STA3431 (Monte Carlo Methods) Lecture Notes, Winter 2010

by Jeffrey S. Rosenthal, University of Toronto

(Last updated: March 29, 2010.)

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
  - If not Stat Dept grad student, must REQUEST enrolment (by e-mail); need strong probability/statistics background, plus some computer programming experience.
  - 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 $z_1, \ldots, z_n$ of Normal(0,1) random variables, and let $x_i = z_i^4 \cos(z_i^3)$.
  - Their mean $\overline{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i$ 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 $\overline{x}$ close to -0.085 ...
  - Variability??

– Well, can estimate standard deviation of $\overline{x}$ by "standard error" of $\overline{x}$, which is:

$$se \;=\; n^{-1/2}\,\mathrm{sd}(x) \;=\; n^{-1/2}\,\sqrt{\mathrm{var}(x)} \;=\; n^{-1/2}\,\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i-\overline{x})^2}\;.$$

[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? (Can't do numerical integration!)

- 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! ... including Metropolis, Gibbs, tempered, trans-dimensional, ...)

- OTHER EXAMPLES:

- 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 service times $\sim$ Exponential(μ) [mean $1/\mu$], and interarrival times $\sim$ Exponential(λ) ("M/M/1 queue"), so $\{Q(t)\}$ Markovian. Then well known:

– If $\mu \leq \lambda$, then $Q(t) \to \infty$ as $t \to \infty$.

– If $\mu > \lambda$, then $Q(t)$ converges in distribution as $t \to \infty$:

– $\mathbf{P}(Q(t) = i) \to (1 - \frac{\lambda}{\mu})(\frac{\lambda}{\mu})^i$, for $i = 0, 1, 2, \ldots$.

– Easy! (e.g. $\mu = 3$, $\lambda = 2$, $t = 1000$) [file "Rqueue"]

- Now suppose instead 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"]
- Or, to make the means the same as the first example, suppose service times $\sim$ Uniform$[0,2/3]$, and interarrival times have distribution of $Z^2/2$ where $Z \sim$ Normal$(0,1)$. Now what? [file "Rqueue3"]
- HISTORICAL EXAMPLE – BUFFON'S NEEDLE:
  - Have series of parallel lines ... line spacing w, needle length $\ell \leq w$ ... what is prob that needle lands touching line? [file buffon.html]
  - Let θ be angle counter-clockwise from line direction, and h distance of top end above nearest line.
  - Then $h \sim$ Uniform$[0,w]$ and $\theta \sim$ 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 $\pi \approx 2n\ell/mw \doteq 3.1423$.]

PSEUDORANDOM NUMBERS:

- Goal: generate an i.i.d. sequence $U_1, U_2, U_3, \ldots \sim$ 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, $0 \le x_n \le m-1$.
  - 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; "dieharder", www.phy.duke.edu/$\sim$rgb/General/dieharder.php)
  - One common "good" choice: $m = 2^{32}$, $a = 69,069$, $b = 23,606,797$.
- Thm: has full period (m) iff $\gcd(b, m) = 1$, and every "prime or 4" divisor of m also divides $a - 1$.
  - So, if $m = 2^{32}$, then if b odd and $a - 1$ is a multiple of 4, then has full period $m = 2^{32} \doteq 4 \times 10^9$; good.
- 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 / entropy of kernel (e.g. Linux's "urandom").
  - Or the randomness of *quantum mechanics*, e.g. `www.fourmilab.ch/hotbits`.
  - Or of atmospheric noise, e.g. `random.org`.
  - But for most purposes, standard pseudorandom numbers are pretty good ...

- 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 $x_{n+2} = 6x_{n+1} - 9x_n \bmod m$ ... 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 $1 \le s < k$ where $2^{kw-r} - 1$ is Mersenne prime, and A is $w \times w$ (e.g. 32×32) 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.)

——————— **END WEEK #1**———————

[Reminder: e-mail me if you're from another dept (not Stats) and want to take this class for credit.]

Summary of Previous Class:

∗ Examples of Monte Carlo:

—— $\mathbf{E}[Z^4 \cos(Z^3)]$

—— computer vision, code breaking, Buffon's needle, queue simulation, ...

∗ Pseudorandom number generation:

—— Want $U_1, U_2, \ldots \approx$ 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. $m = 2^{32}$, $a = 69,069$, $b = 23,606,797$.

—— THM: full period iff ...

—— RNG tests ...

SIMULATING OTHER DISTRIBUTIONS:

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

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

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

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

or set

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

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

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

$$Y \;=\; \max\{x_j \,;\, \sum_{k=1}^{j-1} p_k \;\le\; U_1\}\,.$$

- e.g. to make $Z \sim$ 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 $W \sim$ Exponential(λ), 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 $f_{X,Y}(x,y) = f_{U_1,U_2}\big(h^{-1}(x,y)\big) \,/\, |J\big(h^{-1}(x,y)\big)|$. Here $f_{U_1,U_2}(u_1,u_2) = 1$ for $0 < u_1, u_2 < 1$ (otherwise 0), and

$$J(u_1,u_2) \;=\; \det\begin{pmatrix} \frac{\partial x}{\partial u_1} & \frac{\partial x}{\partial u_2} \\ \frac{\partial y}{\partial u_1} & \frac{\partial y}{\partial u_2} \end{pmatrix}$$

$$= \det\begin{pmatrix} -\cos(2\pi u_2)\,/\,u_1\sqrt{2\log(1/u_1)} & -2\pi\sin(2\pi u2)\sqrt{2\log(1/u_1)} \\ -\sin(2\pi u_2)\,/\,u_1\sqrt{2\log(1/u_1)} & 2\pi\cos(2\pi u2)\sqrt{2\log(1/u_1)} \end{pmatrix}$$

$$= \;-2\pi\,/\,u_1\,.$$

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

$$f_{X,Y}(x,y) \;=\; 1/|J(h^{-1}(x,y))| \;=\; 1/\big|-2\pi\,/\,e^{-(x^2+y^2)/2}\big| \;=\; e^{-(x^2+y^2)/2}/2\pi$$

$$= \;\Big(\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\Big)\Big(\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\Big)\,,$$

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

- Another approach: "INVERSE CDF METHOD":

  - Suppose <u>want</u> $\mathbf{P}(X \le x) = F(x)$. ("CDF")

  - For $0 < t < 1$, set $F^{-1}(t) \;=\; \min\{x\,;\; F(x) \ge t\}$. ("inverse CDF")

  - Then set $X = F^{-1}(U_1)$.

  - Then $X \le x$ if and only if $U_1 \le F(x)$.

  - So, $\mathbf{P}(X \le x) \;=\; \mathbf{P}\big(U_1 \le F(x)\big) \;=\; F(x)$.

- 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: $se = M^{-1/2}\; sd\big(5 \cdot 4 \cdot g(X_1,Y_1),\, \ldots,\, 5 \cdot 4 \cdot g(X_M,Y_M)\big)$.
  - With $M = 10^6$, get about $-4.11 \pm 0.01 \ldots$ (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 $0.767 \pm 0.0004 \ldots$ very accurate! (file "RMCint3")
  - (Check: Numerical integration [Mathematica] gives 0.767211.)

- 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 $X \sim$ Uniform$[0,1]$ and $Y \sim$ 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 $x_i \sim$ Uniform$[0,1]$ and $y_i \sim$ Exponential(5) (i.i.d.).
  - With $M = 10^6$, get about 0.767 ± 0.0016 ... larger standard error ... (file "RMCint4").
  - If replace 5 by 1/5, get about 0.767 ± 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 π, 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. ("Importance Sampling")
  - Can then do classical (iid) Monte Carlo integration, get standard errors etc.
  - Good if easier to sample from f than π, 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 $X \sim$ Exponential(3), so $I = 1/3$ (error = 0, no MC needed).

UNNORMALISED DENSITIES:

- Suppose now that $\pi(y) = c\, g(y)$, where we know g but <u>don't</u> know c or π. ("Unnormalised density", e.g. Bayesian posterior.)
  - Obviously, $c = \frac{1}{\int g(y)\, dy}$, but this might be hard to compute.

