MARKOV CHAIN MONTE CARLO: *A Workhorse for Modern Scientific Computation*

Xiao-Li Meng

Department of Statistics

Harvard University

Introduction

The Markov chain Monte Carlo (MCMC) methods, originated in computational physics about half a century ago, have seen an enormous range of applications in recent statistical literature, due to their ability to simulate from very complex distributions such as the ones needed in realistic statistical models. This talk provides an introductory tutorial of the two most frequently used MCMC algorithms: the Gibbs sampler and the Metropolis-Hastings algorithm. Using simple yet non-trivial examples, we show, step by step, how to implement these two algorithms. The examples involve a family of bivariate distributions whose full conditional distributions are all normal but whose joint densities are not only non-normal, but also bimodal.

Applications of Monte Carlo

Physics	Sociology	Economics
Chemistry	Education	Finance
Astronomy	Psychology	Management
Biology	Arts	Policy
Environment	Linguistics	Military
Engineering	History	Government
Traffic	Medical Science	Business

…

Monte Carlo 之应用
A Chinese version of the previous slide

物理	社会	经济
化学	教育	金融
天文	心理	管理
生物	人文	政策
环境	语言	军事
工程	历史	政府
交通	医学	商务

…

Monte Carlo Integration

- Suppose we want to compute

$$I = \int g(x) f(x) dx,$$

where f(x) is a probability density. If we have samples $x_1, \ldots, x_n \sim f(x)$, we can estimate I by

$$I_n = \frac{1}{n} \sum_{i=1}^{n} g(x_i)$$

Monte Carlo Optimization

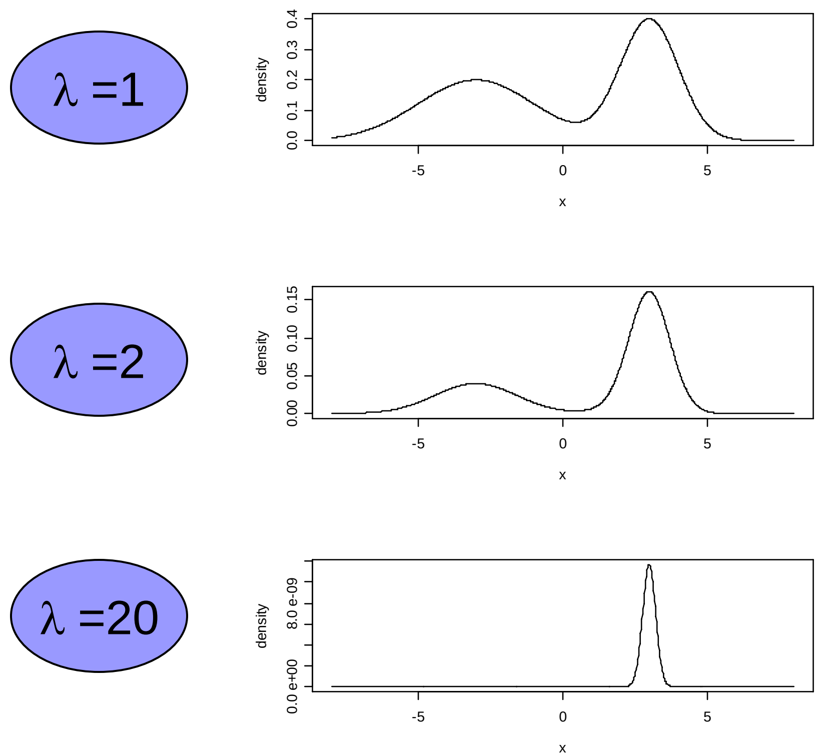
- We want to maximize p(x)
- Simulate from
  $f(x) \propto p^{\lambda}(x)$.

  As $\lambda \to \infty$, the simulated

  draws will be more and

  more concentrated around

  the maximizer of p(x)



Simulating from a Distribution

- What does it mean?

