

# STA3431 (Monte Carlo Methods) Lecture Notes, Winter 2009

by Jeffrey S. Rosenthal, University of Toronto

(Last updated: April 6, 2009.)

**Note:** I will update these notes regularly (on-line). However, they are just rough, point-form notes, with no guarantee of completeness or accuracy. They should in no way be regarded as a substitute for attending the lectures, doing the homework exercises, or reading the reference books.

## INTRODUCTION:

- Introduction to course, handout, references, prerequisites, etc. [file “index.html”]
  - Course web page: [probability.ca/sta3431](http://probability.ca/sta3431)
  - If not Stat Dept grad student, must REQUEST enrolment (by e-mail); need strong probability/statistics background (and some programming, everyone!).
  - How many of you are stat grad students? undergrads? math? computer science? physics? economics? management? engineering? other?
- Theme of the course: use (pseudo)randomness on computer to simulate (and hence estimate).
- Example: Suppose want to estimate  $\mathbf{E}[Z^4 \cos(Z^3)]$ , where  $Z \sim \text{Normal}(0, 1)$ .
  - Monte Carlo solution: replicate a large number  $z_1, \dots, z_n$  of  $\text{Normal}(0,1)$  random variables, and let  $x_i = z_i^4 \cos(z_i^3)$ .
  - Their mean  $\bar{x} \equiv \frac{1}{n} \sum_{i=1}^n x_i$  is an (unbiased) estimate of  $\mathbf{E}[X] \equiv \mathbf{E}[Z^4 \cos(Z^3)]$ .
  - R: `Z = rnorm(100); X = Z^4 * cos(Z^3); mean(X)` [file “RMC”]
  - unstable ... but if replace “100” with “10000” then  $\bar{x}$  close to  $-0.09$  ...
  - Variance??

- Well, can estimate standard deviation of  $\bar{x}$  by “standard error” of  $\bar{x}$ , which is:

$$se = n^{-1/2} \text{sd}(x) = n^{-1/2} \sqrt{\text{var}(x)} = n^{-1/2} \sqrt{\frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2}.$$

[file “RMC”]

- Alternatively, could compute expectation as

$$\int_{-\infty}^{\infty} z^4 \cos(z^3) \frac{e^{-z^2/2}}{\sqrt{2\pi}} dz.$$

Analytic? Numerical? Better? Worse? [file “RMC”]

- What about higher-dimensional versions?
- How do we generate Normal(0,1) random variables, etc.? (pseudorandomness, random variates ... we’ll start here ... )
- What if distribution too complicated to sample from? (MCMC!)
- OTHER EXAMPLES:
- HISTORICAL EXAMPLE: BUFFON’S NEEDLE:

- Have series of parallel lines ... line width  $w$ , needle length  $\ell \leq w$  ... what is prob that needle lands touching line? [file buffon.html]
- Let  $\theta$  be angle counter-clockwise from line direction, and  $h$  distance of top end above nearest line.
- Then  $h \sim \text{Uniform}[0, w]$  and  $\theta \sim \text{Uniform}[0, \pi]$ .
- Touches line iff  $h < \ell \sin(\theta)$ .
- So, prob =  $\frac{1}{\pi} \int_0^\pi \frac{1}{w} \int_0^w \mathbf{1}_{h < \ell \sin(\theta)} dh d\theta = \frac{1}{\pi} \int_0^\pi \frac{1}{w} \ell \sin(\theta) d\theta = 2\ell/w\pi$ .
- Hence, by LLN, if throw needle  $n$  times, of which it touches a line  $m$  times, then  $m/n \approx 2\ell/w\pi$ , so  $\pi \approx 2n\ell/mw$ .
- [e.g. recuperating English Captain O.C. Fox, 1864:  $\ell = 3$ ,  $w = 4$ ,  $n = 530$ ,  $m = 253$ , so  $\pi \approx 2n\ell/mw \doteq 3.1423$ .]

- COMPUTER VISION, e.g. “faces” Java applet. [“faces.html”]
- CODE BREAKING, e.g. “decipherit oliver”, “decipheritfast oliver”. [files “decipher.c”, “decipheroutput”]
- e.g. QUEUEING THEORY
  - $Q(t)$  = number of people in queue at time  $t \geq 0$ .
- Suppose first that service times  $\sim$  Exponential( $\mu$ ), and interarrival times  $\sim$  Exponential( $\lambda$ ) (“M/M/1 queue”), so  $\{Q(t)\}$  Markovian. Then well known:
  - If  $\mu \leq \lambda$ , then  $Q(t) \rightarrow \infty$  as  $t \rightarrow \infty$ .
  - If  $\mu > \lambda$ , then  $Q(t)$  converges in distribution as  $t \rightarrow \infty$ :
  - $\mathbf{P}(Q(t) = i) \rightarrow (1 - \frac{\lambda}{\mu})(\frac{\lambda}{\mu})^i$ , for  $i = 0, 1, 2, \dots$
  - Easy! (e.g.  $\mu = 3, \lambda = 2, t = 1000$ ) [file “Rqueue”]
- But suppose now that service times  $\sim$  Uniform[0, 1], and interarrival times have distribution of  $|Z|$  where  $Z \sim$  Normal(0, 1). Limits not easily computed. Now what?
  - Simulate it! [file “Rqueue2”]

## PSEUDORANDOM NUMBERS:

- Goal: generate an i.i.d. sequence  $U_1, U_2, U_3, \dots \sim$  Uniform[0, 1].
- One method: LINEAR CONGRUENTIAL GENERATOR.
  - Choose (large) positive integers  $m, a$ , and  $b$ .
  - Start with a “seed” value,  $x_0$ . (e.g., current time in milliseconds)
  - Then, recursively,  $x_n = (ax_{n-1} + b) \bmod m$ , i.e.  $x_n$  = remainder when  $ax_{n-1} + b$  is divided by  $m$ .
  - So,  $0 \leq x_n \leq m - 1$ .
  - Then let  $U_n = x_n/m$ .
  - Then  $\{U_n\}$  will “seem” to be approximately i.i.d.  $\sim$  Uniform[0, 1]. (file “Rrng”)

- Choice of  $m$ ,  $a$ , and  $b$ ?
- Many issues:
  - need  $m$  large (so many possible values);
  - need  $a$  large enough that no obvious “pattern” between  $U_{n-1}$  and  $U_n$ .
  - need  $b$  to avoid short “cycles” of numbers.
  - many statistical tests, to try to see which choices provide good randomness, avoid correlations, etc. (e.g. “diehard tests”, [www.stat.fsu.edu/pub/diehard](http://www.stat.fsu.edu/pub/diehard); “dieharder”, [www.phy.duke.edu/~rgb/General/dieharder.php](http://www.phy.duke.edu/~rgb/General/dieharder.php))
  - One common “good” choice:  $m = 2^{32}$ ,  $a = 69,069$ ,  $b = 23,606,797$ .
- Not “really” random, just “pseudorandom” ...
  - Can cause problems!
  - Will fail certain statistical tests ...
  - Some implementations also use external randomness, e.g. current temperature of computer’s CPU.
  - Or the randomness of *quantum mechanics*, e.g. [www.fourmilab.ch/hotbits](http://www.fourmilab.ch/hotbits).
  - Or of atmospheric noise, e.g. [random.org](http://random.org).
  - But for most purposes, standard pseudorandom numbers are pretty good ...

---

**END WEEK #1**

---

[See updated web page, notes, evaluation, supp, [preliminary](#) hw1 & project, ...]

**Summary of Previous Class:**

\* Examples of Monte Carlo:

—  $\mathbf{E}[Z^4 \cos(Z^3)]$

— Buffon’s needle (proof in notes, more on hw1)

— computer vision (faces)

— cipher code breaking

— queue simulation

\* Pseudorandom number generation:

— Want  $U_1, U_2, \dots \approx$  i.i.d. Uniform $[0, 1]$

\* e.g. linear congruential generator

—  $x_n = (ax_{n-1} + b) \bmod m$

— Then  $U_n = x_n/m$ .

— e.g.  $m = 2^{32}$ ,  $a = 69,069$ ,  $b = 23,606,797$ .

- Has full period  $m = 2^{32} \doteq 4 \times 10^9$ ; good. [Thm: Full period iff  $\gcd(b, m) = 1$ , and every “prime or 4” divisor of  $m$  also divides  $a - 1 \dots$  so if  $m = 2^{32}$ , then suffices that  $b$  odd and  $a - 1$  a multiple of 4.]
- We’ll consider this “good enough for now”, but:
  - Many other choices, e.g. C programming language (glibc) uses  $m = 2^{32}$ ,  $a = 1,103,515,245$ ,  $b = 12,345$ .
  - One bad choice:  $m = 2^{31}$ ,  $a = 65539 = 2^{16} + 3$ ,  $b = 0$  (“RANDU”) ... used for many years (esp. early 1970s) ... but then  $x_{n+2} = 6x_{n+1} - 9x_n \bmod m \dots$  too much serial correlation. [Proof:  $x_{n+2} = (2^{16} + 3)^2 x_n = (2^{32} + 6(2^{16}) + 9)x_n \equiv (0 + 6(2^{16} + 3) - 9)x_n \pmod{2^{31}} = 6x_{n+1} - 9x_n$ .]
  - (Microsoft Excel pre-2003: period  $< 10^6$ , too small ... Excel 2003 used floating-point “version” of LCG, sometimes gave negative numbers – bad!)
  - Other generators include “Multiply-with-Carry” [ $x_n = (ax_{n-r} + b_{n-1}) \bmod m$  where  $b_n = \lfloor (ax_{n-r} + b_{n-1})/m \rfloor$ ]; and ‘Kiss’ [ $y_n = (x_n + J_n + K_n) \bmod 2^{32}$ , where  $x_n$  as above, and  $J_n$  and  $K_n$  are “shift register generators”, given in bit form by  $J_{n+1} = (I + L^{15})(I + R^{17})J_n \bmod 2^{32}$ , and  $K_{n+1} = (I + L^{13})(I + R^{18})K_n \bmod 2^{31}$ ]; and “Mersenne Twister” [ $x_{n+k} = x_{n+s} \oplus (x_n^{(\text{upper})} | x_{n+1}^{(\text{lower})})A$ , where  $1 \leq s < k$  where  $2^{kw-r} - 1$  is Mersenne prime, and  $A$  is  $w \times w$  (e.g.  $32 \times 32$ ) with  $(w-1) \times (w-1)$  identity in upper-right, with matrix mult. done bit-wise mod 2], and many others too.
  - (R implementation: see “?.Random.seed” ... default is Mersenne Twister.)

## SIMULATING OTHER DISTRIBUTIONS:

- Once we have  $U_1, U_2, \dots$  i.i.d.  $\sim \text{Uniform}[0, 1]$  (at least approximately), how do we generate other distributions?
- With transformations, using “change-of-variable” theorem!
- e.g. to make  $X \sim \text{Uniform}[L, R]$ , set  $X = (R - L)U_1 + L$ .
- e.g. to make  $X \sim \text{Bernoulli}(p)$ , set

$$X = \begin{cases} 1, & U_1 \leq p \\ 0, & U_1 > p \end{cases}$$

- e.g. to make  $Y \sim \text{Binomial}(n, p)$ , either set  $Y = X_1 + \dots + X_n$  where

$$X_i = \begin{cases} 1, & U_i \leq p \\ 0, & U_i > p \end{cases},$$

or set

$$Y = \max \left\{ j : \sum_{k=0}^{j-1} \binom{n}{k} p^k (1-p)^{n-k} \leq U_1 \right\}$$

(where by convention  $\sum_{k=0}^{-1} (\dots) = 0$ ).