– Still, $I = \int h(x)\,\pi(x)\,dx = \int h(x)\,c\,g(x)\,dx = \dfrac{\int h(x)\,g(x)\,dx}{\int g(x)\,dx}$.

– If sample $\{x_i\} \sim f$ (i.i.d.), then $\int h(x)\,g(x)\,dx = \int \Big(h(x)\,g(x)\,/\,f(x)\Big) f(x)\,dx = \mathbf{E}[h(X)\,g(X)\,/\,f(X)]$ where $X \sim f$.

– So, $\int h(x)\,g(x)\,dx \approx \frac{1}{M}\sum_{i=1}^{M}\Big(h(x_i)\,g(x_i)\,/\,f(x_i)\Big)$.

– Similarly, $\int g(x)\,dx \approx \frac{1}{M}\sum_{i=1}^{M}\Big(g(x_i)\,/\,f(x_i)\Big)$.

– So, $I \approx \dfrac{\sum_{i=1}^{M}\Big(h(x_i)\,g(x_i)\,/\,f(x_i)\Big)}{\sum_{i=1}^{M}\Big(g(x_i)\,/\,f(x_i)\Big)}$. ("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 $U \sim$ Uniform$[0,1]$, then $X \equiv U^{1/6} \sim f$, since then for $0 < x < 1$, $\mathbf{P}(X \le x) = \mathbf{P}(U^{1/6} \le x) = \mathbf{P}(U \le x^6) = x^6$, so $f_X(x) = \frac{d}{dx}x^6 = f(x)$.]

  – Then $I \approx \dfrac{\sum_{i=1}^{M}\Big(h(x_i)\,g(x_i)\,/\,f(x_i)\Big)}{\sum_{i=1}^{M}\Big(g(x_i)\,/\,f(x_i)\Big)} = \dfrac{\sum_{i=1}^{M}\Big(\sin(x_i^4)\,\cos(x_i^5)\Big)}{\sum_{i=1}^{M}\Big(\sin(x_i^4)\,\cos(x_i^5)\,/\,x_i^2\Big)}$. (file "Rimp" ... get about 0.766 ...)

  – Or, let $f(y) = 4\,y^3\mathbf{1}_{0<y<1}$. [Then if $U \sim$ Uniform$[0,1]$, then $U^{1/4} \sim f$.]

  – Then $I \approx \dfrac{\sum_{i=1}^{M}\Big(h(x_i)\,g(x_i)\,/\,f(x_i)\Big)}{\sum_{i=1}^{M}\Big(g(x_i)\,/\,f(x_i)\Big)} = \dfrac{\sum_{i=1}^{M}\Big(\sin(x_i^4)\,\cos(x_i^5)\,x_i^2\Big)}{\sum_{i=1}^{M}\Big(\sin(x_i^4)\,\cos(x_i^5)\Big)}$. (file "Rimp")

——————— **END WEEK #2**———————

[Assign HW#1 (due Feb 8) and Project (due March 29).]

Summary of Previous Class:

∗ Transforming uniforms to binomial, exponential, normal, ...

—— Inverse CDF method.

∗ Monte Carlo integration

—— Convert integral to some expectation

—— Then use classical Monte Carlo

—— Better if integrand less variable

$*$ Unnormalised densities: $\pi(x) = c\, g(x)$

- What other methods to iid sample from π?

REJECTION SAMPLER:

- Assume $\pi(x) = c\, g(x)$, with π and c unknown, g known but hard to sample from.
- Want to sample $X \sim \pi$.
  - Then if $X_1, X_2, \ldots, X_M \sim \pi$ iid, then can estimate $\mathbf{E}_\pi(h)$ by $\frac{1}{M}\sum_{i=1}^{M} h(X_i)$, etc.
- 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 π).
  - Otherwise, reject X and start over again.
- Conditional on accepting, we have [since $\mathbf{P}(U \leq g(X)/Kf(X) \,|\, X = x) = g(x)/Kf(x)$] that

$$\mathbf{P}\Big(X \leq y \,\Big|\, U \leq \frac{g(X)}{Kf(X)}\Big) \;=\; \frac{\mathbf{P}\Big(X \leq y,\;\; U \leq \frac{g(X)}{Kf(X)}\Big)}{\mathbf{P}\Big(U \leq \frac{g(X)}{Kf(X)}\Big)}$$

$$= \frac{\int_{-\infty}^{y} f(x)\, \frac{g(x)}{Kf(x)}\, dx}{\int_{-\infty}^{\infty} f(x)\, \frac{g(x)}{Kf(x)}\, dx} \;=\; \frac{\int_{-\infty}^{y} g(x)\, dx}{\int_{-\infty}^{\infty} g(x)\, dx} \;=\; \int_{-\infty}^{y} \pi(x)\, dx\,.$$

  - 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 $|x| \le 2$, $Kf(x) = 8\,\frac{1}{2}\,\exp(-|x|) \ge 8\,\frac{1}{2}\,\exp(-2) \ge (2\pi)^{-1/2} \ge \pi(x) = g(x)$.
  - For $|x| \ge 2$, $Kf(x) = 8\,\frac{1}{2}\,\exp(-|x|) \ge 8\,\frac{1}{2}\,\exp(-x^2/2) \ge (2\pi)^{-1/2}\exp(-x^2/2) = \pi(x) = g(x)$.
- 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")
- For Rejection Sampler, $P(\text{accept}) = \mathbf{E}[P(\text{accept}|X)] = \mathbf{E}[\frac{g(X)}{Kf(X)}] = \int \frac{g(x)}{Kf(x)}\,f(x)\,dx = \frac{1}{K}\int g(x)\,dx = \frac{1}{cK}$. (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 APPROACH: (related: "slice sampler")
  - Suppose $\pi(x) = c\,g(x)$, and (X, Y) chosen uniformly under the graph of g.
  - i.e., $(X, Y) \sim \text{Uniform}\{(x, y) \in \mathbf{R}^2 : 0 \le y \le g(x)\}$.
  - Then $X \sim \pi$, i.e. we have sampled from π.
  - 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 π, can estimate $\mathbf{E}_\pi(h)$ etc.
- Auxiliary Variable rejection sampler:
  - If support of g contained in $[L, R]$, and $|g(x)| \le K$, 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 $(X, Y) \sim \text{Uniform}\{(x, y) : 0 \le y \le g(x)\}$, 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!

MARKOV CHAIN MONTE CARLO (MCMC):

- Suppose have complicated, high-dimensional density $\pi = c\, g$.
- Want samples $X_1, X_2, \ldots \sim \pi$. (Then can do Monte Carlo.)
- Define a Markov chain (random process) $X_0, X_1, X_2, \ldots$, 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 $n = 1, 2, 3, \ldots, M$.
  - (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 π, 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" ...

——————— END WEEK #3———————

[Reminders: HW#1 due Feb 8, Project due March 29. Office hours?]

[HW#1, Q3: "f" should be "g" (of course!).]

[HW#1, Q2: "different" algorithms, not just different function choices.]

Summary of Previous Class:

∗ Rejection sampler

∗ Auxiliary variable approach

∗ MCMC

—— Metropolis algorithm

—— burn-in, starting distribution, standard error, ...

- Review of Java Applet example ("rwm.html").
- EXAMPLE: $\pi(x_1, x_2) = C\,|\cos(\sqrt{x_1\,x_2})|\; I(0 \le x_1 \le 5,\, 0 \le x_2 \le 4)$.
  - 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 $x_1 x_2 \approx (\pi/2)^2 \doteq 2.5$ ...
- OPTIMAL SCALING:
  - Can change proposal distribution to $Y_n \sim MVN(X_n,\, \sigma^2 I)$ for any $\sigma > 0$.
  - Which is best?
  - If σ too small, then usually accept, but chain won't move much.
  - If σ too large, then will usually reject proposals, so chain still won't move much.
  - Optimal: need σ "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 high-dimensional 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?)
- Basic fact: if a Markov chain is "irreducible" and "aperiodic", with "stationarity distribution" π, then $\mathcal{L}(X_n) \to \pi$ as $n \to \infty$.

- Let's figure out what this all means ...