  Suppose a random variable (随机变量) X can only take two values:

$$P(X=0)=\frac{1}{4} \qquad P(X=1)=\frac{3}{4}$$

  Simulating from the distribution of X *means that we want a collection of 0's and 1's:*

$$x_1, x_2, ..., x_n$$

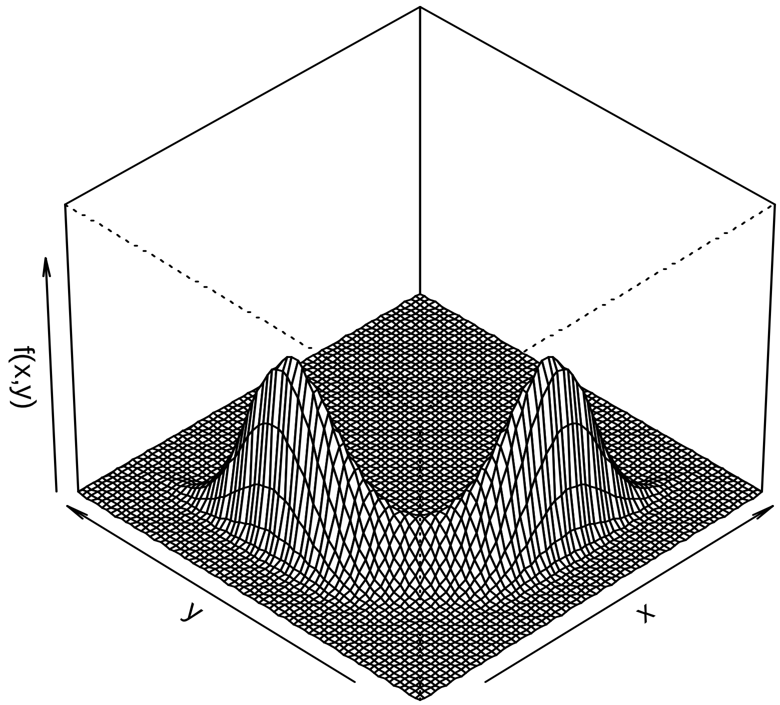
  such that about 25% of them are 0's and about 75%of them are 1's, when n, the simulation size is large.

- The $\{x_i, i = 1,\ldots,n\}$ don't have to be independent

Simulating from a Complex Distribution

- Continuous variable X, described by a density function f(x)

- Complex:
  - *the form of* f(x)
  - *the dimension of* x



$$f(x,y) \propto \exp(-\frac{1}{2}(x^2y^2 + x^2 + y^2 - 8x - 8y))$$

Markov Chain Monte Carlo

$$x^{(t)} = \varphi(x^{(t-1)}, U^{(t)}),$$

where $\{U^{(t)}, t=1,2,\ldots\}$ are identically and independently distributed.

- Under regularity conditions,

$$f(x^{(t)}) \stackrel{t\rightarrow\infty}{\longrightarrow} f(x)$$

So We can treat $\{x^{(t)}, t= N_0, \ldots, N\}$ as an approximate sample from f(x), the stationary/limiting distribution.

Gibbs Sampler

- Target density: $f(x, y)$
- We know how to simulate form the conditional distributions

