STA3431 (Monte Carlo Methods) Lecture Notes, Fall 2020

by Jeffrey S. Rosenthal, University of Toronto

(Last updated: November 30, 2020)

Note: I will update these notes regularly (on the course web page). 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, studying the course material, or reading the reference books.

INTRODUCTION:

- Introduction to course, handout, references, prerequisites, etc.
  - Course web page: probability.ca/sta3431
  - Lectures: Online on Zoom, synchronous, Mondays 10:10–12:00.
  - If not Stat Dept grad student, must REQUEST enrolment (by e-mail); need advanced undergraduate probability/statistics background, plus basic computer programming (including "R"; see e.g. this page).
  - Conversely, if you already know lots about MCMC etc., then this course might not be right for you since it's an INTRODUCTION to these topics.
  - How many of you are graduate students in Statistics? Computer Science? Math? Engineering? Physics? Economics? Management? Other?
- Theme of the course: use (pseudo)randomness on a computer to simulate, and hence estimate, important/interesting quantities.
- Example: Suppose we want to estimate $m := \mathbf{E}[Z^4 \cos(Z)]$, where $Z \sim \text{Normal}(0, 1)$.
  - "Classical" Monte Carlo solution: replicate a large number $z_1, \ldots, z_n$ of Normal(0,1) random variables, and let $x_i = z_i^4 \cos(z_i)$.
  - Their mean $\overline{x} \equiv \frac{1}{n} \sum_{i=1}^{n} x_i$ is an estimate of $\mathbf{E}[X] \equiv \mathbf{E}[Z^4 \cos(Z)]$.
  - R: Z = rnorm(100); X = Z^4 * cos(Z); mean(X) [file "RMC"]
  - Unbiased (good) ... but unstable ... but if replace "100" with "1000000" then $\overline{x}$ is consistently close to -1.213 ... good ...
  - [Aside: In this course we will often use R to automatically sample from simple distributions like Normal, Uniform, Exponential, etc. But how does it work? Discussed later!]
  - Can we quantify the variability?

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

$$se \;=\; \sqrt{\mathrm{Var}(\overline{x})} \;=\; \sqrt{\mathrm{Var}\left(\frac{1}{n}\sum_{i=1}^{n} x_i\right)} \;=\; \sqrt{\left(\frac{1}{n}\right)^2 \left(n\ \mathrm{var}(x)\right)}$$

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

[file "RMCse"]

- Then what is, say, a 95% confidence interval for m?
- Well, by by the Central Limit Theorem (CLT), for large n, we have $\overline{x} \approx N(m, v) \approx N(m, se^2)$.
  – (Strictly speaking, should use "t" distribution, not normal distribution ... but if n large that doesn't really matter – ignore it for now.)
  – So $\frac{m-\overline{x}}{se} \approx N(0,1)$.
  – So, $\mathbf{P}(-1.96 < \frac{m-\overline{x}}{se} < 1.96) \approx 0.95$.
  – So, $\mathbf{P}(\overline{x} - 1.96\,se\ < m < \overline{x} + 1.96\,se\,) \approx 0.95$.
  – i.e., approximate 95% confidence interval is [file "RMCci"]

$$(\overline{x} - 1.96\,se,\ \overline{x} + 1.96\,se)\,.$$

- Alternatively, could compute expectation as

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

Analytic? Numerical? Better? Worse? [file "RMCcomp": -1.213]

- [Aside: In fact, by considering it as the real part of $\mathbf{E}(Z^4 e^{iZ})$, with extra work this expectation can be computed exactly, to be $-2/\sqrt{e} \doteq -1.213061$. But not for harder examples.]
- What about higher-dimensional examples? (Can't do numerical integration!) Can we still sample?
- What if the distribution is too complicated to sample from?
  – (MCMC! Metropolis, Gibbs, etc.... Soon!)

HISTORICAL EXAMPLE – BUFFON'S NEEDLE:

– Have series of parallel lines ... line spacing w, needle length $\ell \le w$ (say $\ell = w$) ... what is prob that needle lands touching line? [https://mste.illinois.edu/activity/buffon/]

- Let h be the vertical distance from the bottom end to the nearest line above, and θ be the needle's angle counter-clockwise from horizontal.
- Then $h \sim \text{Uniform}[0, w]$ and $\theta \sim \text{Uniform}[0, \pi]$, independent.
- Touches line iff $\ell \sin(\theta) > h$, i.e. $h < \ell \sin(\theta)$.
- So, the probability the needle touches the line is:

$$\frac{1}{\pi}\int_0^\pi \frac{1}{w}\int_0^w \mathbf{1}_{h<\ell\sin(\theta)}\, dh\, d\theta = \frac{1}{\pi}\int_0^\pi \frac{1}{w}\ell\sin(\theta)\, d\theta$$

$$= \frac{1}{\pi}\frac{1}{w}\ell\left[-\cos(\theta)\right]_{\theta=0}^{\theta=\pi} = \frac{1}{\pi}\frac{1}{w}\ell\left[-(-1)+(1)\right] = \frac{2\ell}{w\pi}\,.$$