- More generally, to make  $\mathbf{P}(Y = x_i) = p_i$  for any  $x_1 < x_2 < x_3 < \dots$ , where  $\sum_i p_i = 1$ , simply set

$$Y = \max \left\{ x_j ; \sum_{k=0}^{j-1} p(x_k) \leq U_1 \right\}.$$

- e.g. to make  $Z \sim \text{Exponential}(1)$ , set  $Z = -\log(U_1)$ .
  - Then  $\mathbf{P}(Z > x) = \mathbf{P}(-\log(U_1) > x) = \mathbf{P}(\log(U_1) < -x) = \mathbf{P}(U_1 < e^{-x}) = e^{-x}$ .
  - Then, to make  $W \sim \text{Exponential}(\lambda)$ , set  $W = Z/\lambda = -\log(U_1)/\lambda$ .
- What about normal dist.? Fact: If

$$X = \sqrt{2 \log(1/U_1)} \cos(2\pi U_2),$$

$$Y = \sqrt{2 \log(1/U_1)} \sin(2\pi U_2),$$

then  $X, Y \sim N(0, 1)$  (independent!). [“Box-Muller transformation”, Ann Math Stat 1958, 29, 610-611]

- Proof: By multidimensional change-of-variable theorem, if  $(x, y) = h(u_1, u_2)$  and  $(u_1, u_2) = h^{-1}(x, y)$ , then  $f_{X,Y}(x, y) = f_{U_1,U_2}(h^{-1}(x, y)) / J(h^{-1}(x, y))$ . Here  $f_{U_1,U_2}(u_1, u_2) = 1$  for  $0 < u_1, u_2 < 1$  (otherwise 0), and

$$\begin{aligned} J(u_1, u_2) &= \det \begin{pmatrix} \frac{\partial x}{\partial u_1} & \frac{\partial x}{\partial u_2} \\ \frac{\partial y}{\partial u_1} & \frac{\partial y}{\partial u_2} \end{pmatrix} \\ &= \det \begin{pmatrix} -\cos(2\pi u_2) / u_1 \sqrt{2 \log(1/u_1)} & -2\pi \sin(2\pi u_2) \sqrt{2 \log(1/u_1)} \\ -\sin(2\pi u_2) / u_1 \sqrt{2 \log(1/u_1)} & 2\pi \cos(2\pi u_2) \sqrt{2 \log(1/u_1)} \end{pmatrix} \\ &= -2\pi / u_1. \end{aligned}$$

But  $u_1 = e^{-(x^2+y^2)/2}$ , so density of  $(X, Y)$  is

$$\begin{aligned} f_{X,Y}(x, y) &= 1/|J(h^{-1}(x, y))| = 1/|-2\pi / e^{-(x^2+y^2)/2}| = e^{-(x^2+y^2)/2} / 2\pi \\ &= \left( \frac{1}{\sqrt{2\pi}} e^{-x^2/2} \right) \left( \frac{1}{\sqrt{2\pi}} e^{-y^2/2} \right), \end{aligned}$$

i.e.  $X \sim N(0, 1)$  and  $Y \sim N(0, 1)$  are independent.

- Another approach: “INVERSE CDF METHOD”:

- Suppose want  $\mathbf{P}(X \leq x) = F(x)$ . (“CDF”)
- For  $0 < t < 1$ , set  $F^{-1}(t) = \min\{x; F(x) \geq t\}$ . (“inverse CDF”)
- Then set  $X = F^{-1}(U_1)$ .
- Then  $X \leq x$  if and only if  $U_1 \leq F(x)$ .
- So,  $\mathbf{P}(X \leq x) = \mathbf{P}(U_1 \leq F(x)) = F(x)$ .

- So, generating (pseudo)random numbers for most “standard” one-dimensional distributions is pretty easy ...

- So, can get Monte Carlo estimates of expectations involving standard one-dimensional distributions, e.g.  $\mathbf{E}[Z^4 \cos(Z^3)]$  where  $Z \sim \text{Normal}(0, 1)$ .

- But what if distribution is complicated, multidimensional, etc.?

## MONTE CARLO INTEGRATION:

- How to compute an integral with Monte Carlo?

- Re-write it as an expectation!

- e.g. estimate  $I = \int_0^5 \int_0^4 g(x, y) dy dx$ , where  $g(x, y) = \cos(\sqrt{xy})$ .

- Here

$$\int_0^5 \int_0^4 g(x, y) dy dx = \int_0^5 \int_0^4 5 \cdot 4 \cdot g(x, y) (1/4) dy (1/5) dx = \mathbf{E}[5 \cdot 4 \cdot g(X, Y)],$$

where  $X \sim \text{Uniform}[0, 5]$  and  $Y \sim \text{Uniform}[0, 4]$ .

- So, let  $X_i \sim \text{Uniform}[0, 5]$ , and  $Y_i \sim \text{Uniform}[0, 4]$  (all independent).

- Estimate  $I$  by  $\frac{1}{M} \sum_{i=1}^M (5 \cdot 4 \cdot g(X_i, Y_i))$ .

- Standard error:  $se = M^{-1/2} sd(5 \cdot 4 \cdot g(X_1, Y_1), \dots, 5 \cdot 4 \cdot g(X_M, Y_M))$ .

- With  $M = 10^6$ , get about  $-4.11 \pm 0.01 \dots$  (file “RMCint2”)

- e.g. estimate  $\int_0^1 \int_0^\infty h(x, y) dy dx$ , where  $h(x, y) = e^{-y^2} \cos(\sqrt{xy})$ .

- (Can’t use “Uniform” expectations.)

- Instead, write this as  $\int_0^1 \int_0^\infty (e^y h(x, y)) e^{-y} dy dx$ .

- This is the same as  $\mathbf{E}[e^Y h(X, Y)]$ , where  $X \sim \text{Uniform}[0, 1]$  and  $Y \sim \text{Exponential}(1)$  are independent.

- So, estimate it by  $\frac{1}{M} \sum_{i=1}^M e^{Y_i} h(X_i, Y_i)$ , where  $X_i \sim \text{Uniform}[0, 1]$  and  $Y_i \sim \text{Exponential}(1)$  (i.i.d.).

- With  $M = 10^6$  get about  $0.767 \pm 0.0004 \dots$  very accurate! (file “RMCint3”)

- (Check: Numerical integration [Mathematica] gives 0.767211.)

---

**END WEEK #2**

---



[New room: BA2145. (I didn't have to, but ... )]

[Hand out HW#1 assignment (and project?). Due Feb 2, at 2:10 sharp.]

[Reminder: probability.ca/sta3431, on-line notes, programs, ... ]

### Summary of Previous Class:

\* More on RNG:

— period, correlations

— Multiply-with-Carry, Kiss, Mersenne Twister

\* Transforming uniforms to binomial, exponential, normal, ...

— Inverse CDF method.

\* Monte Carlo integration

— e.g.  $I \equiv \int_0^1 \int_0^\infty h(x, y) dy dx = \mathbf{E}[e^Y h(X, Y)]$  where  $X \sim \text{Uniform}[0, 1]$  and  $Y \sim \text{Exponential}(1)$ , indep. [ $h(x, y) = e^{-y^2} \cos(\sqrt{xy})$ : get about  $0.767 \pm 0.0004$  ... ]

- Previous example cont'd:

- Alternatively, could write this as  $\int_0^1 \int_0^\infty (\frac{1}{5} e^{5y} h(x, y)) (5 e^{-5y}) dy dx = \mathbf{E}[\frac{1}{5} e^{5Y} h(X, Y)]$  where  $X \sim \text{Uniform}[0, 1]$  and  $Y \sim \text{Exponential}(5)$  (indep.).

- Then, estimate it by  $\frac{1}{M} \sum_{i=1}^M \frac{1}{5} e^{5y_i} h(x_i, y_i)$ , where  $x_i \sim \text{Uniform}[0, 1]$  and  $y_i \sim \text{Exponential}(5)$  (i.i.d.).

- With  $M = 10^6$ , get about  $0.767 \pm 0.0016$  ... larger standard error ... (file “RMCint4”).

- If replace 5 by 1/5, get about  $0.767 \pm 0.0015$  ... about the same.

- So which choice is best?

- Whichever one minimises the standard error! ( $\lambda \approx 1.5$ ,  $se \approx 0.00025$ ?)

- In general, to evaluate  $I \equiv \mathbf{E}[h(Y)] = \int h(y) \pi(y) dy$ , where  $Y$  has density  $\pi$ , could instead re-write this as  $I = \int h(x) \frac{\pi(x)}{f(x)} f(x) dx$ , where  $f$  is easily sampled from, with  $f(x) > 0$  whenever  $\pi(x) > 0$ .

- Then  $I = \mathbf{E}\left(h(X) \frac{\pi(X)}{f(X)}\right)$ , where  $X$  has density  $f$ .

- Can then do classical (iid) Monte Carlo integration, get standard errors etc.

- Good if easier to sample from  $f$  than  $\pi$ , and/or if the function  $h(x) \frac{\pi(x)}{f(x)}$  is less variable than  $h$  itself.
- In general, best to make  $h(x) \frac{\pi(x)}{f(x)}$  approximately constant.
  - e.g. extreme case: if  $I = \int_0^\infty e^{-3x} dx$ , then  $I = \int_0^\infty (1/3)(3e^{-3x})dx = \mathbf{E}[1/3]$  where  $X \sim \text{Exponential}(3)$ , so  $I = 1/3$  (error = 0, no MC needed).
- Suppose now that  $\pi(y) = c g(y)$ , where we know  $g$  but don't know  $c$  or  $\pi$ . (“Unnormalised density”, e.g. Bayesian posterior.)
  - Obviously,  $c = \frac{1}{\int g(y) dy}$ , but this might be hard to compute.
  - Still,  $I = \int h(x) \pi(x) dx = \int h(x) c g(x) dx = \frac{\int h(x) g(x) dx}{\int g(x) dx}$ .
  - If sample  $\{x_i\} \sim f$  (i.i.d.), then  $\int h(x) g(x) dx = \int \left( h(x) g(x) / f(x) \right) f(x) dx = \mathbf{E}[h(X) g(X) / f(X)]$  where  $X \sim f$ .
  - So,  $\int h(x) g(x) dx \approx \frac{1}{M} \sum_{i=1}^M \left( h(x_i) g(x_i) / f(x_i) \right)$ .
  - Similarly,  $\int g(x) dx \approx \frac{1}{M} \sum_{i=1}^M \left( g(x_i) / f(x_i) \right)$ .
  - So,  $I \approx \frac{\sum_{i=1}^M \left( h(x_i) g(x_i) / f(x_i) \right)}{\sum_{i=1}^M \left( g(x_i) / f(x_i) \right)}$ . (“Importance Sampling”: weighted average)
  - (Not unbiased, standard errors less clear, but still consistent.)
- Example: compute  $I \equiv \mathbf{E}(Y^2)$  where  $Y$  has density  $c y^3 \sin(y^4) \cos(y^5) \mathbf{1}_{0 < y < 1}$ , where  $c > 0$  unknown (and hard to compute!).
  - Here  $g(y) = y^3 \sin(y^4) \cos(y^5) \mathbf{1}_{0 < y < 1}$ , and  $h(y) = y^2$ .
  - Let  $f(y) = 6 y^5 \mathbf{1}_{0 < y < 1}$ . [Note: if  $U \sim \text{Uniform}[0, 1]$ , then  $X \equiv U^{1/6} \sim f$ , since then for  $0 < x < 1$ ,  $\mathbf{P}(X \leq x) = \mathbf{P}(U^{1/6} \leq x) = \mathbf{P}(U \leq x^6) = x^6$ , so  $f_X(x) = \frac{d}{dx} x^6 = f(x)$ .]
  - Then  $I \approx \frac{\sum_{i=1}^M \left( h(x_i) g(x_i) / f(x_i) \right)}{\sum_{i=1}^M \left( g(x_i) / f(x_i) \right)} = \frac{\sum_{i=1}^M \left( \sin(x_i^4) \cos(x_i^5) \right)}{\sum_{i=1}^M \left( \sin(x_i^4) \cos(x_i^5) / x_i^2 \right)}$ . (file “Rimp” ... get about 0.766 ...)

- Or, let  $f(y) = 4y^3 \mathbf{1}_{0 < y < 1}$ . [Then if  $U \sim \text{Uniform}[0, 1]$ , then  $U^{1/4} \sim f$ .]
- Then  $I \approx \frac{\sum_{i=1}^M (h(x_i) g(x_i) / f(x_i))}{\sum_{i=1}^M (g(x_i) / f(x_i))} = \frac{\sum_{i=1}^M (\sin(x_i^4) \cos(x_i^5) x_i^2)}{\sum_{i=1}^M (\sin(x_i^4) \cos(x_i^5))}$ . (file “Rimp”)

- What other methods to iid sample from  $\pi$ ?