- BEGIN WITH DISCRETE CASE, FROM JAVA APPLET EXAMPLE (rwm.html):
  - Here proposal is $q(x, x+1) = q(x, x-1) = 1/2$.
  - Acceptance probability is $\min(1, \frac{\pi(y)}{\pi(x)})$.
  - State space is $\mathcal{X} \equiv \{1, 2, 3, 4, 5, 6\}$.
  - So, for $i, j \in \mathcal{X}$ with $|j - i| = 1$,

$$P(i,j) \equiv P(i, \{j\}) = (1/2) \min(1, \frac{\pi(j)}{\pi(i)}) = \min(1/2, \frac{\pi(j)}{2\pi(i)}) .$$

  - So, chain is "reversible": for all $i, j \in \mathcal{X}$,

$$\pi(i) P(i,j) = \min(\pi(i)/2, \pi(j)/2) = \pi(j) P(j,i) . \qquad \text{(by symmetry)}$$

  - Then, if $X_0 \sim \pi$, then as for X_1,

$$\mathbf{P}(X_1 = j) = \sum_{i \in \mathcal{X}} \mathbf{P}(X_0 = i) P(i,j) = \sum_{i \in \mathcal{X}} \pi(i) P(i,j) = \sum_{i \in \mathcal{X}} \pi(j) P(j,i)$$

$$= \pi(j) \sum_{i \in \mathcal{X}} P(j,i) = \pi(j) .$$

    So, the Markov chain "preserves" π, i.e. π is a <u>stationary distribution</u>.
  - Also, it's irreducible (you can eventually get from anywhere to anywhere else).
  - And, it's aperiodic (no forced cycles).
  - So, as $n \to \infty$, $\mathcal{L}(X_n) \to \pi$. (file "rwm.html")

- SO WHAT ABOUT THE MORE GENERAL, CONTINUOUS CASE?

- Some notation:
  - 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, $q(x, y) = (2\pi\sigma)^{-d/2} \exp\big(-\sum_{i=1}^{d} (y_i - x_i)^2 / 2\sigma^2\big)$.) <u>Symmetric</u>: $q(x, y) = q(y, x)$.

– Let $\alpha(x, y)$ be probability of accepting a proposed move from x to y, i.e.

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

– Let $P(x, S) = \mathbf{P}(X_1 \in S \,|\, X_0 = x)$ be the transition probabilities.

- 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 <u>all</u> $x, y \in \mathcal{X}$. ("<u>reversible</u>")

– Shorthand: $P(x, dy)\, \Pi(dx) = P(y, dx)\, \Pi(dy)$.

- How does "reversible" help?

- Well, suppose $X_0 \sim \Pi$, i.e. we "start in stationarity". Then

$$\mathbf{P}(X_1 \in S) \;=\; \int_{x \in \mathcal{X}} \mathbf{P}(X_1 \in S \,|\, X_0 = x)\, \pi(x)\, dx \;=\; \int_{x \in \mathcal{X}} \int_{y \in S} P(x, dy)\, \pi(x)\, dx$$

$$= \int_{x \in \mathcal{X}} \int_{y \in S} P(y, dx)\, \pi(y)\, dy \;=\; \int_{y \in S} \pi(y)\, dy \;\equiv\; \Pi(S)\,,$$

so also $X_1 \sim \pi$. So, chain "preserves" π, i.e. π is <u>stationary</u> distribution.

- Also <u>irreducible</u>, 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. $\mathbf{P}(X_n \in S \,|\, X_0 = 0) > 0$. (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".)

- Aperiodic? Convergence Theorem? Coming up!

——————— **END WEEK #4**———————

[Reminders: HW#1 due Feb 8, Project due March 29. Office hours?]

[HW#1, Q3: just identify a, b, c, d. Also, a reminder that "f" $\to$ "g".]

[MCMC seminar: this Thurs (Feb 4), 3:30pm, SS1085.]

Summary of Previous Class:

$*$ Metropolis Algorithm (cont'd)

$*$ choice of σ^2

$*$ Theory, discrete case:

—— irreducible, aperiodic, reversible, π stationary

—— So, $\mathcal{L}(X_n) \to \pi$.

$*$ Theory, general case:

—— irreducible: if $\Pi(S) > 0$, then $P^n(x, S) > 0$ for some n (here $n = 1$)

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

—— So, π is stationary distribution

- Also aperiodic, i.e. there do not exist disjoint subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_j$ 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 $P(x, \{x\}) > 0$, e.g. 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 π 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 stationarity probability density π, then for π-a.e. initial value $X_0 = x$,
  (a) if $\mathbf{E}_\pi(|h|) < \infty$, then $\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^n h(X_i) = \mathbf{E}_\pi(h) \equiv \int h(x)\, \pi(x)\, dx$; and
  (b) if chain aperiodic, then also $\lim_{n\to\infty} \mathbf{P}(X_n \in S) = \int_S \pi(x)\, dx$ for all $S \subseteq \mathcal{X}$.
- Note: key facts about $q(x, y)$ are symmetry, and irreducibility.
  - So, could replace $Y_n \sim N(0, 1)$ by 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.

- 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!
  - π 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 π – good.
- EXAMPLE #2: Same as #1, except now $\pi(x) = 2^{-|x|-1}$ for $x \neq 0$, with $\pi(0) = 0$.
  - Still reversible, π 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, π 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 $\pi(x) = C\, e^{-x^6}$, and proposals $Y_n \sim \text{Uniform}[X_{n-1} - 1, X_{n-1} + 1]$.
  - Reversible? Yes since $q(x,y)$ still symmetric.
  - π 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 $P(x, \mathcal{X}_1) > 0$. (Or, even simpler: since $P(x, \{x\}) > 0$ for all $x \in \mathcal{X}$.)
  - So, by theorem, probabilities and expectations converge to those of π – good.
- EXAMPLE #5: Same as #4, except now $\pi(x) = C_1\, e^{-x^6}(\mathbf{1}_{x<2} + \mathbf{1}_{x>4})$.
  - 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 <u>not</u> 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 <u>can</u> 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!
- QUESTION: Why does Theorem say "π-a.e." $X_0 = x$?
- Example: $\mathcal{X} = \{1, 2, 3, \ldots\}$, 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 π-a.e. X_0, have $\lim_{n \to \infty} \mathbf{P}(X_n \in S) = \Pi(S)$, i.e. $\lim_{n \to \infty} \mathbf{P}(X_n = 1) = 1$.
  - 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 $\lim_{n \to \infty} \mathbf{P}(X_n = 1) \neq 1$.
  - Convergence holds if $X_0 = 1$, which is π-a.e. since $\Pi(1) = 1$, but not from $X_0 = x \geq 2$.
- So, convergence subtle. But <u>usually</u> holds from any $x \in \mathcal{X}$. ("Harris recurrent")

BAYESIAN STATISTICS:

- Have unknown parameter(s) θ, and a statistical model (likelihood function) for how the distribution of the data Y depends on θ: $\mathcal{L}(Y \,|\, \theta)$.
- Have a prior distribution, representing our "initial" (subjective?) probabilities for θ: $\mathcal{L}(\theta)$.
- Combining these gives a full joint distribution for θ and Y, i.e. $\mathcal{L}(\theta, Y)$.
- Then posterior distribution of θ, $\pi(\theta)$, is then the conditional distribution of θ, 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

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

——————— **END WEEK #5** ———————

[Collect HW#1 (at 2:10 sharp). Reminder: project due March 29.]

[Reminder: no class next week (Reading Week); return on Feb 22.]

Summary of Previous Class:
$*$ THM: if MC irreducible, aperiodic, and π stationary, then from π-a.e. starting point $X_0 = x$, the probabilities and averages converge to those of π.
—— Many examples
$*$ Bayesian Statistics (intro)

- Bayesian Statistics Example: VARIANCE COMPONENTS MODEL (a.k.a. "random effects model"):
  - Suppose some population has overall mean μ (unknown).
  - Population consists of K groups.
  - Observe $Y_{i1}, \ldots, Y_{iJ_i}$ from group i, for $1 \le i \le K$.
  - Assume $Y_{ij} \sim N(\theta_i, W)$ (cond. ind.), where θ_i and W unknown.