- Hence, by LLN, if throw needle n times, of which it touches a line m times, then for n large, $m/n \approx 2\ell/w\pi$, so $\pi \approx 2n\ell/mw$.
- (e.g. if $\ell = w$, then $\pi \approx 2n/m$)
- [e.g. recuperating English Captain O.C. Fox, 1864: $\ell = 3$, $w = 4$, $n = 530$, $m = 253$, so $\pi \approx 2n\ell/mw \doteq 3.1423$.]
- But for modern simulations, use computer. How to randomise??

PSEUDORANDOM NUMBERS:

- Goal: generate an i.i.d. sequence $U_1, U_2, U_3, \ldots \sim \text{Uniform}[0, 1]$.
- One method: LINEAR CONGRUENTIAL GENERATOR (LCG).
  - Choose (large) positive integers m, a, and b.
  - Start with a "seed" value, x_0. (e.g., the 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. $\sim \text{Uniform}[0, 1]$. (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.
  - Want large "period", i.e. number of iterations before repeat.
  - Many statistical tests, to try to see which choices provide good randomness, avoid correlations, etc. (e.g. "diehard tests", "dieharder": www.phy.duke.edu/~rgb/General/dieharder.php)
  - One common "good" choice: $m = 2^{32}$, $a = 69,069$, $b = 23,606,797$.

- Theorem: the LCG has full period (m) if and only if both:
  (i) $\gcd(b, m) = 1$, and
  (ii) 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 (like above), then the LCG has full period $m = 2^{32} \doteq 4.3 \times 10^9$; good.
- Many other choices, some good, some bad (ASIDE):
  - e.g. C programming language "glibc" uses $m = 2^{32}$, $a = 1,103,515,245$, $b = 12,345$. Pretty good.
  - "RANDU" used $m = 2^{31}$, $a = 65539 = 2^{16} + 3$, $b = 0$ for many years, around the 1970s. Seemed okay. But then people noticed:
    $x_{n+2} = ax_{n+1} = a^2 x_n = (2^{16} + 3)^2 x_n = (2^{32} + 6(2^{16}) + 9)x_n$
    $\equiv (0 + 6(2^{16} + 3) - 9)x_n \pmod{2^{31}} = 6x_{n+1} - 9x_n$.
    So, $x_{n+2} = 6x_{n+1} - 9x_n \bmod m$. Too much serial correlation! Bad!
  - Microsoft Excel pre-2003: had period $< 10^6$, too small ...
  - Excel 2003 instead used a floating-point "version" of LCG ...
    which sometimes gave negative numbers! Bad!
- These numbers are not "really" random, just "pseudorandom" ...
  - Can cause problems! Will fail certain statistical tests!
  - Some implementations also use external randomness, e.g. 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 (see "myhotbits").
  - Or, of atmospheric noise (from lightning etc.), e.g. random.org.
  - But mostly, pseudorandom numbers work pretty well ...
- LCG's are "good enough". But other generators include (ASIDE):
  - "Multiply-with-Carry": $x_n = (ax_{n-r} + b_{n-1}) \bmod m$ where $b_n = \lfloor (ax_{n-r} + b_{n-1})/m \rfloor$.
  - '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 $K_{n+1} = (I + L^{13})(I + R^{18})K_n \bmod 2^{31}$, where L means "shift left" and R means "shift right".
  - 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, and where the matrix multiplication is done bit-wise mod 2.
  - And many others, too. An entire research area!
  - R's choice? See "?RNGkind". Default is Mersenne Twister.

- So, just need computer to do simple arithmetic. No problem, right?

LIMITATIONS OF COMPUTER ARITHMETIC:

- Consider the following computations in R:
  > 2 + 1 - 2
  > 2∧10 + 1 - 2∧10
  > 2∧100 + 1 - 2∧100
- Why??
- Question for next class: In R, for what values of n does:
  > 2∧n + 1 - 2∧n
  give 0 instead of 1?
- (Similarly in many other computer languages too, e.g. C (powertest.c), Java (powertest.java) ... and Python with floating numbers ... but not Python with *integer* variables (powertest.py), because it then does dynamic memory allocation ...)
- Also, overflow/Inf/underflow problems: 2∧10000, 2∧(−10000), etc.
- So, numerical computations are approximations, with their own errors.
- We'll usually ignore these issues, but you should BE CAREFUL!
- So how to use pseudorandomness?
  - With LCG etc, we can simulate Uniform[0,1] random variables.
  - What about other random variables?