$$f(x|y) \quad \text{and} \quad f(y|x)$$

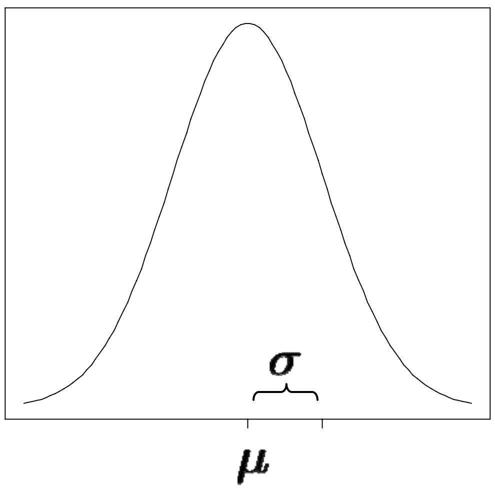
- For the previous example,

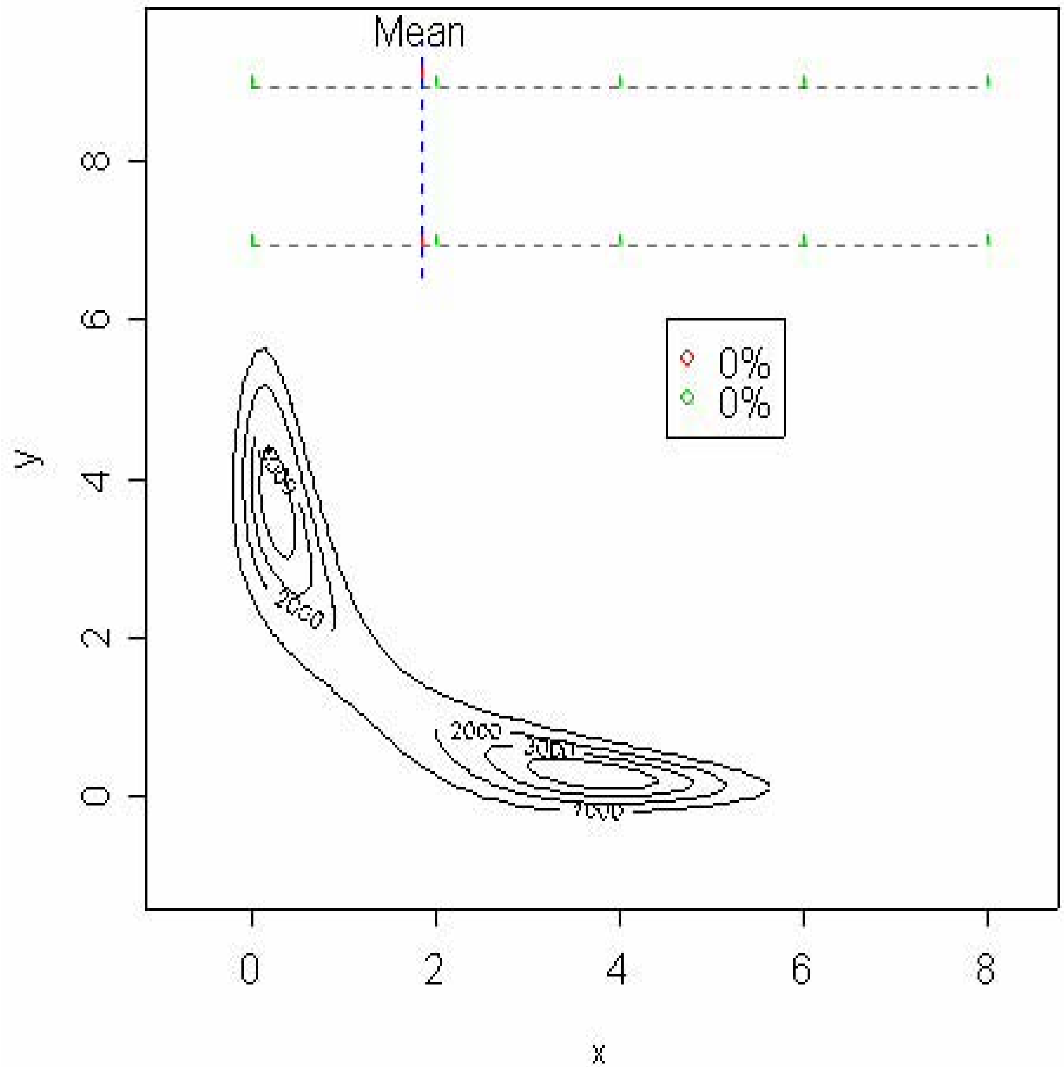
$$f(x,y) \propto \exp(-\frac{1}{2}(x^2y^2 + x^2 + y^2 - 8x - 8y))$$

$$f(x|y) = N(\frac{4}{1+y^2}, \frac{1}{1+y^2})$$

$$f(y|x) = N(\frac{4}{1+x^2}, \frac{1}{1+x^2})$$

$N(\mu,\sigma^2)$
Normal Distribution
“Bell Curve”