- Assume the different θ_i are "linked" by $\theta_i \sim N(\mu, V)$ (cond. ind.), with μ and V also unknown.
- Want to estimate some or all of $V, W, \mu, \theta_1, \ldots, \theta_K$.
- Bayesian approach: use prior distributions, e.g. ("conjugate"):

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

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 $x > 0$.

- Many applications, e.g.:
  - Predicting success at law school (D. Rubin, JASA 1980), $K = 82$ schools.
  - Melanoma recurrence (http://www.mssanz.org.au/modsim07/papers/52_s24/Analysing_Clinicals24_Bartolucci_.pdf), $K = 19$ patient catagories.
  - Comparing baseball home-run hitters (J. Albert, The American Statistician 1992), $K = 12$ players.
  - Analysing fabric dyes (Davies 1967; Box/Tiao 1973; Gelfand/Smith JASA 1990), $K = 6$ batches of dyestuff.
- Related to all of the above is our statistical model:

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

where

$$\theta_i \sim N(\mu, V); \qquad (1 \le i \le K),$$

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

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

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

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

$$= C_1 \left(e^{-b_1/V} V^{-a_1-1}\right)\left(e^{-b_2/W} W^{-a_2-1}\right)\left(e^{-(\mu-a_3)^2/2b_3}\right) \times$$
$$\times \left(\prod_{i=1}^{K} V^{-1/2} e^{-(\theta_i-\mu)^2/2V}\right)\left(\prod_{i=1}^{K}\prod_{j=1}^{J_i} W^{-1/2} e^{-(Y_{ij}-\theta_i)^2/2W}\right)$$

$$= C_2 e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-\frac{1}{2}\sum_{i=1}^{K} J_i} \times$$
$$\times \exp\left[-\sum_{i=1}^{K}(\theta_i-\mu)^2/2V - \sum_{i=1}^{K}\sum_{j=1}^{J_i}(Y_{ij}-\theta_i)^2/2W\right].$$

- Then

$$\pi(V, W, \mu, \theta_1, \ldots, \theta_K)$$
$$= C_3 \left(e^{-b_1/V} V^{-a_1-1}\right)\left(e^{-b_2/W} W^{-a_2-1}\right)\left(e^{-(\mu-a_3)^2/2b_3}\right) \times$$
$$\times \left(\prod_{i=1}^{K} V^{-1/2} e^{-(\theta_i-\mu)^2/2V}\right)\left(\prod_{i=1}^{K}\prod_{j=1}^{J_i} W^{-1/2} e^{-(Y_{ij}-\theta_i)^2/2W}\right)$$

- After a bit of simplifying,

$$\pi(V, W, \mu, \theta_1, \ldots, \theta_K)$$
$$= C e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-\frac{1}{2}\sum_{i=1}^{K} J_i} \times$$
$$\times \exp\left[-\sum_{i=1}^{K}(\theta_i-\mu)^2/2V - \sum_{i=1}^{K}\sum_{j=1}^{J_i}(Y_{ij}-\theta_i)^2/2W\right].$$

- Dimension: $d = K + 3$, e.g. $K = 19$, $d = 22$.

- How to compute/estimate, say, $\mathbf{E}_\pi(W/V)$?

  - 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!

- Try it out, with "dyestuffs" data ... (file "Rvarcomp")

– Sensitivity to choice of e.g. b_1?

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 (to make the Markov chain be reversible).
- If instead $\alpha(x, y) = \min\left[1,\ \frac{\pi(y)\,q(y,x)}{\pi(x)\,q(x,y)}\right]$, then

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

  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 $q(x, y) >> q(y, x)$, then Metropolis chain would spend too much time at y and not enough at x, so need to accept fewer moves $x \to y$.)

- EXAMPLE: again $\pi(x_1, x_2) \;=\; C\,|\cos(\sqrt{x_1\,x_2}\,)|\ I(0 \le x_1 \le 5, 0 \le x_2 \le 4)$, 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 ≈ 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 π, then $A_n \equiv 1$, 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?

- LANGEVIN ALGORITHM:
  - $Y_n \sim MVN\big(X_{n-1} + \frac{1}{2}\,\sigma^2\,\nabla \log \pi(X_{n-1}),\ \sigma^2 I\big)$.
  - Special case of Metropolis-Hastings algorithm.
  - Intuition: tries to move in direction where π increasing.
  - Based on discrete approximation to Langevin diffusion.
  - Usually more efficient, but requires knowledge and computation of $\nabla \log \pi$. (Hard.)

MCMC CONVERGENCE RATES:

- $\{X_n\}$: Markov chain on $\mathcal{X}$, with stationary distribution $\Pi(\cdot)$.
- Let $P^n(x, S) = \mathbf{P}[X_n \in S \,|\, X_0 = x]$.
  - 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 $\lim_{n\to\infty} D(x,n) = 0$, for Π-a.e. $x \in \mathcal{X}$.
  - Sometimes say chain "converges" in n iterations if $D(x,n) < 0.01 \ldots$
- DEFN: chain is geometrically ergodic if there is $\rho < 1$, and $M : \mathcal{X} \to [0,\infty]$ which is Π-a.e. finite, such that $D(x,n) \leq M(x)\,\rho^n$ for all $x \in \mathcal{X}$ and $n \in \mathbf{N}$.

——————— **END WEEK #6**———————

[Return HW#1: generally very good, but don't forget to consider accuracy of estimates – especially when estimating the same quantity by different methods! And, MCMC standard error is more complicated (discuss today!). When in doubt, run it again! Also, with RNG, use fresh values each time, and don't re-seed!]

Summary of Previous Class:

$*$ Variance Component Model

—— Estimate of W/V ...

$*$ Metropolis-Hastings algorithm

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

—— Independence sampler: $q(x,y) = q(y)$.

—— Langevin algorithm

$*$ MCMC Convergence Rates

—— $D(x,n) = \sup_{S\subseteq\mathcal{X}} |P^n(x,S) - \Pi(S)|$.

—— ergodic, geometrically ergodic, convergence in n iterations

- previous theorem: chain ergodic if irreducible and aperiodic and $\Pi(\cdot)$ stationary.
  - Hence: independence sampler is ergodic provided $q(x) > 0$ whenever $\pi(x) > 0$.
- 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 π-a.e. $x \in \mathcal{X}$, in which case $D(x,n) \leq (1-\delta)^n$ for π-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 <u>not</u> 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, "Quantitative Non-Geometric Convergence Bounds for Independence Samplers", MCAP, to appear.]

- 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 π 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: π and q continuous and positive; q has finite first moment; and π non-increasing in the tails, with (in higher dims) bounded Gaussian curvature ...)

  [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: 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 $v \equiv \mathrm{Var}\left(\frac{1}{M-B}\sum_{i=B+1}^{M} h(X_i)\right)$?
  - Let $\overline{h}(x) = h(x) - \mathbf{E}_\pi(h)$, so $\mathbf{E}_\pi(\overline{h}) = 0$.
  - And, assume B large enough that $X_i \approx \pi$ for $i > B$.
  - Then, for large $M - B$,

$$v \;\approx\; \mathbf{E}_\pi\Big[\Big(\Big(\frac{1}{M-B}\sum_{i=B+1}^{M} h(X_i)\Big) - \mathbf{E}_\pi(h)\Big)^2\Big] \;=\; \mathbf{E}_\pi\Big[\Big(\frac{1}{M-B}\sum_{i=B+1}^{M} \overline{h}(X_i)\Big)^2\Big]$$

$$= \frac{1}{(M-B)^2}\Big[(M-B)\mathbf{E}_\pi(\overline{h}(X_i)^2) + 2(M-B-1)\mathbf{E}_\pi(\overline{h}(X_i)\overline{h}(X_{i+1})) + 2(M-B-2)\mathbf{E}_\pi(\overline{h}(X_i)\overline{h}(X_{i+2})) + \ldots\Big]$$

$$\approx \frac{1}{M-B}\Big(\mathbf{E}_\pi(\overline{h}^2) + 2\,\mathbf{E}_\pi(\overline{h}(X_i)\overline{h}(X_{i+1})) + 2\,\mathbf{E}_\pi(\overline{h}(X_i)\overline{h}(X_{i+2})) + \ldots\Big)$$

$$= \frac{1}{M-B}\,\mathbf{E}_\pi(\overline{h}^2)\Big(1 + 2\,\mathrm{Corr}_\pi(\overline{h}(X_i)\overline{h}(X_{i+1})) + 2\,\mathrm{Corr}_\pi(\overline{h}(X_i)\overline{h}(X_{i+2})) + \ldots\Big)$$

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

    where

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

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

  - Then can estimate both iid variance, and varfact, from the sample run, as usual.
  - e.g. "acf" commands in files Rnorm, Rheavy, Rheavy2.