END WEEK #1

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 the "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:

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

(where by convention $\sum_{k=0}^{-1}(\cdots) = 0$). ("Inverse CDF method"; see below)

- More generally, to make $\mathbf{P}(Y = x_i) = p_i$ for some $x_1 < x_2 < x_3 < \ldots$, where $p_i \ge 0$ and $\sum_i p_i = 1$, set:

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

(discrete version of "Inverse CDF method")

- e.g. to make $Z \sim$ Exponential(1), set:
  $Z = -\log(U_1)$.
  - Then for $x > 0$, $\mathbf{P}(Z > x) = \mathbf{P}(-\log(U_1) > x) = \mathbf{P}(\log(U_1) < -x)$ $= \mathbf{P}(U_1 < e^{-x}) = e^{-x}$. Then CDF $= 1 - e^{-x}$, and density $= e^{-x}$.
  - Then, to make $W \sim$ Exponential(λ), set:
    $W = Z/\lambda = -\log(U_1)/\lambda$. [So that W has density $\lambda e^{-\lambda x}$ for $x > 0$.]
- Suppose we want X to have density $6\,x^5 \mathbf{1}_{0<x<1}$.
  - Let $X = U_1^{1/6}$.
  - Then for $0 < x < 1$, $\mathbf{P}(X \le x) = \mathbf{P}(U_1^{1/6} \le x) = \mathbf{P}(U_1 \le x^6) = x^6$.
  - Hence, $f_X(x) = \frac{d}{dx}\Big[\mathbf{P}(X \le x)\Big] = \frac{d}{dx} x^6 = 6\,x^5$ for $0 < x < 1$.
  - More generally, for $r > 1$, if $X = U_1^{1/r}$, then $f_X(x) = r\,x^{r-1}$ for $0 < x < 1$. [CHECK!]
- 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 $X, Y \sim N(0,1)$, and X and Y are independent! ["Box-Muller transformation": Ann Math Stat 1958, 29, 610-611]

  - Proof (Aside): 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

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

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

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

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

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

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

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

- Most general approach: the general "INVERSE CDF METHOD":
  - Suppose want $\mathbf{P}(X \le x) = F(x)$. ("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 $U_1 \le F(x)$. [Subtle; see e.g. Rosenthal, *A First Look at Rigorous Probability Theory*, 2nd ed., Lemma 7.1.2.]
  - So, $\mathbf{P}(X \le x) \ = \ \mathbf{P}(U_1 \le F(x)) \ = \ F(x)$.
  - Very general, but computing $F^{-1}(t)$ is often very difficult ...
- Overall, generating (pseudo)random numbers for most "standard" one-dimensional distributions is mostly pretty easy or well-known ...
  - So, can get Monte Carlo estimates of expectations involving standard one-dimensional distributions, e.g. $\mathbf{E}[Z^4 \cos(Z)]$ where $Z \sim \text{Normal}(0,1)$.
- What about other Monte Carlo estimates?

MONTE CARLO INTEGRATION:

- How to compute an integral with Monte Carlo?
  - Re-write it as an expectation!
- EXAMPLE: Want to compute $\int_0^1 \int_0^1 g(x,y)\,dx\,dy$.
  - Regard this as $\mathbf{E}[g(X,Y)]$, where X, Y i.i.d. $\sim \text{Uniform}[0,1]$.
  - Then, similar to before, estimate $\mathbf{E}[g(X,Y)]$ by $\frac{1}{M}\sum_{i=1}^M g(x_i, y_i)$, where $x_i \sim \text{Uniform}[0,1]$ and $y_i \sim \text{Uniform}[0,1]$ (all independent).
  - e.g. $g(x,y) = \cos(\sqrt{xy}\,)$. (file "RMCint") Easy!
  - Get about 0.88 ± 0.003 ... Mathematica gives 0.879544.
- 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 $\frac{1}{M}\sum_{i=1}^{M}(5\cdot 4\cdot g(X_i,Y_i))$. (file "RMCint2")
- Standard error: $se = M^{-1/2}\ sd(5\cdot 4\cdot g(X_1,Y_1),\ \ldots,\ 5\cdot 4\cdot g(X_M,Y_M))$.
- With $M = 10^6$, get about -4.11 ± 0.01 ... Mathematica gives -4.11692.