## REJECTION SAMPLER:

- Suppose want to sample  $X$  from the density  $\pi$ .
  - (Then if  $X_1, X_2, \dots, X_M \sim \pi$ , then can estimate  $\mathbf{E}_\pi(h)$  by  $\frac{1}{M} \sum_{i=1}^M h(X_i)$ , etc.)
  - Assume  $\pi(x) = c g(x)$ , with  $\pi$  and  $c$  unknown,  $g$  known but hard to sample from.
  - Find some other, easily-sampled density  $f$ , and known  $K > 0$ , such that  $K f(x) \geq g(x)$  for all  $x$ .
  - Sample  $X \sim f$ , and  $U \sim \text{Uniform}[0, 1]$  (indep.).
  - If  $U \leq g(X)/Kf(X)$ , then accept  $X$  (as a draw from  $\pi$ ).
  - Otherwise, reject  $X$  and start over again.
- Conditional on accepting, we have

$$\begin{aligned} \mathbf{P}\left(X \leq y \mid U \leq \frac{g(X)}{Kf(X)}\right) &= \frac{\mathbf{P}\left(X \leq y, U \leq \frac{g(X)}{Kf(X)}\right)}{\mathbf{P}\left(U \leq \frac{g(X)}{Kf(X)}\right)} \\ &= \frac{\int_{-\infty}^y f(x) \frac{g(x)}{Kf(x)} dx}{\int_{-\infty}^{\infty} f(x) \frac{g(x)}{Kf(x)} dx} = \frac{\int_{-\infty}^y g(x) dx}{\int_{-\infty}^{\infty} g(x) dx} = \int_{-\infty}^y \pi(x) dx. \end{aligned}$$

- So, conditional on accepting,  $X \sim \pi$ . Good! iid!
- However, prob. of accepting may be very small, then get very few samples.
- Example:  $\pi = N(0, 1)$ , i.e.  $g(x) = \pi(x) = (2\pi)^{-1/2} \exp(-x^2/2)$ .
  - Want:  $\mathbf{E}_\pi(X^4)$ , i.e.  $h(x) = x^4$ .
  - Let  $f$  be double-exponential distribution, i.e.  $f(x) = \frac{1}{2} e^{-|x|}$ .
- If  $K = 8$ , then:

- For  $|x| \leq 2$ ,  $Kf(x) = 8 \frac{1}{2} \exp(-|x|) \geq 8 \frac{1}{2} \exp(-2) \geq (2\pi)^{-1/2} \geq \pi(x) = g(x)$ .
- For  $|x| \geq 2$ ,  $Kf(x) = 8 \frac{1}{2} \exp(-|x|) \geq 8 \frac{1}{2} \exp(-x^2/2) \geq (2\pi)^{-1/2} \exp(-x^2/2) = \pi(x) = g(x)$ .
- So, can apply rejection sampler with this  $f$  and  $K$ , to get samples, estimate of  $\mathbf{E}[X]$ , estimate of  $\mathbf{E}[h(X)]$ , estimate of  $\mathbf{P}[X < -1]$ , etc. (file “Rrej”)
- A RELATED APPROACH: (“auxiliary variable”, “slice sampler”)
  - Suppose  $\pi(x) = c g(x)$ , and  $(X, Y)$  chosen uniformly under the graph of  $g$ .
  - i.e.,  $(X, Y) \sim \text{Uniform}\{(x, y) : 0 \leq y \leq g(x)\}$ .
  - Then  $X \sim \pi$ , i.e. we have sampled from  $\pi$ .
  - Why? For  $a < b$ ,  $\mathbf{P}(a < X < b) = \frac{\text{area with } a < X < b}{\text{total area}} = \frac{\int_a^b g(x) dx}{\int_{-\infty}^{\infty} g(x) dx} = \int_a^b \pi(x) dx$ .
  - So, if repeat, get i.i.d. samples from  $\pi$ , can estimate  $\mathbf{E}_\pi(h)$  etc.

---

**END WEEK #3**

---

[Reminder: HW#1 due next Monday, 2:10 sharp; Project due April 6.]

[HW#1 Q3: “two algorithms” could be just two different  $f$ .]

[Office hours? This Wed 12:30? 4:30? Friday 2:30?]

[Kung Hay Fat Choy!]

**Summary of Previous Class:**

- \* Monte Carlo integration “choices”
- \* Unnormalised densities:  $\pi(x) = c g(x)$
- \* Importance sampling
- \* Rejection sampling:  $K f(x) \geq g(x)$
- \* Auxiliary variable trick:

— if  $(X, Y) \sim \text{Uniform}\{(x, y) : 0 \leq y \leq g(x)\}$ , then  $X \sim \pi$ .

- For Rejection Sampler,  $P(\text{accept}) = \mathbf{E}[P(\text{accept}|X)] = \mathbf{E}\left[\frac{g(X)}{Kf(X)}\right] = \int \frac{g(x)}{Kf(x)} f(x) dx = \frac{1}{K} \int g(x) dx = \frac{1}{cK}$ . (Only depends on  $K$ , not  $f$ .)

- So, in  $M$  attempts, get about  $M/cK$  iid samples.
- (“Rrej” example:  $c = 1$ ,  $K = 8$ ,  $M = 10,000$ , so get about  $M/8 = 1250$  samples.)
- Since  $c$  fixed, try to minimise  $K$ .
- Extreme case:  $f(x) = \pi(x)$ , so  $g(x) = \pi(x)/c = f(x)/c$ , and can take  $K = 1/c$ , whence  $P(\text{accept}) = 1$ , iid sampling: optimal.
- Auxiliary Variable rejection sampler:
  - If support of  $g$  contained in  $[L, R]$ , and  $|g(x)| \leq K$ , then can first sample  $(X, Y) \sim \text{Uniform}([L, R] \times [0, K])$ , then reject if  $Y > g(X)$ , otherwise accept as sample with  $(X, Y) \sim \text{Uniform}\{(x, y) : 0 \leq y \leq g(x)\}$ , hence  $X \sim \pi$ .
- Example:  $g(y) = y^3 \sin(y^4) \cos(y^5) \mathbf{1}_{0 < y < 1}$ .
  - Then  $L = 0$ ,  $R = 1$ ,  $K = 1$ .
  - So, sample  $X, Y \sim \text{Uniform}[0, 1]$ , then keep  $X$  iff  $Y \leq g(X)$ .
  - If  $h(y) = y^2$ , could compute e.g.  $\mathbf{E}_\pi(h)$  as the mean of the squares of the accepted samples. (file “Raux”)
- Can iid / importance / rejection / auxiliary sampling solve all problems? No!

## BAYESIAN STATISTICS:

- Have unknown parameter(s)  $\theta$ , and a statistical model (likelihood function) for how the distribution of the data  $Y$  depends on  $\theta$ :  $\mathcal{L}(Y | \theta)$ .
- Have a prior distribution, representing our “initial” (subjective?) probabilities for  $\theta$ :  $\mathcal{L}(\theta)$ .
- Combining these gives a full joint distribution for  $\theta$  and  $Y$ , i.e.  $\mathcal{L}(\theta, Y)$ .
- Then posterior distribution of  $\theta$ ,  $\pi(\theta)$ , is then the conditional distribution of  $\theta$ , conditioned on the observed data  $y$ , i.e.  $\pi(\theta) = \mathcal{L}(\theta | Y = y)$ .
  - In terms of densities, if have prior density  $f_\theta(\theta)$ , and likelihood  $f_{Y|\theta}(y, \theta)$ , then

joint density is  $f_{\theta,Y}(\theta, y) = f_{\theta}(\theta) f_{Y|\theta}(y, \theta)$ , and posterior density is

$$\pi(\theta) = \frac{f_{\theta,Y}(\theta, y)}{f_Y(y)} = c f_{\theta,Y}(\theta, y) = c f_{\theta}(\theta) f_{Y|\theta}(y, \theta).$$

- Example: VARIANCE COMPONENTS MODEL (“random effects model”):

- Suppose some population has overall mean  $\mu$  (unknown).
- Population consists of  $K$  groups.
- Observe  $Y_{i1}, \dots, Y_{iJ_i}$  from group  $i$ , for  $1 \leq i \leq K$ .
- Assume  $Y_{ij} \sim N(\theta_i, W)$  (cond. ind.), where  $\theta_i$  and  $W$  unknown.
- Assume the different  $\theta_i$  are “linked” by  $\theta_i \sim N(\mu, V)$  (cond. ind.), with  $\mu$  and  $V$  also unknown.
- Want to estimate some or all of  $V, W, \mu, \theta_1, \dots, \theta_K$ .
- Bayesian approach: use prior distributions, e.g. (“conjugate”):

$$V \sim IG(a_1, b_1); \quad W \sim IG(a_2, b_2); \quad \mu \sim N(a_3, b_3),$$

where  $a_i, b_i$  known constants, and  $IG(a, b)$  is “inverse gamma” distribution, with density  $\frac{b^a}{\Gamma(a)} e^{-b/x} x^{-a-1}$  for  $x > 0$ .

- Related application #1 (D. Rubin, JASA 1980):

- Success in first-year at law school  $i$  depends on student’s previous LSAT score and on undergraduate GPA:  $F \propto L + c_i G$ .
- Want to estimate  $c_i$ , for  $1 \leq i \leq 82$  law schools.
- Here  $200 \leq L \leq 800$ , and  $1.0 \leq G \leq 4.0$ . So,  $c_i = 200$ ? Smaller?
- Instead assume  $c_i = 200 \tan(\theta_i)$ , where the  $\theta_i$  are unknown but “linked”.

- Related application #2 (Bartolucci, Singh, Bae: [www.mssanz.org.au/modsim07/MODSIM07.htm](http://www.mssanz.org.au/modsim07/MODSIM07.htm)):

- 724 early-stage post-operative melanoma (skin cancer) patients, from 19 different patient “catagories” (distinguished by thickness of the lesion, whether the lesion was ulcerated, ...).

- Category  $i$  has  $n_i$  patients, of which  $r_i$  had cancer recurrence.
- Model  $r_i \sim \text{Binomial}(n_i, \theta_i)$ , where  $\theta_i$  is risk of recurrence for  $i^{\text{th}}$  category.
- Again, the  $\theta_i$  ( $1 \leq i \leq 19$ ) are unknown but “linked”.
- Related application #3 (J. Albert, The American Statistician 1992):
  - 12 leading baseball home-run hitters: who was best?
  - Player  $i$  hit  $H_i$  home-runs in  $A_i$  attempts (“at-bats”).
  - Model e.g.  $H_i \sim \text{Binomial}(A_i, \theta_i)$  or  $\text{Poisson}(\theta_i A_i)$ , where  $\theta_i$  is “true” prob.
  - Again,  $\theta_i$  ( $1 \leq i \leq 12$ ) are unknown and “linked”. [HW#2?]
- Related application #4 (Davies, 1967; Box/Tiao, 1973; Gelfand/Smith, JASA 1990):
  - $K = 6$  batches of “dyestuff”.
  - $J = 5$  samples from each batch.
  - $Y_{ij}$  = amount of “yield” from  $j^{\text{th}}$  sample of  $i^{\text{th}}$  batch.
  - Assume  $Y_{ij} \sim N(\theta_i, W)$ , where  $\theta_i \sim N(\mu, V)$ .
  - Comparison of  $W$  and  $V$ ? [HW#2?]
- “Related” to all of the above is our statistical model:

$$Y_{ij} \sim N(\theta_i, W); \quad (1 \leq i \leq K; \quad 1 \leq j \leq J_i)$$

where

$$\theta_i \sim N(\mu, V); \quad (1 \leq i \leq K),$$

with  $V, W, \mu, \theta_1, \dots, \theta_K$  all unknown, but  $\{Y_{ij}\}$  observed data (known), and prior distributions

$$V \sim \text{IG}(a_1, b_1); \quad W \sim \text{IG}(a_2, b_2); \quad \mu \sim N(a_3, b_3).$$