Mean
0%
0%
2000
2000
2000
2000
2000
y
x
0
2
4
6
8

Statistical Inference

- Point Estimator: $\bar{g}_n = \frac{1}{n}\sum_{t=1}^{n} g(x^{(t)})$
- Variance Estimator: $V(\bar{g}_n) \approx \frac{\sigma^2}{n}\frac{1+\rho}{1-\rho}$,

$\sigma^2 = Var(g(x))$ estimated by $\hat{\sigma}_n^2 = \frac{1}{n-1}\sum_{t=1}^{n}(g(x^{(t)}) - \bar{g}_n)^2$,

$\rho = corr(g(x^{(t)}), g(x^{(t-1)}))$ estimated by

$$\hat{\rho} = \frac{1}{n-1}\frac{\sum_{t=2}^{n}(g(x^{(t)}) - \bar{g}_n)(g(x^{(t-1)}) - \bar{g}_n)}{\sqrt{\sum_{t=1}^{n-1}(g(x^{(t)}) - \bar{g}_n)^2 \sum_{t=2}^{n}(g(x^{(t)}) - \bar{g}_n)^2}}.$$

- Interval Estimator:

$$(\bar{g}_n - t_d\sqrt{\widehat{V}(\bar{g}_n)}, \quad \bar{g}_n + t_d\sqrt{\widehat{V}(\bar{g}_n)}),$$

where $d = n\frac{1-\rho}{1+\rho} - 1$, and $t_d \to 1.96$ as $n \to \infty$.

Gibbs Sampler (k steps)

- Select an initial value $(x_1^{(0)}, x_2^{(0)}, \ldots, x_k^{(0)})$.
- For t = 0,1,2, …, N
  - Step 1: Draw $x_1^{(t+1)}$ from $f(x_1|x_2^{(t)}, x_3^{(t)}, \ldots, x_k^{(t)})$
  - Step 2: Draw $x_2^{(t+1)}$ from $f(x_2|x_1^{(t+1)}, x_3^{(t)}, \ldots, x_k^{(t)})$

  …
  - Step K:Draw $x_k^{(t+1)}$ from $f(x_k|x_1^{(t+1)}, x_2^{(t+1)}, \ldots, x_{k-1}^{(t+1)})$
- Output $\{(x_1^{(t)}, x_2^{(t)}, \ldots, x_k^{(t)}): t= 1,2,\ldots,N\}$
- Discard the first N_0 draws

  Use $\{(x_1^{(t)}, x_2^{(t)}, \ldots, x_k^{(t)}): t= N_0+1,2,\ldots,N\}$ as (approximate) samples from $f(x_1, x_2, \ldots, x_k)$.

Data Augmentation

- We want to simulate from

$$f(x) \propto \frac{1}{\sqrt{1+x^2}} \exp\{-\frac{1}{2}(x^2 - 8x - \frac{16}{1+x^2})\}.$$

But this is just the marginal distribution of

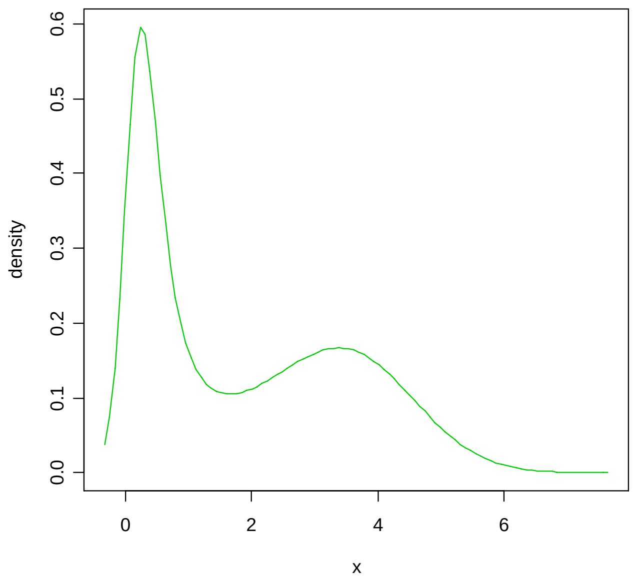
$$f(x,y) \propto \exp(-\frac{1}{2}(x^2y^2 + x^2 + y^2 - 8x - 8y)).$$