- 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, say, $\int_0^1\int_0^\infty (e^y\,h(x,y))\,e^{-y}\,dy\,dx$.
  - This is the same as $\mathbf{E}[e^Y\,h(X,Y)]$, where $X \sim$ Uniform$[0,1]$ and $Y \sim$ Exponential(1) are independent.
  - So, estimate it by $\frac{1}{M}\sum_{i=1}^{M} e^{Y_i}h(X_i,Y_i)$, where $X_i \sim$ Uniform$[0,1]$ and $Y_i \sim$ Exponential(1) (i.i.d.). (file "RMCint3")
  - With $M = 10^6$ get about 0.767 ± 0.0004 ... Small error!
  - Mathematica: 0.767211.

- 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 $I \equiv \int s(y)\,dy$, could write it as $I = \int \frac{s(x)}{f(x)}\,f(x)\,dx$, where f is easily sampled from, with $f(x) > 0$ whenever $s(x) > 0$.
  - Then $I = \mathbf{E}\left(\frac{s(X)}{f(X)}\right)$, where X has density f.
  - ("Importance Sampling")
  - So, $I \approx \frac{1}{M}\sum_{i=1}^{M}\frac{s(x_i)}{f(x_i)}$ where $x_i \sim f$.
  - Can then do classical (iid) Monte Carlo integration, and also get standard errors, confidence intervals, etc.
  - Good if it's easier to sample from f, and/or if the function $\frac{s(x)}{f(x)}$ is less variable than h itself.

- In general, best to make $\frac{s(x)}{f(x)}$ approximately constant if possible.
  - 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). [Here $s(x) = e^{-3x}$, and $f(x) = 3e^{-3x}$.]

UNNORMALISED DENSITIES:

- Suppose now that $\pi(y) = c\,g(y)$, where we know g but don't 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, $I = \int h(x)\,\pi(x)\,dx = \int h(x)\,c\,g(x)\,dx = \frac{\int h(x)\,g(x)\,dx}{\int g(x)\,dx}$.
  - Here $\int h(x)\,g(x)\,dx = \int \Big(h(x)\,g(x)\,/\,f(x)\Big)\,f(x)\,dx = \mathbf{E}[h(X)\,g(X)\,/\,f(X)]$ where $X \sim f$.
  - So, $\int h(x)\,g(x)\,dx \approx \frac{1}{M}\sum_{i=1}^{M}\Big(h(x_i)\,g(x_i)\,/\,f(x_i)\Big)$ if $\{x_i\} \sim f$ (i.i.d.).
  - Similarly, $\int g(x)\,dx \approx \frac{1}{M}\sum_{i=1}^{M}\Big(g(x_i)\,/\,f(x_i)\Big)$ if $\{x_i\} \sim f$ (i.i.d.).
  - So, must $I \approx \dfrac{\sum_{i=1}^{M}\Big(h(x_i)\,g(x_i)\,/\,f(x_i)\Big)}{\sum_{i=1}^{M}\Big(g(x_i)\,/\,f(x_i)\Big)}$.
  - ("Importance Sampling": weighted average)
  - (Because we are taking ratios of (unbiased) estimates, the resulting estimate is not unbiased, and its standard errors are less clear. But it is still consistent as $M \to \infty$.)

END WEEK #2

- 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 $c > 0$ is 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 $f(y) = 6\,y^5\mathbf{1}_{0<y<1}$.
  - [Recall: if $U \sim \text{Uniform}[0,1]$, and if $X = U^{1/6}$, then $X \sim f$.]
  - Then $I \approx \dfrac{\sum_{i=1}^{M}\Big(h(x_i)\,g(x_i)\,/\,f(x_i)\Big)}{\sum_{i=1}^{M}\Big(g(x_i)\,/\,f(x_i)\Big)} = \dfrac{\sum_{i=1}^{M}\Big(\sin(x_i^4)\,\cos(x_i^5)\Big)}{\sum_{i=1}^{M}\Big(\sin(x_i^4)\,\cos(x_i^5)\,/\,x_i^2\Big)}$, where $\{x_i\}$ are i.i.d. $\sim f$. (file "Rimp1" ... get about 0.766 ...)
  - Or, let $f(y) = 4\,y^3\mathbf{1}_{0<y<1}$. [Then if $U \sim \text{Uniform}[0,1]$, 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 "Rimp2")
  - Numerical integration: 0.7661155 (file "Rimp3").
- With importance sampling, is it important to use the same samples $\{x_i\}$ in both numerator and denominator?
  - What if independent samples are used instead?
  - Let's try it! (file "Rimpind")
  - Both ways work, but usually(?) the same samples work better.
  - Overall, good to use same sample $\{x_i\}$ for both numerator and denominator: easier computationally, and leads to smaller variance.