- Then for  $V, W > 0$ , joint density is:

$$f(V, W, \mu, \theta_1, \dots, \theta_K, Y_{11}, Y_{12}, \dots, Y_{KJ_K})$$

$$\begin{aligned}
&= C_1 \left( e^{-b_1/V} V^{-a_1-1} \right) \left( e^{-b_2/W} W^{-a_2-1} \right) \left( e^{-(\mu-a_3)^2/2b_3} \right) \times \\
&\quad \times \left( \prod_{i=1}^K V^{-1/2} e^{-(\theta_i-\mu)^2/2V} \right) \left( \prod_{i=1}^K \prod_{j=1}^{J_i} W^{-1/2} e^{-(Y_{ij}-\theta_i)^2/2W} \right) \\
&= C e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-\frac{1}{2} \sum_{i=1}^K J_i} \times \\
&\quad \times \exp \left[ -\sum_{i=1}^K (\theta_i - \mu)^2/2V - \sum_{i=1}^K \sum_{j=1}^{J_i} (Y_{ij} - \theta_i)^2/2W \right].
\end{aligned}$$

- Then

$$\begin{aligned}
&\pi(V, W, \mu, \theta_1, \dots, \theta_K) \\
&= c \left( e^{-b_1/V} V^{-a_1-1} \right) \left( e^{-b_2/W} W^{-a_2-1} \right) \left( e^{-(\mu-a_3)^2/2b_3} \right) \times \\
&\quad \times \left( \prod_{i=1}^K V^{-1/2} e^{-(\theta_i-\mu)^2/2V} \right) \left( \prod_{i=1}^K \prod_{j=1}^{J_i} W^{-1/2} e^{-(Y_{ij}-\theta_i)^2/2W} \right)
\end{aligned}$$

---

**END WEEK #4**

---

[Collect HW#1; hand out HW#2. (Count HW#2 for 12%? “R” on in-class test?)]

**Summary of Previous Class:**

- \* more re Rejection Samplers, Auxiliary Variables
- \* Bayesian statistics: likelihood, prior, posterior
- \* Examples where  $\{\theta_i\}_{i=1}^K$  unknown but “linked”
  - law school admissions, skin cancer patients, baseball, dyestuffs
  - dimension  $K + 3$ , where  $K = 82$  or  $19$  or  $12$  or  $6$  or ...
- \* Variance components model:
  - $Y_{ij} \sim N(\theta_i, W)$ , where  $\theta_i \sim N(\mu, V)$ .
  - Priors:  $V \sim IG(a_1, b_1)$ ;  $W \sim IG(a_2, b_2)$ ;  $\mu \sim N(a_3, b_3)$ .

- After a bit of simplifying,

$$\begin{aligned}
&\pi(V, W, \mu, \theta_1, \dots, \theta_K) \\
&= C e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-\frac{1}{2} \sum_{i=1}^K J_i} \times
\end{aligned}$$



$$\times \exp \left[ - \sum_{i=1}^K (\theta_i - \mu)^2 / 2V - \sum_{i=1}^K \sum_{j=1}^{J_i} (Y_{ij} - \theta_i)^2 / 2W \right].$$

- Dimension:  $d = K + 3$ , e.g.  $K = 19$ ,  $d = 22$ .
  - Numerical integration? No, too high-dimensional!
  - Importance sampling? Perhaps, but what “ $f$ ”? Not very efficient!
  - Rejection sampling? What “ $f$ ”? What “ $K$ ”? Virtually no samples!
  - Alternative: MCMC ...

## MARKOV CHAIN MONTE CARLO (MCMC):

- Suppose have complicated, high-dimensional density  $\pi = c g$ .
- Want samples  $X_1, X_2, \dots \sim \pi$ . (Then can do Monte Carlo.)
- Define a Markov chain (random process)  $X_0, X_1, X_2, \dots$ , so for large  $n$ ,  $X_n \approx \pi$ .
- METROPOLIS ALGORITHM (1953):
  - Choose some initial value  $X_0$  (perhaps random).
  - Then, given  $X_{n-1}$ , choose a proposal move  $Y_n \sim MVN(X_{n-1}, \sigma^2 I)$  (say).
  - Let  $A_n = \pi(Y_n) / \pi(X_{n-1}) = g(Y_n) / g(X_{n-1})$ , and  $U_n \sim \text{Uniform}[0, 1]$ .
  - Then, if  $U_n < A_n$ , set  $X_n = Y_n$  (“accept”), otherwise set  $X_n = X_{n-1}$  (“reject”).
  - Repeat, for  $n = 1, 2, 3, \dots, M$ .
  - (Note: only need to compute  $\pi(Y_n) / \pi(X_{n-1})$ , so multiplicative constants cancel.)
- Fact: Then, for large  $n$ , have  $X_n \approx \pi$ . (“rwm.html” Java applet)
- Then can estimate  $\mathbf{E}_\pi(h) \equiv \int h(x) \pi(x) dx$  by:

$$\mathbf{E}_\pi(h) \approx \frac{1}{M - B} \sum_{i=B+1}^M h(X_i),$$

where  $B$  (“burn-in”) chosen large enough so  $X_B \approx \pi$ , and  $M$  chosen large enough to get good Monte Carlo estimates.

- How large  $B$ ? Difficult to say! (Some theory ... active area of research [see e.g. Rosenthal, Elec Comm Prob 2002] ... usually use trial-and-error ... )
- What initial value  $X_0$ ?
  - Virtually any one will do, but “central” ones best.
  - Ideal: “overdispersed starting distribution”, i.e. choose  $X_0$  randomly from some distribution that “covers” the “important” part of the state space.
- What about standard error, i.e. uncertainty?
  - It’s usually larger than in iid case (due to correlations), and harder to quantify.
  - Simplest (not only!) way to estimate standard error: redo the simulation several times, with same  $B$  and  $M$ , but from different initial values (from same “overdispersed starting distribution”).
  - Then can analyse the estimates obtained as iid ...
- (Comment: for big complicated  $\pi$ , often better to work with the LOGARITHMS, i.e. accept if  $\log(U_n) < \log(A_n) = \log(\pi(Y_n)) - \log(\pi(X_{n-1}))$ .)
- EXAMPLE:  $g(y) = y^3 \sin(y^4) \cos(y^5) \mathbf{1}_{0 < y < 1}$ .
  - Want to compute (again!)  $\mathbf{E}_\pi(h)$  where  $h(y) = y^2$ .
  - Use Metropolis algorithm with proposal  $Y \sim N(X, 1)$ . [file “Rmet”]
  - Works pretty well, but lots of variability!
  - Plot: appears to have “good mixing” ...
- EXAMPLE:  $\pi(x_1, x_2) = C |\cos(\sqrt{x_1 x_2})| I(0 \leq x_1 \leq 5, 0 \leq x_2 \leq 4)$ .
  - Want to compute  $\mathbf{E}_\pi(h)$ , where  $h(x_1, x_2) = e^{x_1} + (x_2)^2$ .
  - Metropolis algorithm ... works ... gets between about 34 and 44 ... but large uncertainty ... (file “Rmet2”) (Mathematica gets 38.7044)
  - Individual plots appear to have “good mixing” ...

– Joint plot shows fewer samples where  $x_1 x_2 \approx (\pi/2)^2 \doteq 2.5 \dots$

- OPTIMAL SCALING:

- Can change proposal distribution to  $Y_n \sim MVN(X_n, \sigma^2 I)$  for any  $\sigma > 0$ .
- Which is best?
- If  $\sigma$  too small, then chain won't move much.
- If  $\sigma$  too large, then will usually reject proposals, so chain still won't move much.
- Optimal: need  $\sigma$  “just right” to avoid both extremes. (“Goldilocks Principle”)
- Can experiment ... (“rwm.html” applet, files “Rmet”, “Rmet2”) ...
- Some theory ... limited ... active area of research ...
- General principle: the acceptance rate should be far from 0 and far from 1.
- In a certain idealised limit, optimal acceptance rate is 0.234 (!). [Roberts et al., Ann Appl Prob 1997; Roberts and Rosenthal, Stat Sci 2001]

## SO WHY DOES IT WORK?:

- (Need Markov chain theory ... STA447/2006 ... already know?)
- Let  $P(x, S) = \mathbf{P}(X_1 \in S | X_0 = x)$  be the transition probabilities.
  - Let  $\mathcal{X}$  be the state space of all possible values. (Usually  $\mathcal{X} \subseteq \mathbf{R}^d$ , e.g. for Variance Components Model,  $\mathcal{X} = (0, \infty) \times (0, \infty) \times \mathbf{R} \times \mathbf{R}^K \subseteq \mathbf{R}^{K+3}$ .)
  - Let  $q(x, y)$  be the proposal density for  $y$  given  $x$ . (So, in this case,  $q(x, y) = (2\pi\sigma)^{-d/2} \exp(-\sum_{i=1}^d (y_i - x_i)^2 / 2\sigma^2)$ .) Symmetric:  $q(x, y) = q(y, x)$ .
  - Let  $\alpha(x, y)$  be probability of accepting a proposed move from  $x$  to  $y$ , i.e.

$$\alpha(x, y) = \mathbf{P}(U_n < A_n | X_{n-1} = x, Y_n = y) = \min\left[1, \frac{\pi(y)}{\pi(x)}\right].$$

- Then if  $x \notin S$ , then

$$P(x, S) = \mathbf{P}(Y_1 \in S, U_1 < A_1 | X_0 = x) = \int_S q(x, y) \min[1, \pi(y)/\pi(x)] dy.$$

- Shorthand: for  $x \neq y$ ,  $P(x, dy) = q(x, y) \min[1, \pi(y)/\pi(x)] dy$ .
- Then for  $x \neq y$ ,  $P(x, dy) \pi(x) dx = q(x, y) \min[1, \pi(y)/\pi(x)] dy \pi(x) dx = q(x, y) \min[\pi(x), \pi(y)] dy dx = P(y, dx) \pi(y) dy$ . (symmetric)
- Follows that  $P(x, dy) \pi(x) dx = P(y, dx) \pi(y) dy$  for all  $x, y \in \mathcal{X}$ . (“reversible”)
- Shorthand:  $P(x, dy) \Pi(dx) = P(y, dx) \Pi(dy)$ .

---

**END WEEK #5**

---

[No class next week (Reading Week). HW#2 due March 2, 2:10 p.m.]

**Summary of Previous Class:**

\* MCMC, Metropolis Algorithm

\* choice of  $B, M, X_0, \sigma^2$

\* Theory:

—  $P(x, dy)$  so  $\mathbf{P}(X_{n+1} \in S | X_n = x) = \int_S P(x, dy)$ .

—  $\Pi(dx)$  so  $\int_S \pi(x) dx = \int_S \Pi(dx) = \Pi(S)$ .

— Then  $\Pi(dx) P(x, dy) = \Pi(dy) P(y, dx)$  (“reversible”).

- How does “reversible” help?
- Well, suppose  $X_0 \sim \Pi$ , i.e. we “start in stationarity”. Then

$$\begin{aligned} \mathbf{P}(X_1 \in S) &= \int_{x \in \mathcal{X}} \mathbf{P}(X_1 \in S | X_0 = x) \pi(x) dx = \int_{x \in \mathcal{X}} \int_{y \in S} P(x, dy) \pi(x) dx \\ &= \int_{x \in \mathcal{X}} \int_{y \in S} P(y, dx) \pi(y) dy = \int_{y \in S} \pi(y) dy \equiv \Pi(S), \end{aligned}$$

so also  $X_1 \sim \pi$ . So, chain “preserves”  $\pi$ , i.e.  $\pi$  is “stationary” density.

- Also irreducible, i.e. possible to eventually get anywhere.
  - More precisely: for every  $x \in \mathcal{X}$ , and every  $S \subseteq \mathcal{X}$  with  $\Pi(S) > 0$ , there is  $n$  such that  $P^n(x, S) > 0$ , i.e.  $\mathbf{P}(X_n \in S | X_0 = x) > 0$ . (Here, can even take  $n = 1$ .)
  - (Makes sense on discrete space, too; then requires ability to eventually reach every point of positive stationary measure; here “density” is with respect to “counting measure”.)