So once we have simulations:

$\{(x^{(t)}, y^{(t)}: t = 1,2,\ldots,N)\}$,

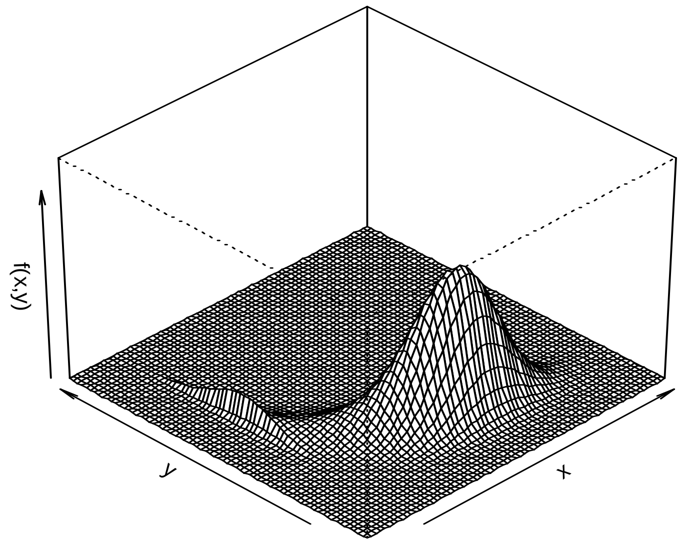
we also obtain draws:

$\{x^{(t)}: t = 1,2,\ldots,N)\}$



A More Complicated Example

$$f(x,y) \propto \exp(-\frac{1}{2}(|x|y^2 + x^2 + y^2 - 8x - 8y))$$



$$f(x,y) = \exp\{-\frac{1}{2}(x-4)^2\}\exp\{-\frac{1}{2}(y-4)^2\}\exp\{-\frac{1}{2}|x|y^2\}$$

Metropolis-Hastings algorithm

- Simulate from an approximate distribution $q(z_1|z_2)$, then
  - Step 0: Select $z^{(0)}$;

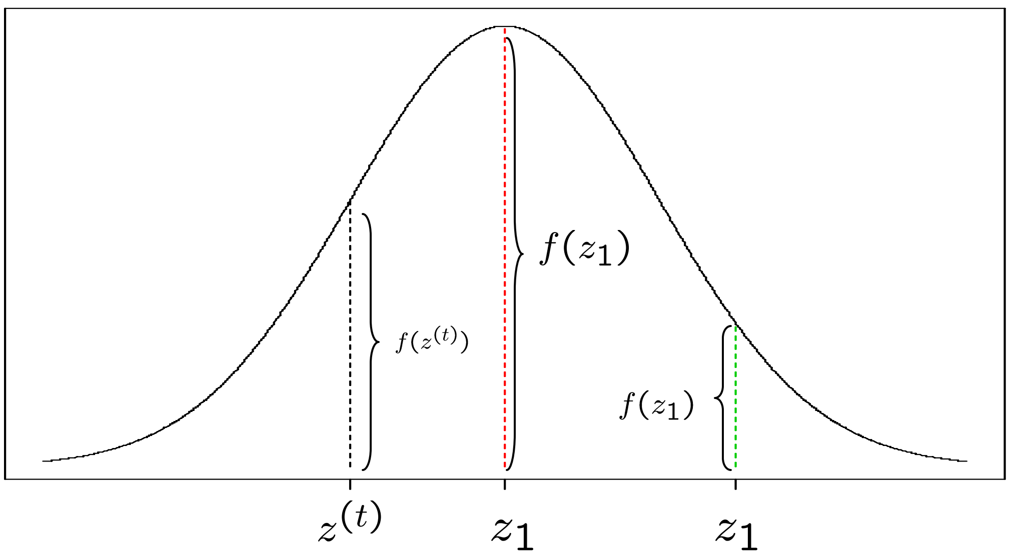
    Now for t = 1,2,…,N, repeat
  - Step 1: draw z_1 from $q(z_1|z_2=z^{(t)})$
  - Step 2: Calculate

$$\alpha(z_1, z^{(t)}) = \frac{f(z_1)q(z^{(t)}|z_1)}{f(z^{(t)})q(z_1|z^{(t)})}$$

  - Step 3: set $z^{(t+1)} = \begin{cases} z_1, & with\ p = \min\{\alpha, 1\} \\ z^{(t)}, & with\ 1-p \end{cases}$ Accept / reject
- Discard the first N_0 draws