- What other methods are available 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. (i.e., $K\,f(x) \geq \pi(x)\,/\,c$, i.e. $c\,K\,f(x) \geq \pi(x)$)
- Sample $X \sim f$, and $U \sim \text{Uniform}[0,1]$ (indep.).
  - If $U \leq \frac{g(X)}{Kf(X)}$, then accept X (as a draw from π).
  - Otherwise, reject X and start over again.
- Does this algorithm give valid samples?
  - Well, conditional on accepting, we have for any $y \in \mathbf{R}$ that

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

  - Since $0 \leq \frac{g(x)}{Kf(x)} \leq 1$, therefore $\mathbf{P}(U \leq \frac{g(X)}{Kf(X)} \,|\, X = x) = \frac{g(x)}{Kf(x)}$.
  - Hence, by the double expectation formula, $\mathbf{P}\Big(U \leq \frac{g(X)}{Kf(X)}\Big) = \mathbf{E}\Big[\mathbf{P}\Big(U \leq \frac{g(X)}{Kf(X)}\Big|X\Big)\Big] = \mathbf{E}\Big[\frac{g(X)}{Kf(X)}\Big] = \int_{-\infty}^{\infty} \frac{g(x)}{Kf(x)}\; f(x)\,dx = \frac{1}{K}\int_{-\infty}^{\infty} g(x)\,dx$.
  - Similarly, for any $y \in \mathbf{R}$, $\mathbf{P}\Big(X \leq y,\; U \leq \frac{g(X)}{Kf(X)}\Big) = \mathbf{E}\Big[\mathbf{1}_{X\leq y}\,\mathbf{1}_{U \leq \frac{g(X)}{Kf(X)}}\Big] = \mathbf{E}\Big[\mathbf{1}_{X\leq y}\,\mathbf{P}\Big(U \leq \frac{g(X)}{Kf(X)}\Big|X\Big)\Big] = \mathbf{E}\Big[\mathbf{1}_{X\leq y}\;\frac{g(X)}{Kf(X)}\Big] = \int_{-\infty}^{y} \frac{g(x)}{Kf(x)}\; f(x)\,dx = \frac{1}{K}\int_{-\infty}^{y} g(x)\,dx$.
- So,

$$\mathbf{P}\Big(X \leq y \,\Big|\, U \leq \frac{g(X)}{Kf(X)}\Big) \;=\; \frac{\frac{1}{K}\int_{-\infty}^{y} g(x)\,dx}{\frac{1}{K}\int_{-\infty}^{\infty} g(x)\,dx} \;=\; \int_{-\infty}^{y} \pi(x)\,dx\,.$$

  - So, conditional on accepting, $X \sim \pi$. Good! iid!
- However, prob. of accepting may be very small.
  - If so, then get very few samples – bad.
- 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. $h(x) = x^4$. (Should be 3.)
  - Let f be double-exponential (Laplace) 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 $|x| \ge 2$, $Kf(x) = 8\,\frac{1}{2}\,\exp(-|x|) \ge 8\,\frac{1}{2}\,\exp(-x^2/2) \ge (2\pi)^{-1/2}\exp(-x^2/2) = \pi(x) = g(x)$.
- See graph: file "Rrejgraph".

- 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.
  - Try it: 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.
- Note: these algorithms all work in discrete case too.
  - Can let π, f, etc. be "probability functions", i.e. probability densities with respect to counting measure.
  - Then the algorithms proceed exactly as before.

AUXILIARY VARIABLE APPROACH:

- (related: "slice sampler")
- Suppose $\pi(x) = c\,g(x)$, and (X, Y) chosen uniformly under 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? Well, 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 $Y \le g(X)$.
- 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")

DIGRESSION – QUEUEING THEORY:

- Consider a long line (queue) of customers.
  - Let $Q(t)$ = number of people in queue at time $t \ge 0$.
- 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 [e.g. STA447/2006]:
  - If $\mu \le \lambda$, 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"]

END WEEK #3

DIGRESSION – MONTE CARLO IN FINANCE:

- X_t = stock price at time t
- Assume that $X_0 = a > 0$, and $dX_t = bX_t dt + \sigma X_t dB_t$, where $\{B_t\}$ is Brownian motion. ("diffusion")
  - 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 price 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 $\mathbf{E}(e^{-rT} \max(0,\ X_T - q))$, 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 formula ("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 one run is highly variable. (file "RBS", with $M = 1$)
  - But many runs give good estimate. (file "RBS", with $M = 1000$)
- 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 formulas not so simple ... but can still simulate! (file "RAO")
- So, can iid / importance / rejection / auxiliary sampling solve ALL of our problems? No!
  - Many challenging cases arise, e.g. from Bayesian statistics (later).
  - Some are high-dimensional, and the above algorithms fail.
  - Alternative algorithm: MCMC!