- And aperiodic, i.e. there do not exist disjoint subsets  $\mathcal{X}_1, \mathcal{X}_2, \dots, \mathcal{X}_j$  for  $j \geq 2$ , with  $\Pi(\mathcal{X}_i) > 0$ , such that  $P(x, \mathcal{X}_{i+1}) = 1$  for all  $x \in \mathcal{X}_i$  (where  $\mathcal{X}_{j+1} \equiv \mathcal{X}_1$ ). (Diagram.)
  - Aperiodicity always holds if  $P(x, \{x\}) > 0$ , i.e. if positive prob of rejection.
  - Or if  $P(x, \cdot)$  has positive density throughout  $S$ , for all  $x \in S$ , for some  $S \subseteq \mathcal{X}$  with  $\Pi(S) > 0$ .
  - Not quite guaranteed, e.g.  $\mathcal{X} = \{0, 1, 2, 3\}$ , and  $\pi$  uniform on  $\mathcal{X}$ , and  $Y_n = X_{n-1} \pm 1 \pmod{4}$ . But almost always holds.
- THEOREM: If Markov chain is irreducible, with stationary probability density  $\pi$ , then for  $\pi$ -a.e. initial value  $X_0 = x$ ,
  - (a) if  $\mathbf{E}_\pi(|h|) < \infty$ , then  $\lim_{n \rightarrow \infty} \frac{1}{n} \sum_{i=1}^n h(X_i) = \mathbf{E}_\pi(h) \equiv \int h(x) \pi(x) dx$ ; and
  - (b) if chain aperiodic, then  $\lim_{n \rightarrow \infty} \mathbf{P}(X_n \in S) = \int_S \pi(x) dx$  for all  $S \subseteq \mathcal{X}$ .
- So, key facts about  $q(x, y)$  are symmetry, and irreducibility.
  - So, could use e.g.  $Y_n \sim \text{Uniform}[X_{n-1} - 1, X_{n-1} + 1]$ , or (on discrete space)  $Y_n = X_{n-1} \pm 1$  with prob.  $\frac{1}{2}$  each.
- Why “ $\pi$ -a.e.” starting point??
- Example:  $\mathcal{X} = \{1, 2, 3, \dots\}$ , and  $P(1, \{1\}) = 1$ , and for  $x \geq 2$ ,  $P(x, \{1\}) = 1/x^2$  and  $P(x, \{x+1\}) = 1 - (1/x^2)$ .
  - Stationary distribution:  $\Pi(\cdot) = \delta_1(\cdot)$ , i.e.  $\Pi(S) = \mathbf{1}_{1 \in S}$  for  $S \subseteq \mathcal{X}$ .
  - Irreducible, since if  $\Pi(S) > 0$  then  $1 \in S$  so  $P(x, S) \geq P(x, \{1\}) > 0$  for all  $x \in \mathcal{X}$ .
  - Aperiodic since  $P(1, \{1\}) > 0$ .
  - So, by Theorem, for  $\pi$ -a.e.  $X_0$ , have  $\lim_{n \rightarrow \infty} \mathbf{P}(X_n \in S) = \Pi(S)$ , i.e.  $\lim_{n \rightarrow \infty} \mathbf{P}(X_n = 1) = 1$ .
  - But if  $X_0 = x \geq 2$ , then  $\mathbf{P}[X_n = x + n \text{ for all } n] = \prod_{j=x}^{\infty} (1 - (1/j^2)) > 0$ , so  $\lim_{n \rightarrow \infty} \mathbf{P}(X_n = 1) \neq 1$ .
  - Convergence holds if  $X_0 = 1$ , which is  $\pi$ -a.e. since  $\Pi(1) = 1$ , but not from

$$X_0 = x \geq 2.$$

- So, convergence subtle. But usually holds from any  $x \in \mathcal{X}$ . (“Harris recurrent”)

---

**END WEEK #6**

---

[HW#2 due March 2, 2:10 p.m. (sims may be tricky/unstable); Proj April 6. Office hours?]

[HW#2 Q1: replace “ $0 < x < 3$ ” by “ $0 < x < 1$ ”, and “[0,3]” by “[0,1]” (sorry!).]

**Summary of Previous Class:**

\* MCMC Theory (apparently harder than I thought ... )

\* Key: Metropolis Algorithm is “reversible”:

—  $\Pi(dx) P(x, dy) = \Pi(dy) P(y, dx)$

— i.e.,  $\pi(x) q(x, y) \alpha(x, y) = \pi(y) q(y, x) \alpha(y, x)$

— i.e., if start in stationarity ( $X_0 \sim \pi$ ), then density for the two events “ $X_0 = x, X_1 = y$ ” and “ $X_0 = y, X_1 = x$ ” are equal.

— Follows since  $q(x, y) = q(y, x)$ , and  $\alpha(x, y) = \min[1, \pi(y)/\pi(x)]$ .

\* Follows from this that  $\pi$  is stationary for the algorithm.

— i.e., if start in stationarity ( $X_0 \sim \pi$ ), then also  $X_1 \sim \pi$ .

\* Also (check) algorithm is irreducible (can eventually get anywhere), and aperiodic (no forced cycles).

\* THM: if irreducible, aperiodic, and  $\pi$  stationary, then from “almost every” starting point  $X_0 = x$  (not from  $X_0 \sim \pi$ ), the probabilities and averages converge to those of  $\pi$ .

- EXAMPLE #1: Metropolis algorithm where  $\mathcal{X} = \mathbf{Z}$ ,  $\pi(x) = 2^{-|x|}/3$ , and  $q(x, y) = \frac{1}{2}$  if  $|x - y| = 1$ , otherwise 0.

– Reversible? Yes, it’s a Metropolis algorithm!

–  $\pi$  stationary? Yes, follows from reversibility!

– Aperiodic? Yes, since  $P(x, \{x\}) > 0$ !

– Irreducible? Yes:  $\pi(x) > 0$  for all  $x \in \mathcal{X}$ , so can get from  $x$  to  $y$  in  $|x - y|$  steps.

– So, by theorem, probabilities and expectations converge to those of  $\pi$  – good.

- EXAMPLE #2: Same as #1, except now  $\pi(x) = 2^{-|x|-1}$  for  $x \neq 0$ , with  $\pi(0) = 0$ .

- Still reversible,  $\pi$  stationary, aperiodic, same as before.
- Irreducible? No – can't go from positive to negative!
- EXAMPLE #3: Same as #2, except now  $q(x, y) = \frac{1}{4}$  if  $1 \leq |x - y| \leq 2$ , otherwise 0.
  - Still reversible,  $\pi$  stationary, aperiodic, same as before.
  - Irreducible? Yes – can “jump over 0” to get from positive to negative, and back!
- EXAMPLE #4: Metropolis algorithm with  $\mathcal{X} = \mathbf{R}$ , and  $\pi(x) > 0$  for all  $x \in \mathcal{X}$ , and proposals  $Y_n \sim \text{Uniform}[X_{n-1} - 1, X_{n-1} + 1]$ .
  - Reversible? Yes since  $q(x, y)$  still symmetric.
  - $\pi$  stationary? Yes since reversible!
  - Irreducible? Yes since  $P^n(x, dy)$  has positive density whenever  $|y - x| \leq n$ .
  - Aperiodic? Yes since if periodic, then if e.g. if  $\mathcal{X}_1 \cap [0, 1]$  has positive measure, then possible to go from  $\mathcal{X}_1$  directly to  $\mathcal{X}_1$ , i.e. if  $x \in \mathcal{X}_1 \cap [0, 1]$ , then  $P(x, \mathcal{X}_1) > 0$ .
  - So, by theorem, probabilities and expectations converge to those of  $\pi$  – good.
- EXAMPLE #5: Same as #4, except now  $\pi(x) = 0$  for  $2 < x < 4$ , with  $\pi(x) > 0$  otherwise.
  - Still reversible and stationary and aperiodic, same as before.
  - But no longer irreducible: cannot jump from  $[4, \infty)$  to  $(-\infty, 2]$  or back.
  - So, does not converge.
- EXAMPLE #6: Same as #5, except now proposals are  $Y_n \sim \text{Uniform}[X_{n-1} - 5, X_{n-1} + 5]$ .
  - Still reversible and stationary and aperiodic, same as before.
  - And now irreducible, too: now can jump from  $[4, \infty)$  to  $(-\infty, 2]$  or back.
- EXAMPLE #7: Same as #6, except now  $Y_n \sim \text{Uniform}[X_{n-1} - 5, X_{n-1} + 10]$ .
  - Makes no sense – proposals not symmetric, so not a Metropolis algorithm!

## METROPOLIS-HASTINGS ALGORITHM:

- (Hastings [Canadian!], Biometrika 1970)
- Previous Metropolis algorithm works provided proposal distribution is symmetric, i.e.  $q(x, y) = q(y, x)$ . But what if it isn't?
- For Metropolis, key was that  $q(x, y) \alpha(x, y) \pi(x)$  was symmetric.
- If instead  $\alpha(x, y) = \min[1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}]$ , then

$$q(x, y) \alpha(x, y) \pi(x) = q(x, y) \min[1, \frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}] \pi(x) = \min[\pi(x)q(x, y), \pi(y)q(y, x)].$$

So, still symmetric, even if  $q$  wasn't.

- So, for Metropolis-Hastings algorithm, replace “ $A_n = \pi(Y_n) / \pi(X_{n-1})$ ” by  $A_n = \frac{\pi(Y_n)q(Y_n, X_{n-1})}{\pi(X_{n-1})q(X_{n-1}, Y_n)}$ , then still reversible, and everything else remains the same.
- i.e., still accept if  $U_n < A_n$ , otherwise reject.
- (Intuition: if  $q(x, y) \gg q(y, x)$ , then Metropolis chain would spend too much time at  $y$  and not enough at  $x$ , so need to accept fewer moves  $x \rightarrow y$ .)
- EXAMPLE: again  $\pi(x_1, x_2) = C |\cos(\sqrt{x_1 x_2})| I(0 \leq x_1 \leq 5, 0 \leq x_2 \leq 4)$ , and  $h(x_1, x_2) = e^{x_1} + (x_2)^2$ .
  - Proposal distribution:  $Y_n \sim MVN(X_{n-1}, \sigma^2 (1 + |X_{n-1}|^2)^2 I)$ .
  - (Intuition: larger proposal variance if farther from center.)
  - So,  $q(x, y) = C (1 + |x|^2)^{-2} \exp(-|y - x|^2 / 2 \sigma^2 (1 + |x|^2)^2)$ .
  - So, can run Metropolis-Hastings algorithm for this example. (file “RMH”)
  - Usually get between 34 and 43, with claimed standard error  $\approx 2$ . (Recall: Mathematica gets 38.7044.)
- INDEPENDENCE SAMPLER:
  - Proposals  $\{Y_n\}$  i.i.d. from some fixed distribution (say,  $Y_n \sim MVN(0, I)$ ). (Easy.)
  - Another special case of Metropolis-Hastings algorithm.



- Then  $q(x, y) = q(y)$ , depends only on  $y$ .
- So, now  $A_n = \frac{\pi(Y_n)q(X_{n-1})}{\pi(X_{n-1})q(Y_n)}$ .
- Very special case: if  $q(y) \equiv \pi(y)$ , i.e. propose exactly from target density  $\pi$ , then  $\alpha_n \equiv 1$ , i.e. make great proposals, and always accept them (iid).
- EXAMPLE: independence sampler with  $\pi(x) = e^{-x}$  and  $q(x) = ke^{-kx}$ .
  - Then if  $X_{n-1} = x$  and  $Y_n = y$ , then  $A_n = \frac{e^{-y}ke^{-kx}}{e^{-x}ke^{-ky}} = e^{(k-1)(y-x)}$ . (file “Rind”)
  - $k = 1$ : iid sampling (great).
  - $k = 0.01$ : proposals way too large (so-so).
  - $k = 5$ : proposals somewhat too small (terrible).
  - Why is large  $k$  so much worse than small  $k$ ?

---

**END WEEK #7**

---

[Collect HW#2; hand out HW#3; Project due April 6.]