– 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 can also write your own – better.)

– Usually varfact $\gg 1$; try to get "better" chains so varfact smaller.

– Sometimes even try to design chain to get varfact < 1 ("antithetic").

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, if have central limit theorem (CLT), then for large $M - B$, $e \approx N(u, v)$.
  - 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 $(e - 1.96\sqrt{v},\ e + 1.96\sqrt{v})$.
  - Strictly speaking, should use "t" distribution, not normal distribution ... but if $M - B$ large that doesn't really matter (ignore it for now).
- 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 δ: Roberts and Rosenthal, "Geometric ergodicity and hybrid Markov chains", ECP 1997.)

——————— END WEEK #7———————

[Hand out HW #2.]

Summary of Previous Class:

∗ MCMC convergence rates / bounds

—— finite state space

—— independence sampler

—— RWM

∗ MCMC standard error

—— varfact

∗ MCMC CLT's, conf. int.

- e.g. 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")
- e.g. previous "Rind" example: Π = Exponential(1), $h(y) = y$, $\mathbf{E}_\pi(h) = 1$, Q = Exponential(5).
  - Not geometrically ergodic.
  - CLT known not to hold.
  - Confidence intervals often miss 1. (file "Rind2")
- e.g. previous "Rheavy2" example: Π = 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 Π, and might underestimate v.
  - Estimate of v is very important!
  - 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 <u>one</u> coordinate at a time, leaving all the other coordinates <u>fixed</u> (since changing all coordinates at once may be difficult).
  - e.g. proposal Y_n has $Y_{n,i} \sim N(X_{n-1,i}, \sigma^2)$, with $Y_{n,j} = X_{n-1,j}$ for $j \neq i$.
- Then accept/reject with usual Metropolis rule (symmetric case: "Metropolis-within-Gibbs") or Metropolis-Hastings rule (general case: "Metropolis-Hastings-within-Gibbs").
- Need to choose which coordinate to update each time ...
  - Could choose coordinates in sequence $1, 2, \ldots, d, 1, 2, \ldots$ ("systematic-scan").
  - Or, choose coordinate $\sim$ Uniform$\{1, 2, \ldots, d\}$ each time ("random-scan").
  - Note: one systematic-scan iteration corresponds to d random-scan ones ...
- 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 $h(x_1, x_2) = e^{x_1} + (x_2)^2$. (Recall: Mathematica gives $\mathbf{E}_\pi(h) \doteq 38.7044$.)
  - Works with systematic-scan (file "Rmwg") or random-scan (file "Rmwg2").
- GIBBS SAMPLER:
- Special case of Metropolis-Hastings-within-Gibbs.
- Proposal distribution for i^{th} coordinate is equal to the conditional distribution of that coordinate (according to π), conditional on the current values of all the other coordinates.
  - That is, $q_i(x, y) = C(x^{(-i)})\,\pi(y)$ whenever $x^{(-i)} = y^{(-i)}$, where $x^{(-i)}$ means all coordinates <u>except</u> the i^{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, <u>always</u> accept.
  - Can use either systematic or random scan.

- EXAMPLE: Variance Components Model (with $J_i \equiv J$):
  - Update of μ (say) should be from conditional density of μ, conditional on current values of all the other coordinates: $\mathcal{L}(\mu \,|\, V, W, \theta_1, \ldots, \theta_K, Y_{11}, \ldots, Y_{JK})$.
  - This conditional density is proportional to the full joint density, but with everything <u>except</u> μ treated as constant.
  - Full joint density is:

$$= \; Ce^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-JK/2} \; \times$$

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

  - So, conditional density of μ is

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

  - This equals (verify this!)

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

  - 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 $m = (a_3 V + b_3 \sum_{i=1}^{K} \theta_i) \,/\, (V + Kb_3)$.
  - <u>So</u>, in Gibbs Sampler, each time μ is updated, we sample it from $N(m, v)$ for this m and v (and always accept).
- <u>Similarly</u> (HW#2), conditional distribution for V is:

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

  - 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 (HW#2), cond. 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 θ_i $(1 \le i \le K)$, conditional distribution (HW#2) is:

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

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

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. $\frac{1}{2}\,N(0, 10^2) + \frac{1}{2}\,N(20, 10^2)$, then much easier to get between modes.

- So: define a sequence $\Pi_1, \Pi_2, \ldots, \Pi_m$ where $\Pi_1 = \Pi$ ("cold"), and Π_τ is flatter for larger τ ("hot").

- Then define Markov chain on $\mathcal{X} \times \{1, 2, \ldots, m\}$, with stationary distribution $\overline{\Pi}$ defined by $\overline{\Pi}(S \times \{\tau\}) = \frac{1}{m}\,\Pi_\tau(S)$.
  - (Can also use other weights besides $\frac{1}{m}$.)

- Define new Markov chain with both spatial moves (change x) and temperature moves (change τ).
  - e.g. perhaps chain alternates between:

(a) propose $x' \sim N(x, 1)$, accept with prob $\min\left(1,\ \frac{\overline{\pi}(x',\tau)}{\overline{\pi}(x,\tau)}\right) = \min\left(1,\ \frac{\pi_\tau(x')}{\pi_\tau(x)}\right)$.
(b) propose $\tau' = \tau \pm 1$ (prob $\frac{1}{2}$ each), accept with prob
$\min\left(1,\ \frac{\overline{\pi}(x,\tau')}{\overline{\pi}(x,\tau)}\right) = \min\left(1,\ \frac{\pi_{\tau'}(x)}{\pi_\tau(x)}\right)$.

- Chain should converge to $\overline{\Pi}$.
- 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 Π: terrible! (file "Rtempered" with dotempering = FALSE; note the small claimed standard error!)
  - Define $\Pi_\tau = \frac{1}{2}\, N(0, \tau^2) + \frac{1}{2}\, N(20, \tau^2)$, for $\tau = 1, 2, \ldots, 10$.
  - Mixing for Π_{10}: good! (file "Rtempered" started with temp = 10)
  - (Compare graphs of π_1 and π_{10}: plot commands at bottom of "Rtempered" ...)
  - So, use above "(a)–(b)" algorithm; converges <u>fairly</u> well to $\overline{\Pi}$. (file "Rtempered", with dotempering = TRUE)
  - So, conditional on $\tau = 1$, converges to Π. ("points" command at end of file "Rtempered")
  - So, average of those $h(x)$ with $\tau = 1$ gives good estimate of $\mathbf{E}_\pi(h)$.
- HOW TO FIND THE TEMPERED DENSITIES π_τ?
- 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 $\pi_\tau(x) = c_\tau \left(\pi(x)\right)^{1/\tau}$. (Sometimes write $\beta = 1/\tau$.)
  - Then $\Pi_1 = \Pi$, and π_τ <u>flatter</u> for larger τ – 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:

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

– This depends on the c_τ, which are usually unknown – bad.

- What to do?

——————— **END WEEK #8**———————

[Do course evals.]

[Reminders: HW#2 due Mar 22; project due Mar 29.]

Summary of Previous Class:

$*$ Examples of conf. int. for MCMC

—— Work pretty well for geom. erg. chains ...

$*$ Variable-at-a-time Metropolis

—— systematic or random scan

$*$ Gibbs sampler:

—— propose from full conditional dists.

—— example: Variance Components Model (HW#2)

$*$ Tempered MCMC:

—— sequence $\Pi_1 = \Pi, \Pi_2, \ldots, \Pi_m$ getting flatter

—— Define new chain with both x and τ moves

—— Then, only "count" samples where $\tau = 1$

$*$ Problem: how to define Π_τ.