*** 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.) Difficult!
- Define a Markov chain (dependent random process: STA2006) $X_0, X_1, X_2, \ldots$ in such a way that for large enough n, $X_n \approx \pi$.

- 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") is chosen large enough so $X_B \approx \pi$, and M is chosen large enough to get good Monte Carlo estimates.

- How to design such a Markov chain? One good way is:
- METROPOLIS ALGORITHM (1953):
  - Choose some initial value X_0 (perhaps random).
  - Then, given X_{n-1}, choose a proposal state $Y_n \sim MVN(X_{n-1},\, \sigma^2\, I)$ for some fixed $\sigma > 0$ (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 the normalising constant c cancels and is not required.)
  - (Why does it work? Markov chain theory – later!)
  - Try it: www.probability.ca/metropolis (Javascript; formerly Java.)
  - Note: This version is called "random walk Metropolis" (RWM). Why? Because the proposals, if we always accepted them, would form a traditional random walk process.
- How large B? Difficult to say! Some theory (later) ... usually just use trial-and-error / statistical analysis of output, and hope for the best ...
- What initial value X_0?
  - Virtually any one will do, but "central" ones best.
  - Can also use an "overdispersed starting distribution": choose X_0 randomly from some distribution that "covers" the "important" parts of the state space. Good for checking consistency ...
- 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".
  - acf: has some serial autocorrelation $\mathbf{E}(X_n X_{n+k})$. Important! (Soon.)
- 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 (file "Rmet2") ... works, but large uncertainty.
- Gets between about 34 and 44 ... (Mathematica gets 38.7044)
- Individual plots appear to have "good mixing" ...
- Joint plot shows fewer samples where $x_1 x_2 \approx (\pi/2)^2 \doteq 2.5$ ...

END WEEK #4

- OPTIMAL SCALING:
  - What if we change σ? How does that affect estimate? plot? acf?
  - Can change proposal distribution to $Y_n \sim MVN(X_{n-1}, \sigma^2 I)$ for any choice of $\sigma > 0$. 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 ("www.probability.ca/metropolis", "Rmet", "Rmet2").
  - Some theory ... limited ... active area of research ...
  - General principle: the acceptance rate should be far from 0 and far from 1.
  - Surprising Fact: 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] (More later!)

MCMC STANDARD ERROR:

- What about MCMC's standard error, i.e. uncertainty?
  - It's usually larger than in the i.i.d. case (due to the positive correlations), and harder to quantify.
- Simplest: re-run the chain many times, with same M and B, with different initial values drawn from some overdispersed starting distribution, and get a fresh estimate each time, and then compute the standard error of the sequence of estimates.
  - Then can analyse the estimates obtained as iid ...
- But how to estimate standard error from a single run?
- i.e., how to estimate $v \equiv \text{Var}\left(\frac{1}{M-B}\sum_{i=B+1}^{M} h(X_i)\right)$?
  - For simplicity, 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 $M - B$,

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

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

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

$$= \frac{1}{M-B}\Big(\mathrm{Var}_\pi(h) + 2\,\mathrm{Cov}_\pi(h(X_i), h(X_{i+1})) + 2\,\mathrm{Cov}_\pi(h(X_i), h(X_{i+2})) + \dots\Big)$$

$$= \frac{1}{M-B}\,\mathrm{Var}_\pi(h)\Big(1 + 2\,\mathrm{Corr}_\pi(h(X_i), h(X_{i+1})) + 2\,\mathrm{Corr}_\pi(h(X_i), h(X_{i+2})) + \dots\Big)$$

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

where

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

$$= 2\Big(\sum_{k=0}^{\infty}\rho_k\Big) - 1 = \sum_{k=-\infty}^{\infty}\rho_k$$

since $\rho_0 = 1$ and $\rho_{-k} = \rho_k$.

– Also called “integrated auto-correlation time” or “ACT”.

- Then can estimate both iid variance, and varfact, from the sample run.
- Note: to compute varfact, don't sum over <u>all</u> k, just e.g. until, say, $|\rho_k| < 0.05$ or $\rho_k < 0$ or ...
  – (Can use R's built-in “acf” function, hopefully with a good choice of the “lag.max” parameter. Or can write your own – better.)
- Then standard error $= se = \sqrt{v} = (iid\text{-}se)\sqrt{\text{varfact}}$.
- e.g. the files Rmet and Rmet2. (Recall: true answers are about 0.766 and 38.7, respectively.)
  – Usually varfact $\gg 1$; try to get “better” chains so varfact smaller.
  – Sometimes even try to design chain to get varfact < 1 (“antithetic”).
  – Work in parallel? (Antithetically??) Some work, but limited. (Project?)