M-H Algorithm: An Intuitive Explanation

Assume $q(z_1|z_2) = q(z_2|z_1)$, then $\alpha(z_1, z^{(t)}) = \frac{f(z_1)}{f(z^{(t)})}$



M-H: A Terrible Implementation

$$f(x,y) = \Phi(x-4)\Phi(y-4)\exp\{-\tfrac{1}{2}|x|y^2\}$$

[$\Phi(x)$ is the density function of $N(0,1)$]

We choose $q(z|z_2)=q(z)=\Phi(x-4)\Phi(y-4)$

- Step 1: draw $x \sim N(4,1)$, $y \sim N(4,1)$;

  Dnote $z_1=(x,y)$

- Step 2: Calculate

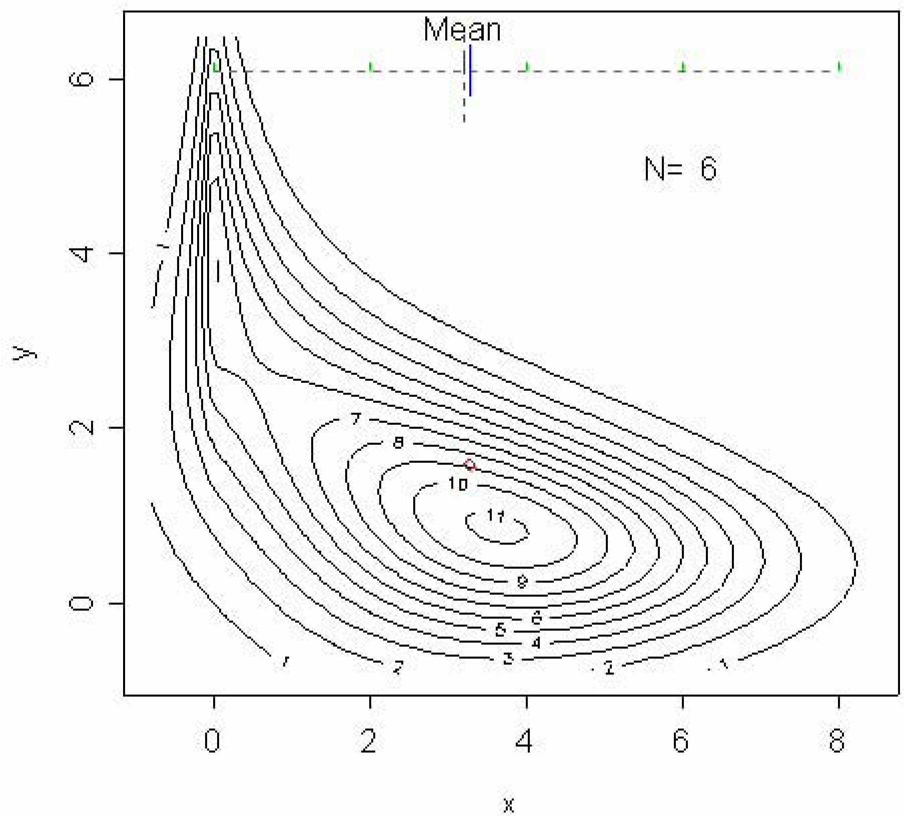
$$\alpha(z_1, z^{(t)}) = \frac{\exp\{-\frac{1}{2}|x|y^2\}}{\exp\{-\frac{1}{2}|x^{(t)}|[y^{(t)}]^2\}}$$

- Step 3: draw $u \sim U[0,1]$

  Let

$$z^{(t+1)} = \begin{cases} z_1, & if \;\; u \le \min\{1,\alpha\} \\ z^{(t)}, & otherwise \end{cases}$$

Why is it so bad?



