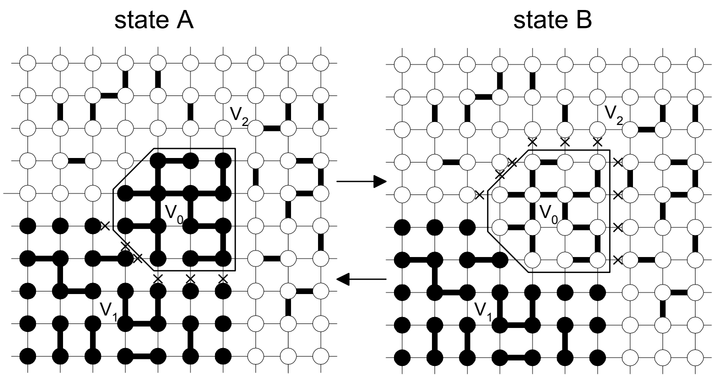
Design ex. 8: Cluster sampling and SW

Swedsen-Wang (1987) is a smart idea for sampling the Potts/Ising model by clustering. It introduces binary (Bernoulli) variables on random fields, so that it can flips a patch/cluster at a time.

$$p(X;\beta) = \frac{1}{Z}\exp\{\beta \sum_{<s,t>} 1(X_s = X_t)\}, \ \beta > 0$$



Design ex. 8: Cluster sampling and SW



Each edge in the lattice e=<s,t> is associated with a variable u_{st} which follows a Bernoulli probability $B(\rho 1(x_s=x_t))$ with $\rho= 1-e^{-\beta}$.

Interpreting SW by data augmentation

One useful interpretation of SW is proposed by Edward and Sokal (1988) using the concept of data augmentation (Tanner and Wang 1987).

Augment the probability with auxiliary variables on the edges of the adjacency graph

$$U = \{u_{st} :< s,t >\in E\}$$

$$(X,U) \sim p_{ES}(X,U)$$

The joint probability is

$$p_{ES} = \frac{1}{Z+} \prod_{(s,t)} [(1-\rho)1(u_{st}=0)+\rho 1(u_{st}=1)\cdot 1(x_s=x_t)]$$

$$\rho = 1 - e^{-\beta}$$

It is not hard to prove that its marginal is the Potts model

$$\sum_U p_{ES}(X,U) = \pi(X)$$

Interpreting SW by data augmentation

Its two conditional probabilities are extremely simple and easy to sample from

1. Flipping the edges by Bernoulli probability,

$$p_{ES}(U|X) = \prod_{<s,t>} p(u_{st}|x_s, x_t)$$

$$p(u_{st}|x_s, x_t) = \text{Bernoulli}(\rho 1(x_s = x_t))$$

2. Flipping a connected component (CCP) by uniform probability,

$$P_{ES}(X|U) = \text{unif}[\frac{\Omega_{\pi_n}}{CP(U))}]$$

CP(U) is a hard constraint that vertices in each connected component according to U has the same color. So we flip the ccp in the quotient space.

Some theoretical results about SW

1. (Gore and Jerrum 97) constructed a "worst case"
   SW does not mix rapidly if G is a complete graph with n>2, and a certain β.

2. (Cooper and Frieze 99) had positive results
   If G is a tree, SW mixing time is O(|G|) for any b.
   If G has constant connectivity O(1), the SW has polynomial mixing time for $\rho<=\rho_0$.

3. (Huber 2002) proposed a method for exact sampling using bounding chain technique
   for small lattice with very low and very high temperature.

To engineers, the real limit of SW is that it is only valid for Ising/Potts models.
(A tiger contained in Potts' cage!)

Furthermore, it makes no use of the data (external fields) in forming clusters.

Design ex. 9: Metropolized Gibbs sampler

Let’s revisit the general idea of hit and run.

$$K(x,y) = \sum_{i=1}^{N} \omega_i K_i(x,y), \quad \omega_i = p(i), \ \sum_{i=1}^{N} \omega_i = 1.$$

We denote the set of states connected to x by the i-th type moves by

$$\Omega_i(x) = \{y \ : \ K_i(x,y) > 0\}$$

x is connected to a set

$$\Omega(x) = \cup_{i=1}^{N} \Omega_i(x)$$

Design ex. 9: Metropolized Gibbs sampler

We know there are two general designs: Gibbs and Metropolis.

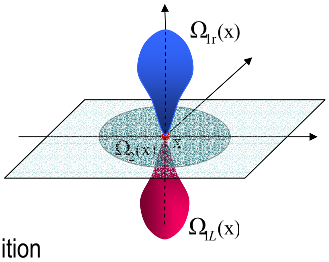
(1) Gibbs design,

we sample a probability in each set

$$y \sim [\pi]_i(y), \quad [\pi]_i(y) = \begin{cases} \sum_{y \in \Omega_i(x)} \pi(y), & y \in \Omega_i(x), \\ 0, & y \notin \Omega_i(x) \end{cases}$$

In this way, the move is symmetric

$$\Omega_i(x) = \Omega_i(y)$$



(2). Metropolis design,

It is no long symmetric, $\Omega_i(x) \neq \Omega_i(y)$

To observe the detailed balance, we need a condition

$$y \in \Omega_i(x) \ \text{ iff } \ x \in \Omega_i(y)$$

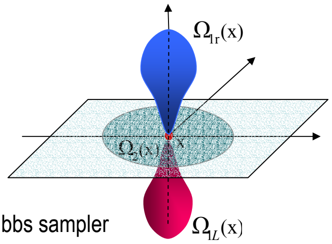
The sub-kernels are designed in pairs

$$K_i(x,y) = \omega_{il} K_{il}(x,y) + \omega_{ir} K_{ir}(x,y)$$

Design ex. 9: Metropolized Gibbs sampler

$$K_{il}(x,y) = Q_{il}(x,y)\,\text{min}(1, \frac{Q_{il}(y,x)}{Q_{ir}(x,y)} \cdot \frac{\pi(y)}{\pi(x)})$$

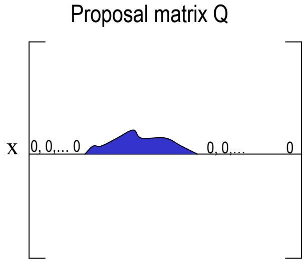
$$\text{for } y \in \Omega_{il}(x)$$



Proposal according to the conditional probabilities --- like a Gibbs sampler

$$\mathrm{Q}_{\mathrm{ir}}(x,y) = \frac{\pi(y)}{\sum_{y' \in \Omega_{ir}(x)} \pi(y')}, \qquad y \in \Omega_{ir}(x);$$

$$\mathrm{Q}_{\mathrm{il}}(y,x) = \frac{\pi(x)}{\sum_{x' \in \Omega_{il}(y)} \pi(x')}, \qquad x \in \Omega_{il}(y);$$


Proposal matrix Q

Discussion

We discussed some design principles and methods.

(1) There is no universal design that is applicable to all problems.

One must analyze the structures of the state space, and the target probability.

(2) In computer vision, it is unlikely to design a Markov chain a priori that works well on all images.

One must look at the data and the Markov chain must by driven by the data.