CONFIDENCE INTERVALS:

- Suppose we estimate $u \equiv \mathbf{E}_\pi(h)$ by the quantity $e = \frac{1}{M-B} \sum_{i=B+1}^{M} h(X_i)$, 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 probability 95%, the interval $(e - 1.96\sqrt{v},\ e + 1.96\sqrt{v})$ will contain u.
  - (Again, strictly speaking, should use something like a "t" distribution, not the normal distribution ... but if $M - B$ large that doesn't really matter – ignore it for now.)
- e.g. the files Rmet and Rmet2. (Recall: true answers are about 0.766 and 38.7, respectively.)
- But does a CLT even hold??
  - Does not follow from classical i.i.d. CLT. Does not always hold. But often does.
  - For example, CLT holds if chain is "geometrically ergodic" (later!) 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.)
  - Can get alternative (slightly larger) confidence intervals even without a CLT, if have consistent variance estimator: Rosenthal, "Simple confidence intervals for MCMC without CLTs", EJS 2017.)
- So MCMC is more complicated than standard Monte Carlo.
  - Why should we bother? Some problems are too challenging for other methods! (e.g. Bayesian – later)

METROPOLIS-HASTINGS ALGORITHM:

- The Metropolis algorithm doesn't always work well.
  - Sometimes other MCMC algorithms can help too.
  - With above theory, can derive other valid algorithms!
- Note: key facts about $q(x, y)$ are symmetry, and irreducibility.
  - So, could replace $Y_n \sim N(X_{n-1}, 1)$ by e.g. $Y_n \sim \text{Uniform}[X_{n-1} - 1, X_{n-1} + 1]$, or (on discrete space) $Y_n = X_{n-1} \pm 1$ prob. $\frac{1}{2}$ each, etc.

- Metropolis algorithm still works provided proposal distribution is symmetric, i.e. $q(x, y) = q(y, x)$.
- But what if q is not symmetric?

- Hastings, Biometrika 1970 [Canadian! see www.probability.ca/hastings]:
  - Claim: If we 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 the algorithm is still valid even if q is not symmetric.
  - That is, we 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 $x \to y$.)
  - Do require that $q(x, y) > 0$ iff $q(y, x) > 0$.
  - Why is it valid? Later!

- 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 $h(x_1, x_2) = e^{x_1} + (x_2)^2$. (Mathematica gives $\mathbf{E}_\pi(h) \doteq 38.7044$.)
  - 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, $q(x, y) = C\,(1 + |x|^2)^{-2}\, \exp(-|y - x|^2 \,/\, 2\,\sigma^2 (1 + |x|^2)^2)$.
  - Then, can run Metropolis-Hastings algorithm. (file "RMH")
  - Usually get between 34 and 43, with claimed standard error ≈ 2.

INDEPENDENCE SAMPLER:

- Propose $\{Y_n\} \sim q(\cdot)$, i.e. the $\{Y_n\}$ are i.i.d. from some fixed density q, independent of X_{n-1}. (e.g. $Y_n \sim MVN(0,\, I_d)$)
  - Then accept if $U_n < A_n$ where $U_n \sim \text{Uniform}[0, 1]$ and $A_n = \frac{\pi(Y_n)\, q(X_{n-1})}{\pi(X_{n-1})\, q(Y_n)}$.
  - Special case of the Metropolis-Hastings algorithm, where $Y_n \sim q(X_{n-1},\, \cdot)$, and $A_n = \frac{\pi(Y_n)\, q(Y_n, X_{n-1})}{\pi(X_{n-1})\, q(X_{n-1}, Y_n)}$.
  - 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).

- e.g. independence sampler with $\pi(x) = e^{-x}$ and $q(y) = ke^{-ky}$ for $x > 0$.
  - 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).
  - And with $k = 5$, confidence intervals often miss 1. (file "Rind2")

- Why is large k so much worse than small k? (Later!)