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

—— Problem: c_τ unknown, and doesn't cancel!

- PARALLEL TEMPERING:
- (a.k.a. Metropolis-Coupled MCMC, or MCMCMC)
- Alternative to tempered MCMC.
- Instead, use state space $\mathcal{X}^m$, with m chains, i.e. one chain for <u>each</u> temperature.
- So, state at time n is $X_n = (X_{n1}, X_{n2}, \ldots, X_{nm})$, where $X_{n\tau}$ is "at" temperature τ.
- Stationary distribution is now $\overline{\Pi} = \Pi_1 \times \Pi_2 \times \ldots \times \Pi_m$, i.e. $\overline{\Pi}(X_1 \in S_1,\ X_2 \in S_2,\ \ldots, X_m \in S_m) = \Pi_1(S_1)\,\Pi_2(S_2)\,\ldots\,\Pi_m(S_m)$.

- Then, can update the chain at temperature τ (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 τ and τ' (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 $\pi_\tau(x) = c_\tau \left(\pi(x)\right)^{1/\tau}$, then acceptance probability is:

$$\min\left(1,\ \frac{c_\tau \pi(X_{n,\tau'})^{1/\tau}\, c_{\tau'} \pi(X_{n,\tau})^{1/\tau'}}{c_\tau \pi(X_{n,\tau})^{1/\tau}\, c_{\tau'} \pi(X_{n,\tau'})^{1/\tau'}}\right) \;=\; \min\left(1,\ \frac{\pi(X_{n,\tau'})^{1/\tau}\, \pi(X_{n,\tau})^{1/\tau'}}{\pi(X_{n,\tau})^{1/\tau}\, \pi(X_{n,\tau'})^{1/\tau'}}\right),$$

    so c_τ and $c_{\tau'}$ are not required.

- EXAMPLE: suppose again that $\Pi_\tau = \frac{1}{2}\, N(0, \tau^2) + \frac{1}{2}\, N(20, \tau^2)$, for $\tau = 1, 2, \ldots, 10$.

  - Can run parallel tempering ... works pretty well. (file "Rpara")

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 Σ.

- What is best choice of Σ??

  - Usually we take $\Sigma = \sigma^2\, I_d$ for some $\sigma > 0$, and then choose σ so acceptance rate not too small, not too large (e.g. 0.234).

  - But can we do better?

- Suppose for now that $\Pi = MVN(\mu_0,\, \Sigma_0)$ for some fixed μ_0 and Σ_0, in dim=5. Try RWM with various proposal distributions (file "Ropt"):

  - first version: $Y_n \sim MVN(X_{n-1},\, I_d)$. ($acc \approx 0.06$; $varfact \approx 220$)

  - second version: $Y_n \sim MVN(X_{n-1},\ 0.1\, I_d)$. ($acc \approx 0.234$; $varfact \approx 300$)

  - third version: $Y_n \sim MVN(X_{n-1},\, \Sigma_0)$. ($acc \approx 0.31$; $varfact \approx 15$)

  - fourth version: $Y_n \sim MVN(X_{n-1},\ 1.4\, \Sigma_0)$. ($acc \approx 0.234$; $varfact \approx 7$)

- Or in dim=20 (file "Ropt2"):

- $Y_n \sim MVN(X_{n-1}, 0.025\, I_d)$. ($acc \approx 0.234$; $varfact \approx 400$ or more)
- $Y_n \sim MVN(X_{n-1}, 0.283\, \Sigma_0)$. ($acc \approx 0.234$; $varfact \approx 50$)

- Conclusion: acceptance rates near 0.234 are better.
- But also, proposals shaped like the target are better.
  - This has been 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).
  - Is "approximately" true for most unimodal targets ...
- Problem: Σ_0 would usually be unknown; then what?
  - Can perhaps "adapt"!

ADAPTIVE MCMC:

- What if target covariance Σ_0 is unknown??
- Can estimate target covariance based on run so far, to get empirical covariance Σ_n.
- Then update proposal covariance "on the fly", by using proposal $Y_n \sim MVN(X_{n-1}, \Sigma_n)$ [or $Y_n \sim MVN(X_{n-1}, 1.4\Sigma_n)$, or $Y_n \sim MVN(X_{n-1}, ((2.38)^2/d)\Sigma_n)$].
  - Hope that for large n, $\Sigma_n \approx \Sigma_0$, so proposals "nearly" optimal.
  - (Usually also add ϵI_d to proposal covariance, to improve stability, e.g. $\epsilon = 0.05$.)
- Resulting "adaptive Metropolis (AM) algorithm" seems to work well in practice (e.g. figure "plotAMx200.png", dim=200).
  - But it takes many iterations before the adaption is helpful.
- Try R version, for the same MVN example as in Ropt (file "Radapt"):
  - Need much longer burn-in, e.g. $B = 20,000$, for adaption to work.
  - Get varfact of last 4000 iterations of about 18 ... "competitive" with Ropt optimal ...

  - The longer the run, the more benefit from adaptation.
  - Can also compute "slow-down factor", $s_n \equiv d \left(\sum_{i=1}^{d} \lambda_{in}^{-2} / (\sum_{i=1}^{d} \lambda_{in}^{-1})^2 \right)$, where $\{\lambda_{in}\}$ eigenvals of $\Sigma_n^{1/2} \Sigma_0^{-1/2}$. Starts large, should converge to 1. [Motivation: if $\Sigma_n = \Sigma_0$, then $\lambda_{in} \equiv 1$, so $s_n = d(d/d^2) \equiv 1$.]

- BUT IS "ADAPTIVE MCMC" A VALID ALGORITHM??
- Not in general: see e.g. "adapt.html"
- Algorithm now non-Markovian, doesn't preserve stationarity at each step.
- However, still converges to Π 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).

———————— END WEEK #9————————

[This week's statistics research seminar is on MCMC: Thurs 3:30, SS1085.]

[Reminders: HW#2 due Mar 22; project due Mar 29.]

Summary of Previous Class:

∗ Parallel tempering

—— like Tempered MCMC, but constants cancel

∗ Optimal RWM proposals

—— acc rate 0.234 good

—— But <u>also</u> good if shape of proposal similar to shape of target

—— Problem: might not KNOW shape of target

∗ Adaptive MCMC

—— <u>Learn</u> shape/size/etc of target as you go.

—— After <u>many</u> iterations, becomes efficient MCMC – good.

—— But requires certain conditions (e.g. "Diminishing Adaptation") or else it might fail to converge – Java applet.

MONTE CARLO IN FINANCE:

- X_t = stock price at time t
- Assume that $X_0 = a > 0$, and $dX_t = bX_t dt + \sigma X_t dB_t$.
  - i.e., for small $h > 0$,

    $$(X_{t+h} - X_t \,|\, X_t) \;\approx\; bX_t(t+h-t) + \sigma X_t(B_{t+h} - B_t) \;\sim\; bX_t(t+h-t) + \sigma X_t\, N(0,h)\,,$$

    so

    $$(X_{t+h} \,|\, X_t) \;\sim\; N\left(X_t + bX_t h,\; \sigma^2 (X_t)^2 h\right)\,. \qquad (*)$$

- A "European call option" is the option to purchase one share of the stock at a fixed time $T > 0$ for a fixed price $q > 0$.
- Question: what is a fair <u>price</u> for this option?
  - At time T, its value is $\max(0,\; X_T - q)$.
  - So, at time 0, its value is $e^{-rT}\, \max(0,\; X_T - q)$, where r is the "risk-free interest rate".
  - But at time 0, X_T is unknown! So, what is fair price??
- FACT: the fair price is equal to $\mathbf{E}\big(e^{-rT} \max(0,\; X_T - q)\big)$, but only after replacing b by r.
  - (Proof: transform to risk-neutral martingale measure ...)
  - Intuition: if b very large, might as well just buy stock itself.
- If σ and r constant, then there's a <u>formula</u> ("Black-Scholes eqn") for this price, in terms of Φ = cdf of $N(0,1)$:

  $$a\;\Phi\left(\frac{1}{\sigma\sqrt{T}}\Big(\log(a/q) + T(r + \frac{1}{2}\sigma^2)\Big)\right) - \;qe^{-rT}\Phi\left(\frac{1}{\sigma\sqrt{T}}\Big(\log(a/q) + T(r - \frac{1}{2}\sigma^2)\Big)\right)$$

- But we can also estimate it through (iid) Monte Carlo!
  - Use $(*)$ above (for fixed small $h > 0$, e.g. $h = 0.05$) to generate samples from the difusion.