**Summary of Previous Class:**

- \* Seven examples about reversible, stationary, irreducible, aperiodic
- \* Metropolis-Hastings algorithm (non-symmetric proposal increments)
- \* Independence sampler: proposals  $\{Y_n\}$  i.i.d.  $\sim q$ .
  - e.g.  $\pi(x) = e^{-x}$ ,  $q(x) = ke^{-kx}$ :
  - $k = 1$  iid,  $0 < k < 1$  okay,  $k > 1$  terrible
  - Why?

**MCMC CONVERGENCE RATES:**

- $\{X_n\}$  : Markov chain on  $\mathcal{X}$ , with stationary distribution  $\Pi(\cdot)$ .
- Let  $P^n(x, S) = \mathbf{P}[X_n \in S \mid X_0 = x]$ .
  - Hope that for large  $n$ ,  $P^n(x, S) \approx \Pi(S)$ .
- Let  $D(x, n) = \|P^n(x, \cdot) - \Pi(\cdot)\| \equiv \sup_{S \subseteq \mathcal{X}} |P^n(x, S) - \Pi(S)|$ .
- DEFN: chain is ergodic if  $\lim_{n \rightarrow \infty} D(x, n) = 0$ , for  $\Pi$ -a.e.  $x \in \mathcal{X}$ .

- Sometimes say chain “converges” in  $n$  iterations if  $D(x, n) < 0.01 \dots$
- previous theorem: chain ergodic if irreducible and aperiodic and  $\Pi(\cdot)$  stationary.
  - Hence: independence sampler is ergodic provided  $q(x) > 0$  whenever  $\pi(x) > 0$ .
- DEFN: chain is geometrically ergodic if there is  $\rho < 1$ , and  $M : \mathcal{X} \rightarrow [0, \infty]$  which is  $\Pi$ -a.e. finite, such that  $D(x, n) \leq M(x) \rho^n$  for all  $x \in \mathcal{X}$  and  $n \in \mathbf{N}$ .
- Quantitative bounds often difficult (though I’ve worked on them a lot), but “geometric ergodicity” often easier to verify.
- FACT: if state space is finite, and chain is irreducible and aperiodic, then always geometrically ergodic.
- FACT: independence sampler is geometrically ergodic IF AND ONLY IF there is  $\delta > 0$  with  $q(x) \geq \delta \pi(x)$  for  $\pi$ -a.e.  $x \in \mathcal{X}$ , in which case  $D(x, n) \leq (1 - \delta)^n$  for  $\pi$ -a.e.  $x \in \mathcal{X}$ .
  - So, if  $\pi(x) = e^{-x}$  and  $q(x) = ke^{-kx}$  for  $x > 0$ , where  $0 < k \leq 1$ , then can take  $\delta = k$ , so  $D(x, n) \leq (1 - k)^n$ .
  - e.g. if  $k = 0.01$ , then  $D(x, 459) \leq (0.99)^{459} \doteq 0.0099 < 0.01$  for all  $x > 0$ , i.e. “converges” after 459 iterations.
- But if  $k > 1$ , then not geometrically ergodic.
  - Fact: if  $k = 5$ , then  $D(0, n) > 0.01$  for all  $n \leq 4,000,000$ , while  $D(0, n) < 0.01$  for all  $n \geq 14,000,000$ , i.e. “convergence” takes between 4 million and 14 million iterations. Slow! [Roberts and Rosenthal (2008), “Quantitative Non-Geometric Convergence Bounds for Independence Samplers”.]
- What about for “random-walk Metropolis algorithm” (RWM), i.e. where  $\{Y_n - X_{n-1}\} \sim q$  for some fixed symmetric density  $q$ ?
  - e.g.  $Y_n \sim N(X_{n-1}, \sigma^2 I)$  or  $Y_n \sim \text{Uniform}[X_{n-1} - \delta, X_{n-1} + \delta]$ .
- FACT: RWM is geometrically ergodic essentially if and only if  $\pi$  has exponential tails, i.e. there are  $a, b, c > 0$  such that  $\pi(x) \leq ae^{-b|x|}$  whenever  $|x| > c$ . (Requires a few technical conditions:  $\pi$  and  $q$  continuous and positive, and  $q$  has finite first moment, and  $\pi$  log-concave in the tails ... )

[Mengersen and Tweedie, Ann Stat 1996; Roberts and Tweedie, Biometrika 1996]

- EXAMPLE: RWM on  $\mathbf{R}$  with usual proposals:  $Y_n \sim N(X_{n-1}, \sigma^2)$ .
  - Case #1:  $\Pi = N(5, 4^2)$ , and functional  $h(y) = y^2$ , so  $\mathbf{E}_\pi(h) = 5^2 + 4^2 = 41$ . (file “Rnorm” ...  $\sigma = 1$  v.  $\sigma = 4$  v.  $\sigma = 16$ )
  - Case #2:  $\pi(y) = c \frac{1}{(1+y^4)}$ , and functional  $h(y) = y^2$ , so

$$\mathbf{E}_\pi(h) = \frac{\int_{-\infty}^{\infty} y^2 \frac{1}{(1+y^4)} dy}{\int_{-\infty}^{\infty} \frac{1}{(1+y^4)} dy} = \frac{\pi/\sqrt{2}}{\pi/\sqrt{2}} = 1.$$

(file “Rheavy”)

- Case #3:  $\pi(y) = \frac{1}{\pi(1+y^2)}$  (Cauchy), and functional  $h(y) = \mathbf{1}_{-10 < y < 10}$ , so  $\mathbf{E}_\pi(h) = \Pi(|X| < 10) = 0.93655$ . [ $\Pi(X < x) = \arctan(x)/\pi$ .] (file “Rheavy2”)

### MCMC STANDARD ERROR:

- With MCMC, cannot use usual iid estimate of standard error.
- Simplest (HW#2): re-run the chain many times, with same  $M$  and  $B$ , with different initial values drawn from some overdispersed starting distribution, and compute standard error of the sequence of estimates.
- But how to estimate standard error from a single run?
- i.e., how to estimate  $V \equiv \text{Var} \left( \frac{1}{M-B} \sum_{i=B+1}^M h(X_i) \right)$ ?
  - Let  $\bar{h}(x) = h(x) - \mathbf{E}_\pi(h)$ , so  $\mathbf{E}_\pi(\bar{h}) = 0$ .
  - And, assume  $B$  large enough that  $X_i \approx \pi$  for  $i > B$ .
  - Then, for large  $M - B$ ,

$$\begin{aligned} V &\approx \mathbf{E}_\pi \left[ \left( \frac{1}{M-B} \sum_{i=B+1}^M h(X_i) - \mathbf{E}_\pi(h) \right)^2 \right] \approx \mathbf{E}_\pi \left[ \left( \frac{1}{M-B} \sum_{i=B+1}^M \bar{h}(X_i) \right)^2 \right] \\ &\approx \frac{1}{M-B} \left( \mathbf{E}_\pi(\bar{h}^2) + 2 \mathbf{E}_\pi(\bar{h}(X_i)\bar{h}(X_{i+1})) + 2 \mathbf{E}_\pi(\bar{h}(X_i)\bar{h}(X_{i+2})) + \dots \right) \\ &= \frac{1}{M-B} \mathbf{E}_\pi(\bar{h}^2) \left( 1 + 2 \text{Corr}_\pi(\bar{h}(X_i)\bar{h}(X_{i+1})) + 2 \text{Corr}_\pi(\bar{h}(X_i)\bar{h}(X_{i+2})) + \dots \right) \end{aligned}$$

$$\equiv \frac{1}{M-B} \text{Var}_\pi(h)(\text{varfact}) = (\text{iid variance})(\text{varfact}),$$

where

$$\text{varfact} = 1 + 2 \sum_{k=1}^{\infty} \text{Corr}_\pi(h(X_0)h(X_k)) \equiv 1 + 2 \sum_{k=1}^{\infty} \rho_k \equiv \sum_{k=-\infty}^{\infty} \rho_k$$

(“integrated auto-correlation time”). (Included in previous R files.)

- Note: to compute varfact, don’t sum over all  $k$ , just e.g. until, say,  $|\rho_k| < 0.05$  or  $\rho_k < 0$  or ...
- (Previous R programs used built-in “acf” function, but on HW#3 you should write your own computation – better.)
- Usually varfact  $\gg 1$ ; try to get “better” chains so varfact smaller.
- Sometimes even try to design chain to get varfact  $< 1$  (“antithetic”).

---

**END WEEK #8**

---

[HW#3 due March 23; Test March 30; Project due April 6]

**Summary of Previous Class:**

\*  $D(x, n)$ , ergodicity

\* geometric ergodicity

— independence sampler: iff  $q(x) \geq \delta\pi(x)$

— RWM: “iff”  $\pi$  has exponentially-small tails

\* approximate variance of MCMC estimator:

—  $\frac{1}{M-B} \text{Var}_\pi(h)(\text{varfact})$

—  $\text{varfact} = 1 + 2 \sum_{k=1}^{\infty} \rho_k$

— in practice, just take FINITE sum (or downweight larger  $k$ )

**CONFIDENCE INTERVALS:**

- Suppose we estimate  $u \equiv \mathbf{E}_\pi(h)$  by  $\frac{1}{M-B} \sum_{i=B+1}^M h(X_i)$ , and obtain an estimate  $e$  and an approximate variance (as above)  $v$ .
- Then what is, say, a 95% confidence interval for  $u$ ?
- Well, suppose first that the  $\{X_i\}$  are i.i.d., with  $\text{Var}_\pi(h) < \infty$ .

- Then, by the classical central limit theorem (CLT), for large  $M - B$ ,  $e \approx N(u, v)$  where  $v = \frac{1}{M-B} \text{Var}_\pi(h)$ .
- So  $(e - u) v^{-1/2} \approx N(0, 1)$ .
- So,  $\mathbf{P}(-1.96 < (e - u) v^{-1/2} < 1.96) \approx 0.95$ .
- So,  $\mathbf{P}(-1.96 \sqrt{v} < e - u < 1.96 \sqrt{v}) \approx 0.95$ .
- i.e., with prob 95%,  $u$  is in the interval  $(e - 1.96 \sqrt{v}, e + 1.96 \sqrt{v})$ .
- Of course, don't actually know  $v$ ; must estimate it empirically.
- So, strictly speaking, should use “t” distribution, not normal distribution ... but if  $M - B$  large that doesn't really matter (ignore it for now).
- What if  $\{X_i\}$  is not i.i.d.??
  - First of all, must replace  $v = \frac{1}{M-B} \text{Var}_\pi(h)$  by  $v = \frac{1}{M-B} \text{Var}_\pi(h)$ (varfact)
  - But does a CLT even hold??
  - Does not follow from classical CLT. Does not always hold. But often does.
  - For example, CLT holds if chain is geometrically ergodic and  $\mathbf{E}_\pi(|h|^{2+\delta}) < \infty$  for some  $\delta > 0$ .
  - (If chain also reversible then don't need  $\delta$ : Roberts and Rosenthal, “Geometric ergodicity and hybrid Markov chains”, ECP 1997.)
- Previous “Rnorm” example:  $\Pi = N(5, 4^2)$ ,  $h(y) = y^2$ ,  $\mathbf{E}_\pi(h) = 41$ .
  - Does CLT hold? Yes! (geometrically ergodic, and  $\mathbf{E}_\pi(|h|^p) < \infty$  for all  $p$ .)
  - Indeed, confidence intervals “usually” contain 41. (file “Rnorm2”)
- Previous “Rind” example:  $\Pi = \text{Exponential}(1)$ ,  $h(y) = y$ ,  $\mathbf{E}_\pi(h) = 1$ ,  $Q = \text{Exponential}(5)$ .
  - Not geometrically ergodic.
  - CLT known not to hold.
  - Confidence intervals often miss 1. (file “Rind2”)

- Previous “Rheavy2” example:  $\Pi = \text{Cauchy}$ ,  $h(y) = \mathbf{1}_{-10 < y < 10}$ ,  $\mathbf{E}_\pi(h) = 0.93655$ .
  - Not geometrically ergodic.
  - Confidence intervals often miss 0.93655. (file “Rheavy3”)
- Even when CLT holds, it’s rather unstable, e.g. requires that chain has converged to  $\Pi$ , and might underestimate  $v$ .
- Estimate of  $v$  is very important (HW #3!).
  - Another approach is “batch means”, whereby chain is broken into  $m$  “batches” which are assumed to be i.i.d.