M-H: A Better Implementation

Starting from some arbitrary $(x^{(0)}, y^{(0)})$

- Step 1: draw $x \sim N(x^{(t)},1)$, $y \sim N(y^{(t)},1)$

  "random walk" $x = x^{(t)} + U_x,\ y = y^{(t)} + U_y$

$$U_x, U_y \overset{iid}{\sim} N(0,1)$$

- Step 2: dnote $z_1=(x,y)$, calculate

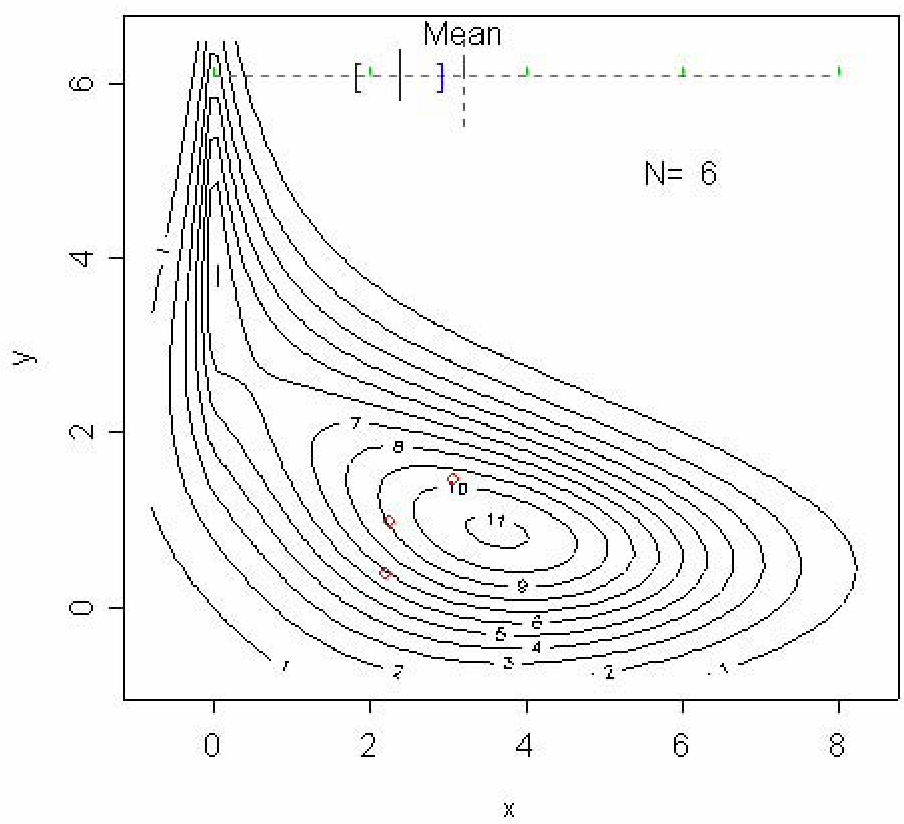
$$\alpha(z_1, z^{(n)}) = \frac{f(z_1)}{f(z^{(t)})}$$

- Step 3: draw $u \sim U[0,1]$

  Let

$$z^{(n+1)} = \begin{cases} z_1, & if\ \ u \le \min\{1, \alpha\} \\ z^{(n)}, & otherwise \end{cases}$$

Much Improved!



Further Discussion

- How large should N_0 and N be?

  Not an easy problem!

- Key difficulty:

  multiple modes in unknown area

- We would like to know all (major) modes, as well as their surrounding mass.

  Not just the global mode

  We need "automatic, Hill-climbing" algorithms.

$\Rightarrow$ The Expectation/Maximization (EM) Algorithm, which can be viewed as a deterministic version of Gibbs Sampler.

Drive/Drink Safely,

Don't become a Statistic;

Go to Graduate School,

Become a Statistician!

"Unfortunately, there's no law against driving after doing triple shifts."