- Any <u>one</u> run is highly variable. (file "RBS", with $M = 1$)
- But <u>many</u> runs give good estimate. (file "RBS", with $M = 1000$)
- Note that it's iid replications, so varfact $\equiv 1$.

- An "Asian call option" is similar, but with X_T replaced by $\overline{X}_{k,t} \equiv \frac{1}{k} \sum_{i=1}^{k} X_{iT/k}$, for some fixed positive integer k (e.g., $k = 8$).
  - Above "FACT" still holds (again with X_T replaced by $\overline{X}_{k,t}$).
  - Now there is no simple formula ... but can still simulate! (file "RAO")

MONTE CARLO MAXIMISATION (OPTIMISATION):

- EXAMPLE #1: CODE BREAKING, e.g. "decipherit oliver". ["decipher.c"]
  - "substitution cipher".
- Data is the coded message text: $s_1 s_2 s_3 \ldots s_N$, where $s_i \in \mathcal{A} = \{A, B, C, \ldots, Z, \text{space}\}$.
- State space $\mathcal{X}$ is set of all bijections of $\mathcal{A}$, i.e. one-to-one onto mappings $f : \mathcal{A} \to \mathcal{A}$, 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 $f \in \mathcal{X}$, let $\pi(f) = \prod_{i=1}^{N-1} M\Big(f(s_i), f(s_{i+1})\Big)$.
  - (Or raise this all to a power, e.g. 0.25.)
- Idea: if $\pi(f)$ is larger, then f leads to pair frequencies which more closely match the reference text, so f is a "better" choice.
- Would like to find f which <u>maximises</u> $\pi(f)$.
- To do this, run a Metropolis algorithm for π:
  - Choose $a, b \in \mathcal{A} \setminus \{\text{space}\}$, uniformly at random.
  - Propose to replace f by g, where $g(a) = f(b)$, $g(b) = f(a)$, and $g(x) = f(x)$ for all $x \neq a, b$.

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

- Easily seen to be irreducible, aperiodic, reversible.
- So, converges (quickly!) to correct answer, breaking the code. (e.g. "decipheroutput")
- References: S. Conner (2003), "Simulation and solving substitution codes". P. Diaconis (2008), "The Markov Chain Monte Carlo Revolution".
- EXAMPLE #2: COMPUTER VISION, e.g. "faces" Java applet. ["faces.html"]
- Data is an image, given in terms of a grid of pixels (each on or off).
- Define the face location by a vector θ of various parameters (face center, eye width, nose height, etc.).
- Then define a score function $S(\theta)$ indicating how well the image agrees with having a face in the location corresponding to the parameters θ.
- Then run a "mixed" Monte Carlo search (sometimes updating by small RWM moves, sometimes starting fresh from a random vector) over the entire parameter space, searching for $\text{argmax}_\theta\, S(\theta)$, i.e. for the parameter values which maximise the score function.
  - Keep track of best θ so far – this allows for greater flexibility in trying different search moves without needing to preserve a stationary distribution.
  - Works pretty well, and fast! ("faces.html" Java applet)
  - For details, see Java applet source code, "faces.java" (or the related paper).

——————— END WEEK #10———————

Summary of Previous Class:

$*$ MC in finance:

—— $dX_t = bX_t dt + \sigma X_t dB_t$

—— $(X_{t+h} \,|\, X_t) \ \sim\ N\left(X_t + bX_t h,\ \sigma^2 (X_t)^2 h\right)$

$*$ European call option:

—— Replace b with r

—— Then price $= \mathbf{E}\left(e^{-rT} \max\left(0,\ X_T - q\right)\right)$.

—— Can compute (BS) or estimate (MC) this – good.

∗ Asian call option:

—— Replace X_T by $\overline{X}_{k,t} \equiv \frac{1}{k}\sum_{i=1}^{k} X_{iT/K}$.

—— Can still estimate by MC.

∗ Code breaking:

—— choose substitution cipher function f to maximise $\pi(f)$.

∗ Face identification:

—— choose face parameters θ to maximise $S(\theta)$.

- In both of these examples, wanted to MAXIMISE π rather than SAMPLE from π.
  - General method??
- SIMULATED ANNEALING:
- General method to find highest <u>mode</u> of π.
- Idea: mode of π is same as mode of flatter version π_τ, for any $\tau > 0$. (e.g. $\pi_\tau \equiv \pi^{1/\tau}$)
  - For large τ, MCMC explores a lot; good at beginning of search.
  - For small τ, MCMC narrows in on local mode; good at end of search.
- So, use tempered MCMC, but where $\tau = \tau_n \searrow 0$, so π_{τ_n} becomes more and more <u>concentrated</u> at mode as $n \to \infty$.
- Need to choose $\{\tau_n\}$, the "cooling schedule".
  - e.g. geometric ($\tau_n = \tau_0\, r^n$ for some $r < 1$).
  - or linear ($\tau_n = \tau_0 - dn$ for some $d > 0$, chosen so that $\tau_M = \tau_0 - dM \geq 0$).
  - or logarithmic ($\tau_n = c/\log(1+n)$). [Thm: if $c \geq \sup \pi$, then simulated annealing with $\tau_n = c/\log(1+n)$ will converge to global maximum as $n \to \infty$.]
  - or ...
- EXAMPLE: $\Pi_\tau = 0.3\, N(0, \tau^2) + 0.7\, N(20, \tau^2)$. (file "Rsimann")
  - Highest mode is at 20 (for any τ).

- If run usual Metropolis algorithm, it will either jump forever between modes (if τ large), or get stuck in one mode or the other with equal probability (if τ small) – bad.
- But if $\tau_n \searrow 0$ slowly, then can usually find the highest mode (20) – good.
- Try both exponential and linear (better?) cooling ... (file "Rsimann")

APPLICATION – SUGAR CANE INFECTIONS:

- Study: a field of sugar cane plants in Guadeloupe, subject to infection.
  - Field of size about 25×100 meters.
  - Grid of size about 17×103 plants.
  - Total number of plants = 1,742.
- Plans were all healthy at time 0.
- Plants were checked for infection after 6, 10, 14, 19, 23, and 30 weeks.
- So, data includes $1742 * 6 = 10,452$ binary variables (0 or 1). (Also includes location coordinates of each plant.)
  - file "canedisplay": time-lag display, 30 weeks $\leftrightarrow$ ten seconds.
  - Will later focus on lower-left 10×10 grid (file "canedisplaysm").
- TASK: develop Bayesian model for infections, and estimate the parameter values.
- Notation: $\mathcal{X}$ = set of all canes, τ_x = time of infection of cane x (with $\tau_x = \infty$ if cane x uninfected throughout study).
- Consider discrete times $k = 0, 1, 2, \ldots, K$. (Could also try continuous times ...)
  - For now take $K = 30$ and measure time in weeks. (Could also take $K = 30 \times 7$ and measure time in days ... more accurate model, but harder to compute.)
- Assume $\mathbf{P}(\tau_x = k \,|\, \tau_x \geq k) = p_{x,k} \equiv 1 - e^{-\lambda_{x,k}}$.
- Take $\lambda_{x,k} = \exp\left(\mu + \sum_{z:\tau_x \leq k-1} e^{-\theta\, d(x,z)}\right)$, where $d(x,z)$ = physical distance between

canes x and z. (Perhaps better is e.g. $\lambda_{x,k} = e^{\mu} + e^{\alpha} \sum_{z:\tau_x \le k-1} e^{-\theta\, d(x,z)}$??)

  - Here μ = spontaneous infection rate, and θ = transmission rate.
  - (Could also add A_k to μ, with e.g. $A_{k+1} = rA_k + B_k$, to make spontaneous infection rate depend on conditions of weather / insects / etc at time k ...)