## VARIABLE-AT-A-TIME MCMC:

- Propose to move just one coordinate at a time, leaving all the other coordinates fixed (since changing all coordinates at once may be difficult).
- Then accept/reject with usual Metropolis-Hastings rule. (“Metropolis-within-Gibbs”)
- Can either choose coordinates in sequence (“systematic-scan”), or uniformly at random each time (“random-scan”).
- EXAMPLE: again  $\pi(x_1, x_2) = C |\cos(\sqrt{x_1 x_2})| I(0 \leq x_1 \leq 5, 0 \leq x_2 \leq 4)$ , and  $h(x_1, x_2) = e^{x_1} + (x_2)^2$ .
  - Works with systematic-scan (file “Rmwg”) or random-scan (file “Rmwg2”).
- GIBBS SAMPLER: Proposal distribution for  $i^{\text{th}}$  coordinate is equal to the conditional distribution of that coordinate (according to  $\pi$ ), conditional on the current values of all the other coordinates.
  - Then  $q_i(x, y) = C(x^{(-i)}) \pi(y)$  whenever  $x^{(-i)} = y^{(-i)}$ , where  $x^{(-i)}$  means all coordinates except  $i^{\text{th}}$  one.
  - Here  $C(x^{(-i)})$  is the appropriate normalising constant (which depends on  $x^{(-i)}$ ). (So  $C(x^{(-i)}) = C(y^{(-i)})$ .)
  - Then  $A_n = \frac{\pi(Y_n) q_i(Y_n, X_{n-1})}{\pi(X_{n-1}) q_i(X_{n-1}, Y_n)} = \frac{\pi(Y_n) C(Y_n^{(-i)}) \pi(X_{n-1})}{\pi(X_{n-1}) C(X_{n-1}^{(-i)}) \pi(Y_n)} = 1$ .

– So, always accept.

• **EXAMPLE:** Variance Components Model (with  $J_i \equiv J$ ):

– Update of  $\mu$  (say) should be from conditional density of  $\mu$ , conditional on current values of all the other coordinates:  $\mathcal{L}(\mu | V, W, \theta_1, \dots, \theta_K, Y_{11}, \dots, Y_{JK})$ .

– This conditional density is proportional to the full joint density, but with everything except  $\mu$  treated as constant.

– Full joint density is:

$$= C e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-JK/2} \times \\ \times \exp \left[ - \sum_{i=1}^K (\theta_i - \mu)^2 / 2V - \sum_{i=1}^K \sum_{j=1}^J (Y_{ij} - \theta_i)^2 / 2W \right].$$

– So, conditional density of  $\mu$  is

$$C_2 e^{-(\mu-a_3)^2/2b_3} \exp \left[ - \sum_{i=1}^K (\theta_i - \mu)^2 / 2V \right].$$

– This equals

$$C_3 \exp \left( - \mu^2 \left( \frac{1}{2b_3} + \frac{K}{2V} \right) + \mu \left( \frac{a_3}{b_3} + \frac{1}{V} \sum_{i=1}^K \theta_i \right) \right).$$

– This is  $N(m, v)$ , where  $1/2v = \frac{1}{2b_3} + \frac{K}{2V}$  and  $m/v = \frac{a_3}{b_3} + \frac{1}{V} \sum_{i=1}^K \theta_i$ .

– Solve:  $v = b_3 V / (V + Kb_3)$ , and  $m = (a_3 V + b_3 \sum_{i=1}^K \theta_i) / (V + Kb_3)$ .

– So, in Gibbs Sampler, each time  $\mu$  is updated, we sample it from  $N(m, v)$  for this  $m$  and  $v$  (and always accept).

• Similarly, conditional distribution for  $V$  is:

$$C_4 e^{-b_1/V} V^{-a_1-1} V^{-K/2} \exp \left[ - \sum_{i=1}^K (\theta_i - \mu)^2 / 2V \right], \quad V > 0.$$

– Recall that “ $IG(r, s)$ ” has density  $\frac{s^r}{\Gamma(r)} e^{-s/x} x^{-r-1}$  for  $x > 0$ .

- So, conditional distribution for  $V$  equals  $IG(a_1 + K/2, b_1 + \frac{1}{2} \sum_{i=1}^K (\theta_i - \mu)^2)$ .
- And, conditional dist. for  $W$  equals  $IG(a_2 + KJ/2, b_2 + \frac{1}{2} \sum_{i=1}^K \sum_{j=1}^J (Y_{ij} - \theta_i)^2)$ .
- And, for  $\theta_i$  ( $1 \leq i \leq K$ ), conditional distribution (exercise: verify this!) is:

$$N\left(\frac{V \sum_{j=1}^J Y_{ij} + W \mu}{JV + W}, \frac{VW}{JV + W}\right).$$

- So, in this case, the systematic-scan Gibbs sampler proceeds by:
  - Update  $V$  from its conditional distribution  $IG(\dots, \dots)$ .
  - Update  $W$  from its conditional distribution  $IG(\dots, \dots)$ .
  - Update  $\mu$  from its conditional distribution  $N(\dots, \dots)$ .
  - Update  $\theta_i$  from its conditional distribution  $N(\dots, \dots)$ , for  $i = 1, 2, \dots, K$ .
  - Repeat all of the above  $M$  times.

---

**END WEEK #9**

---

[Do course evals ... ]

[HW#3 due March 23 [don't change priors/data!]; Test March 30; Project due April 6]

**Summary of Previous Class:**

\* 95% confidence intervals for MCMC:

—  $(e - 1.96\sqrt{v}, e + 1.96\sqrt{v})$ , if CLT holds (e.g. geometrically ergodic, and  $\mathbf{E}_\pi(|h|^{2+\delta}) < \infty$ ), and  $B$  and  $M - B$  are large enough ...

\* variable-at-a-time MCMC:

— propose to change just one coord per iteration

— then cycle through coordinates (systematic or random)

— Gibbs sampler: propose from current cond. dist., always accept

— e.g. Variance Components Model: cond. dists. are  $N$  and  $IG$  ...



## TEMPERED MCMC:

- Suppose  $\Pi(\cdot)$  is multi-modal, i.e. has distinct “parts” (e.g.,  $\Pi = \frac{1}{2} N(0, 1) + \frac{1}{2} N(20, 1)$ )
- Usual RWM with  $Y_n \sim N(X_{n-1}, 1)$  (say) can explore well within each mode, but how to get from one mode to the other?
- Idea: if  $\Pi(\cdot)$  were flatter, e.g.  $\frac{1}{2} N(0, 8^2) + \frac{1}{2} N(20, 8^2)$ , then easier to get between modes.
- So: define a sequence  $\Pi_1, \Pi_2, \dots, \Pi_m$  where  $\Pi_1 = \Pi$  (“cold”), and  $\Pi_\tau$  is flatter for larger  $\tau$  (“hot”).
- Then define Markov chain on  $\mathcal{X} \times \{1, 2, \dots, m\}$ , with stationary distribution  $\bar{\Pi}$  defined by  $\bar{\Pi}(S \times \{\tau\}) = \frac{1}{m} \Pi_\tau(S)$ .
  - (Can also use other weights besides  $\frac{1}{m}$ .)
- Markov chain has both spatial moves (change  $x$ ) and temperature moves (change  $\tau$ ).
- In the end, only “count” those samples where  $\tau = 1$ .
- EXAMPLE:  $\Pi = \frac{1}{2} N(0, 1) + \frac{1}{2} N(20, 1)$ 
  - Assume proposals are  $Y_n \sim N(X_{n-1}, 1)$ .
  - Mixing for  $\Pi$ : terrible! (file “Rtempered”; note the small claimed standard error!)
  - Define  $\Pi_\tau = \frac{1}{2} N(0, \tau^2) + \frac{1}{2} N(20, \tau^2)$ , for  $\tau = 1, 2, \dots, 8$ .
  - Mixing for  $\Pi_8$ : good! (file “Rtempered”)
  - (Compare graphs of  $\pi_1$  and  $\pi_8 \dots$ )
  - So, define new algorithm which alternates between:
    - (a) propose  $x' \sim N(x, 1)$ , accept with prob  $\min\left(1, \frac{\pi_\tau(x')}{\pi_\tau(x)}\right)$ .
    - (b) propose  $\tau' = \tau \pm 1$  (prob  $\frac{1}{2}$  each), accept with prob  $\min\left(1, \frac{\pi_{\tau'}(x)}{\pi_\tau(x)}\right)$ .
  - This converges fairly well to  $\bar{\Pi}$ . (file “Rtempered”, with “dotempering = TRUE”)
  - So, conditional on  $\tau = 1$ , converges to  $\Pi$ .

– So, average of those  $h(x)$  with  $\tau = 1$  gives good estimate of  $\mathbf{E}_\pi(h)$ .

- HOW TO FIND THE TEMPERED DENSITIES  $\pi_\tau$ ?

- Usually won't "know" about e.g.  $\Pi_\tau = \frac{1}{2} N(0, \tau^2) + \frac{1}{2} N(20, \tau^2)$ .

- Instead, can e.g. let  $\pi_\tau(x) = c_\tau (\pi(x))^{1/\tau}$ . (Sometimes write  $\beta = 1/\tau$ .)

- Then  $\Pi_1 = \Pi$ , and  $\pi_\tau$  flatter for larger  $\tau$  – good.

- (e.g. if  $\pi(x)$  density of  $N(\mu, \sigma^2)$ , then  $c_\tau (\pi(x))^{1/\tau}$  density of  $N(\mu, \tau\sigma^2)$ .)

- Then temperature acceptance probability is:

$$\min\left(1, \frac{\pi_{\tau'}(x)}{\pi_\tau(x)}\right) = \min\left(1, \frac{c_{\tau'}}{c_\tau} (\pi(x))^{(1/\tau') - (1/\tau)}\right).$$

- This depends on the  $c_\tau$ , which are usually unknown – bad.

- PARALLEL TEMPERING (a.k.a. Metropolis-Coupled MCMC, or MCMCMC):

- Instead, use state space  $\mathcal{X}^m$ , with  $m$  chains, i.e. one chain for each temperature.

- So, state at time  $n$  is  $X_n = (X_{n1}, X_{n2}, \dots, X_{nm})$ .

- And, now stationary distribution is  $\bar{\Pi} = \Pi_1 \times \Pi_2 \times \dots \times \Pi_m$ , i.e.  $\bar{\Pi}(X_1 \in S_1, X_2 \in S_2, \dots, X_m \in S_m) = \Pi_1(S_1) \Pi_2(S_2) \dots \Pi_m(S_m)$ .

- Then, can update the chain at temperature  $\tau$  (for each  $1 \leq \tau \leq m$ ), by proposing e.g.  $Y_{n,\tau} \sim N(X_{n-1,\tau}, 1)$ , and accepting with probability  $\min\left(1, \frac{\pi_\tau(Y_{n,\tau})}{\pi_\tau(X_{n-1,\tau})}\right)$ .

- And, can also choose temperatures  $\tau$  and  $\tau'$  (e.g., at random), and propose to "swap" the values  $X_{n,\tau}$  and  $X_{n,\tau'}$ , and accept this with probability  $\min\left(1, \frac{\pi_\tau(X_{n,\tau'}) \pi_{\tau'}(X_{n,\tau})}{\pi_\tau(X_{n,\tau}) \pi_{\tau'}(X_{n,\tau'})}\right)$ .

- Now, normalising constants cancel, e.g. if  $\pi_\tau(x) = c_\tau (\pi(x))^{1/\tau}$ , then acceptance probability is  $\min\left(1, \frac{\pi(X_{n,\tau'})^{1/\tau} \pi(X_{n,\tau})^{1/\tau'}}{\pi(X_{n,\tau})^{1/\tau} \pi(X_{n,\tau'})^{1/\tau'}}\right)$ , with no  $c_\tau$  or  $c_{\tau'}$  required.