- Prior distributions: $\theta \sim \text{Exp}(1/10)$, $\mu \sim N(0, 10^2)$.
- Now relevant "data" is L_x = the last time cane x was observed uninfected, and U_x = the first time cane x was observed infected.
  - Restriction: $L_x < \tau_x \le U_x$.
- This gives a joint distribution

$$\pi(\theta, \mu, \{\tau_x\}_{x\in\mathcal{X}}) \;\propto\; e^{-\theta/10} e^{-\mu^2/200} \prod_{k=1}^{K} \left(\prod_{x:\tau_x=k} (1 - e^{-\lambda_{x,k}}) \prod_{x:\tau_x>k} e^{-\lambda_{x,k}} \right) \prod_{x\in\mathcal{X}} \mathbf{1}(L_x < \tau_x \le U_x)\,.$$

- On log scale, if $L_x \ge \tau_x$ or $\tau_x > U_x$ for some $x \in \mathcal{X}$, then $\log \pi \equiv -\infty$, else

$$\log \pi(\theta, \mu, \{\tau_x\}_{x\in\mathcal{X}}) = C - \theta/10 - \mu^2/200 + \sum_{k=1}^{K} \left(\Big(\sum_{x:\tau_x=k} \log(1 - e^{-\lambda_{x,k}}) \Big) - \sum_{x:\tau_x>k} \lambda_{x,k} \right)$$

$$= C - \theta/10 - \mu^2/200 + \sum_{x:\tau_x\le K} \left(\log(1 - e^{-\lambda_{x,\tau_x}}) - \sum_{k=1}^{\min(K,\ \tau_x - 1)} \lambda_{x,k} \right).$$

- How to sample from π??
- Try Metropolis-within-Gibbs ...
- Propose $\theta' = \theta + N(0, \sigma_1^2)$; then accept (i.e., replace θ by θ') iff $\log(U) < \log \pi\,(\theta', \mu, \{\tau\}_{x\in\mathcal{X}}) - \log \pi\,(\theta, \mu, \{\tau\}_{x\in\mathcal{X}})$.
- Propose $\mu' = \mu + N(0, \sigma_2^2)$; accept iff $\log(U) < \log \pi\,(\theta, \mu', \{\tau\}_{x\in\mathcal{X}}) - \log \pi\,(\theta, \mu, \{\tau\}_{x\in\mathcal{X}})$.
- For each $x \in \mathcal{X}$:
  - Propose $\tau_x' = \tau_x \pm 1$ (prob 1/2 each), with $\tau_y' = \tau_y$ for $y \ne x$; accept iff $\log(U) < \log \pi\,(\theta, \mu, \{\tau'\}_{x\in\mathcal{X}}) - \log \pi\,(\theta, \mu, \{\tau\}_{x\in\mathcal{X}})$.

- Repeat M times.
- Works (file "canesR"), but very slow (even just on 10×10 grid).
- Faster in C (file "canes.c"); ran for $M = 110,000$ iterations with burn-in $B = 10,000$.
  - (Don't need to update τ_x at sites where infection never observed.)
  - ARtheta=0.056273, ARmu=0.782545, ARtau=0.712048.
- Estimates of parameters: $\hat{\theta} = 1.092876$, $\hat{\mu} = -4.916916$.
- Estimated τ values in file "tauvals".
  - Can see time-lag display of τ's (file "canetaudisplay").
  - (Provides a "filled-in version" of "canedisplaysm".)
  - Can also investigate histogram of τ's (file "canetaudisplay").
  - Shows SOME "spread" from observed values, but not too much ...

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: autoregressive process: suppose $Y_n = a_1 Y_{n-1} + a_2 Y_{n-2} + \ldots + a_k Y_{n-k}$, but we don't know what k should be.
- EXAMPLE: suppose $\{y_j\}_{j=1}^J$ are known data which come from a mixture distribution $\frac{1}{k}\big(N(a_1, 1) + N(a_2, 1) + \ldots + N(a_k, 1)\big)$.
- Want to estimate unknown $k, a_1, \ldots, a_k$.
  - Here the number of parameters is also unknown, i.e. the dimension is unknown and variable, which makes MCMC more challenging!

——————— END WEEK #11———————

[HW#2 very good. Don't forget repeated runs! Langevin poor. Priors have "meaning".]

[Collect projects. Return HW#2. Refs used? Discuss test.]

Summary of Previous Class:

$*$ Simulated annealing:

—— Run MCMC with $\tau_n \searrow 0$.

—— Hopefully converge to mode.

$*$ Big application: sugar cane infections

—— Complicated model.

—— posterior dim = 1744 (or 122).

—— Try to sample using Metropolis-within-Gibbs.

$*$ Transdimensional MCMC

—— Even dimension of state space is changing.

- EXAMPLE: suppose $\{y_j\}_{j=1}^J$ are known data which come from a mixture distribution $\frac{1}{k}\big(N(a_1,1) + N(a_2,1) + \ldots + N(a_k,1)\big)$.
- Want to estimate unknown $k, a_1, \ldots, a_k$.
- 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 $a|k \sim MVN(0, I_k)$ (say).
- Let $\lambda = \lambda_0 \times (\lambda_1 + \lambda_2 + \lambda_3 + \ldots)$, where λ_0 is counting measure (for k), and for $i \geq 1$, λ_i is Lebesgue measure on $\mathbf{R}^i$.
- Then posterior density (with respect to λ) is:

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

- So, on a log scale,

$$\log\pi(k,a) \ = \ \log C + \log\frac{e^{-2}2^{k-1}}{(k-1)!} - \frac{k}{2}\log(2\pi) - \frac{1}{2}\sum_{i=1}^k a_i^2 - \frac{J}{2}\log(2\pi) +$$

$$\sum_{j=1}^J \log\Big(\sum_{i=1}^k\frac{1}{k}\exp\Big(-\frac{1}{2}(y_j - a_i)^2\Big)\Big).$$

(Can ignore $\log C$ and $\frac{J}{2}\log(2\pi)$, 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 $a' = (a_1, \ldots, a_k, Z)$ where $Z \sim N(0,1)$ (elongate).
  - If $k' = k-1$, then $a' = (a_1, \ldots, a_{k-1})$ (truncate).
- Then accept with usual probability, $\min\Big(1,\ \frac{\pi(k',a')\, q\big((k',a'),(k,a)\big)}{\pi(k,a)\, q\big((k,a),(k',a')\big)}\Big)$.
  - Here if $k' = k+1$, then $q\big((k',a'),(k,a)\big) = \frac{1}{2}$, while $q\big((k,a),(k',a')\big) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-(a'_{k'})^2/2}$.
  - Or, if $k' = k-1$, then $q\big((k,a),(k',a')\big) = \frac{1}{2}$, while $q\big((k',a'),(k,a)\big) = \frac{1}{2}\frac{1}{\sqrt{2\pi}}e^{-(a_k)^2/2}$.
- Seems to work okay; final k usually between 5 and 9 ... (file "Rtrans")
  - (Could also use modification where any coordinate can be added/removed, not just the last one ...)
- Alternative method for "correspondingly change a" step:
  - If $k' = k+1$, then $a' = (a_1, \ldots, a_{k-1}, a_k - Z, a_k + Z)$ where $Z \sim N(0,1)$ ("split").
  - If $k' = k-1$, then $a' = (a_1, \ldots, a_{k-2}, \frac{1}{2}(a_{k-1} + a_k))$ ("merge").
  - What about the densities $q\big((k',a'),(k,a)\big)$?
  - Well, if $k' = k+1$, then $q\big((k',a'),(k,a)\big) = \frac{1}{2}$, while roughly speaking,

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

  - One subtle additional point: The map $(a, Z) \mapsto a' = (a_1, \ldots, a_{k-1}, a_k - Z, a_k + Z)$

has "Jacobian" term:

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

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

– So by Change-of-Variable Thm, actually

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

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

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

– 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).

———————— **END WEEK #12**————————

- SUMMARY: Monte Carlo can be used for nearly everything!
- Good luck on your exams, etc., and have a nice summer.