- EXAMPLE: suppose again that  $\Pi_\tau = \frac{1}{2} N(0, \tau^2) + \frac{1}{2} N(20, \tau^2)$ , for  $\tau = 1, 2, \dots, 8$ .

- Can run parallel tempering ... (file "Rpara")

---

**END WEEK #10**

[Collect HW#3: Rdy posterior means  $V \approx 200,000$ ,  $W \approx 6,000$ ,  $\mu \approx 0.18$ ,  $\theta_i \approx 1500$ .]

[Reminders: test next week (90 mins; R), project due week after (2:10).]

### Summary of Previous Class:

\* (Did course evals – thanks everyone, esp. Avery.)

\* Tempered MCMC:

— Interpolate heavily multimodal  $\Pi$  to flatter dist. by  $\Pi_1 = \Pi, \Pi_2, \dots, \Pi_m$

— use M-H with both  $x$  moves and  $\tau$  moves

— e.g.  $\pi_\tau(x) = c_\tau (\pi(x))^{1/\tau}$

— use Parallel Tempering to avoid computing  $c_\tau$

### TRANSDIMENSIONAL MCMC:

- (a.k.a. “reversible-jump MCMC”: Green, Biometrika 1995)
- What if the state space is a union of parts of different dimension?
  - Can we still apply Metropolis-Hastings then??
- EXAMPLE: suppose  $\{y_j\}_{j=1}^J$  are known data which come from a mixture distribution  $\frac{1}{k}(N(a_1, 1) + N(a_2, 1) + \dots + N(a_k, 1))$ .
- Want to estimate unknown  $k, a_1, \dots, a_k$ ; here the number of parameters is also unknown.
- The state space is  $\mathcal{X} = \{(k, a) : k \in \mathbf{N}, a \in \mathbf{R}^k\}$ .
- Prior distributions:  $k - 1 \sim \text{Poisson}(2)$ , and  $a|k \sim \text{MVN}(0, I_k)$  (say).
- Let  $\lambda = \lambda_0 \times (\lambda_1 + \lambda_2 + \lambda_3 + \dots)$ , where  $\lambda_0$  is counting measure (for  $k$ ), and for  $i \geq 1$ ,  $\lambda_i$  is Lebesgue measure on  $\mathbf{R}^i$ .
- Then posterior density (with respect to  $\lambda$ ) is:

$$\pi(k, a) = c \frac{e^{-2} 2^{k-1}}{(k-1)!} (2\pi)^{-k/2} \exp\left(-\frac{1}{2} \sum_{i=1}^k a_i^2\right) (2\pi)^{-J/2} \prod_{j=1}^J \left(\sum_{i=1}^k \frac{1}{k} \exp\left(-\frac{1}{2}(y_j - a_i)^2\right)\right).$$

- So, on a log scale,

$$\log \pi(k, a) = \log c + \log \frac{e^{-2} 2^{k-1}}{(k-1)!} - \frac{k}{2} \log(2\pi) - \frac{1}{2} \sum_{i=1}^k a_i^2 - \frac{J}{2} \log(2\pi) + \sum_{j=1}^J \log \left( \sum_{i=1}^k \frac{1}{k} \exp \left( -\frac{1}{2} (y_j - a_i)^2 \right) \right).$$

(Can ignore  $\frac{J}{2} \log(2\pi)$ , but not  $\frac{k}{2} \log(2\pi)$ .)

- How to “explore” this posterior distribution??
- For fixed  $k$ , can move around  $\mathbf{R}^k$  in usual RWM way.
- But how to change  $k$ ?
- Can propose to replace  $k$  with, say,  $k' = k \pm 1$  (prob  $\frac{1}{2}$  each).
- Then have to correspondingly change  $a$ . For example:
  - If  $k' = k + 1$ , then  $a' = (a_1, \dots, a_k, Z)$  where  $Z \sim N(0, 1)$  (elongate).
  - If  $k' = k - 1$ , then  $a' = (a_1, \dots, a_{k-1})$  (truncate).
- Then accept with usual probability,  $\min \left( 1, \frac{\pi(k', a') q((k', a'), (k, a))}{\pi(k, a) q((k, a), (k', a'))} \right)$ .
  - Here if  $k' = k + 1$ , then  $q((k', a'), (k, a)) = \frac{1}{2}$ , while  $q((k, a), (k', a')) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-(a'_k)^2/2}$ .
  - Or, if  $k' = k - 1$ , then  $q((k, a), (k', a')) = \frac{1}{2}$ , while  $q((k', a'), (k, a)) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-(a_k)^2/2}$ .
- Seems to work okay; final  $k$  usually between 5 and 9 ... (file “Rtrans”)
- Alternative method for “correspondingly change  $a$ ” step:
  - If  $k' = k + 1$ , then  $a' = (a_1, \dots, a_{k-1}, a_k - Z, a_k + Z)$  where  $Z \sim N(0, 1)$  (“split”).
  - If  $k' = k - 1$ , then  $a' = (a_1, \dots, a_{k-2}, \frac{1}{2}(a_{k-1} + a_k))$  (“merge”).
  - What about the densities  $q((k', a'), (k, a))$ ?
  - Well, if  $k' = k + 1$ , then  $q((k', a'), (k, a)) = \frac{1}{2}$ , while roughly speaking,

$$q((k, a), (k', a')) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-(\frac{1}{2}(a'_k - a_k))^2/2}.$$

- One subtle additional point: The map  $(a, Z) \mapsto a' = (a_1, \dots, a_{k-1}, a_k - Z, a_k + Z)$  has “Jacobian” term:

$$\det \left( \frac{\partial a'}{\partial (a, Z)} \right) = \det \begin{pmatrix} I_{k-1} & 0 & 0 \\ 0 & 1 & -1 \\ 0 & 1 & 1 \end{pmatrix} = 1 - (-1) = 2,$$

i.e. the split moves “spread out” the mass by a factor of 2.

- So by Change-of-Variable Thm, actually

$$q((k, a), (k', a')) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a'_{k'} - a'_k)^2/2} / 2.$$

- Similarly, if  $k' = k - 1$ , then  $q((k, a), (k', a')) = \frac{1}{2}$ , while

$$q((k', a'), (k, a)) = \frac{1}{2} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}(a_k - a_{k'})^2/2} / 2.$$

- Algorithm still seems to work okay ... (file “Rtrans2”)
- For more complicated transformations, need to include more complicated “Jacobian” term (but above it equals 1 or 2).

## OPTIMAL RWM PROPOSALS:

- Consider RWM on  $\mathcal{X} = \mathbf{R}^d$ , where  $Y_n \sim MVN(X_{n-1}, \Sigma)$  for some  $d \times d$  proposal covariance matrix  $\Sigma$ .
- What is best choice of  $\Sigma$ ??
  - Usually we take  $\Sigma = \sigma^2 I_d$  for some  $\sigma > 0$ , and then choose  $\sigma$  so acceptance rate not too small, not too large.
  - But can we do better?
- Suppose for now that  $\Pi = MVN(\mu_0, \Sigma_0)$  for some fixed  $\mu_0$  and  $\Sigma_0$ . Try RWM:
  - first version:  $Y_n \sim MVN(X_{n-1}, I_d)$ . (file “Ropt”)
  - second version:  $Y_n \sim MVN(X_{n-1}, 0.1 I_d)$ . (file “Ropt”)
  - third version:  $Y_n \sim MVN(X_{n-1}, \Sigma_0)$ . (file “Ropt”)

- fourth version:  $Y_n \sim MVN(X_{n-1}, 1.4 \Sigma_0)$ . (file “Ropt”)
- Conclusion: acceptance rates near 0.234 (or 0.25 . . . ) are better.
- But also, proposals shaped like target are better.
  - Proved for targets which are orthogonal transformations of independent components (Roberts et al., Ann Appl Prob 1997; Roberts and Rosenthal, Stat Sci 2001; Bédard, Ann Appl Prob 2007).
  - Seems to be “approximately” true for most unimodal targets . . .
- Problem:  $\Sigma_0$  would usually be unknown; then what?
  - Can perhaps “adapt“!

---

**END WEEK #11**

---

[Return HW#3. (Total # student visits in my THREE office hours: 1.)]

[Write in-class test. (Will post solutions.)]

---

**END WEEK #12**

---

[Collect projects. Return tests (solutions posted; mean 33/45; last question harder).]

[Will grade projects soon . . . should I post grades by student #?]

[Did you make much use of the reference books (library / purchase / on-line)?]

**Summary of Previous Class:**

\* Transdimensional MCMC:

— Version of Metropolis-Hastings when state space dimension changes

— Need to design moves to still preserve reversibility

\* Optimal RWM proposals:

— acceptance rate near 0.234

— proposal covariance proportional to target covariance

## ADAPTIVE MCMC:

- What if target covariance is unknown??
- Can estimate target covariance based on run so far.
- Then update proposal covariance “on the fly”
- Adaptive Metropolis (AM) algorithm (e.g. figure “plotAMx200.png”)
- But is this a valid algorithm??
- Not in general: e.g. “adapt.html”
- Algorithm now non-Markovian, doesn’t preserve stationarity at each step.
- However, still converges to  $\Pi$  provided that the adaption (i) is “diminishing” and (ii) satisfies a technical condition called “containment”.
  - For details see e.g. Roberts & Rosenthal, “Coupling and Convergence of Adaptive MCMC” (J. Appl. Prob. 2007).

## MONTE CARLO OPTIMISATION:

- EXAMPLE: CODE BREAKING, e.g. “decipherit oliver”. [“decipher.c”]
  - “substitution cipher”.
- Data is the coded message text:  $s_1 s_2 s_3 \dots s_N$ , where  $s_i \in \mathcal{A} = \{A, B, C, \dots, Z, \text{space}\}$ .
- State space  $\mathcal{X}$  is set of all bijections of  $\mathcal{A}$ , i.e. one-to-one onto mappings  $f : \mathcal{A} \rightarrow \mathcal{A}$ , subject to  $f(\text{space}) = \text{space}$ .
- Use reference text (e.g. “War and Peace”) to get matrix  $M(x, y) =$  number of times  $y$  follows  $x$ , for  $x, y \in \mathcal{A}$ .
- Then for  $f \in \mathcal{X}$ ,  $\pi(f) = \prod_{i=1}^{N-1} M(f(s_i), f(s_{i+1}))$ .
- Run Metropolis algorithm:
  - Choose  $a, b \in \mathcal{A}$  uniformly at random.
  - Propose to replace  $f$  by  $g$ , where  $g(a) = f(b)$ ,  $g(b) = f(a)$ , and  $g(x) = f(x)$  for

$x \neq a, b$ .

– Accept with probability  $\min\left(1, \frac{\pi(g)}{\pi(f)}\right)$ .

- Still irreducible, aperiodic, reversible.
- So, converges (quickly!) to correct answer, thus breaking the code!
- References: S. Conner (2003), “Simulation and solving substitution codes”. P. Diaconis (2008), “The Markov Chain Monte Carlo Revolution”.
- EXAMPLE: COMPUTER VISION, e.g. “faces” Java applet. [“faces.html”]
- Data is an image, given in terms of pixels (light or dark).
- Define the face location by various parameters  $\theta$  (face center, eye width, nose height, etc.).
- Define a score function  $S(\theta)$  for how well the image agrees with having a face in the location corresponding to the parameters  $\theta$ .
- Run “mixed” Monte Carlo search over parameter space, looking for  $\operatorname{argmax}_{\theta} S(\theta)$ .
- For details, see Java applet source code, “faces.java” (or related paper).
- In both these examples, wanted to MAXIMIZE  $\pi$  rather than SAMPLE from  $\pi$ .
  - General method??
- SIMULATED ANNEALING:
  - General method to find mode of  $\pi$ .
  - Use tempered MCMC, but where  $\tau = \tau_n \searrow 0$ , so  $\Pi_{\tau_n}$  becomes more and more concentrated at mode as  $n \rightarrow \infty$ .
  - Fact: for slow enough “cooling schedule”, the chain should converge to the mode.
- So, in summary, Monte Carlo is used for nearly everything!
- Good luck on your exams, etc., and have a nice summer.

---

**END WEEK #13**

---