LANGEVIN ALGORITHM:

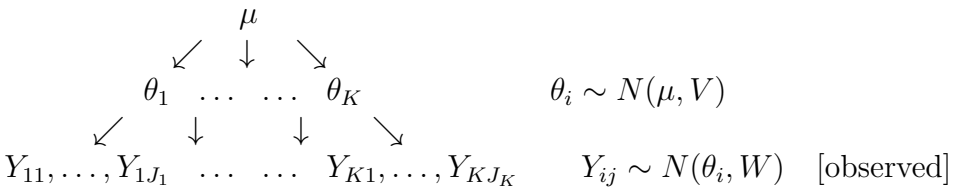
- Special case of Metropolis-Hastings algorithm.
- $Y_n \sim MVN(X_{n-1} + \frac{1}{2}\,\sigma^2\,\nabla \log \pi(X_{n-1}),\ \sigma^2 I)$.
- Intuition: tries to move in direction where π increasing.
- Based on discrete approximation to "Langevin diffusion".
- Usually more efficient, but requires knowledge and computation of $\nabla \log \pi$.
- For theory, see e.g. Roberts & Tweedie, Bernoulli **2(4)**, 341–363, 1996; Roberts & Rosenthal, JRSSB **60**, 255–268, 1998.

COMPONENTWISE (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).
  - 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 $j \neq i$.
  - (Here $Y_{n,i}$ is the i^{th} coordinate of Y_n.)
- Then accept/reject with usual Metropolis rule (symmetric proposals: "Componentwise Metropolis", or "Variable-at-a-time Metropolis", or "Metropolis-within-Gibbs") or Metropolis-Hastings rule (non-symmetric proposals: "Componentwise Metropolis-Hastings", or "Variable-at-a-time Metropolis-Hastings", or "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 "Rcompwise") or random-scan (file "Rcompwise2").
- So, lots of MCMC algorithms to choose from.
  - Why do we need them all?
  - To compute with complicated models! For example ...

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
$$\pi(\theta) \;=\; \frac{f_{\theta,Y}(\theta, y)}{f_Y(y)} \;=\; C\, f_{\theta,Y}(\theta, y) \;=\; C\, f_\theta(\theta)\ f_{Y|\theta}(y, \theta)\,.$$
- 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"):
$$V \sim IG(a_1, b_1); \qquad W \sim IG(a_2, b_2); \qquad \mu \sim N(a_3, b_3)$$
(indep), where a_i, b_i known constants, and $IG(a, b)$ is the "inverse gamma" distribution, with density $\frac{b^a}{\Gamma(a)}\, e^{-b/x}\, x^{-a-1}$ for $x > 0$.
- Combining the above dependencies, we see that the joint density is (for $V, W > 0$):
$$f(V, W, \mu, \theta_1, \ldots, \theta_K, Y_{11}, Y_{12}, \ldots, Y_{KJ_K})$$
$$= \left(\frac{b_1^{a_1}}{\Gamma(a_1)} e^{-b_1/V} V^{-a_1-1}\right) \left(\frac{b_2^{a_2}}{\Gamma(a_2)} e^{-b_2/W} W^{-a_2-1}\right) \left(\frac{1}{\sqrt{2\pi b_3}} e^{-(\mu-a_3)^2/2b_3}\right) \times$$

$$\times \left(\prod_{i=1}^{K} \frac{1}{\sqrt{2\pi V}} e^{-(\theta_i-\mu)^2/2V}\right) \left(\prod_{i=1}^{K}\prod_{j=1}^{J_i} \frac{1}{\sqrt{2\pi W}} e^{-(Y_{ij}-\theta_i)^2/2W}\right)$$

$$= C_2\, e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-\frac{1}{2}\sum_{i=1}^{K} J_i} \times$$

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

- Then

$$\pi(V, W, \mu, \theta_1, \ldots, \theta_K)$$

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

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

$$= C_3\, e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-\frac{1}{2}\sum_{i=1}^{K} J_i} \times$$

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

- NOTE: Many applications of variance components model, e.g.:
  - Predicting success at law school (D. Rubin, JASA 1980), $K = 82$ schools.
  - Melanoma (skin cancer) recurrence (`http://www.mssanz.org.au/MODSIM07/papers/52_s24/Analysing_Clinicals24_Bartolucci_.pdf`), with $K = 19$ different patient categories.
  - 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, $J_i \equiv 5$. (data in file "Rdye")
- Here, the dimension is $d = K + 3$, e.g. $K = 19$, $d = 22$. High!
- How to compute/estimate, say, $\mathbf{E}_\pi(W/V)$, or the effect of changing b_1?
  - Numerical integration? No, too high-dimensional!
  - Importance sampling? Perhaps, but what "f"? Too inefficient!
  - Rejection sampling? What "f"? What "K"? Virtually no samples!
  - Perhaps MCMC can work!
  - But need clever, useful MCMC algorithms!
  - Perhaps Metropolis, or ...
- ASIDE: For big complicated π, often better to work with logarithms, e.g. accept iff $\log(U_n) < \log(A_n) = \log(\pi(Y_n)) - \log(\pi(X_{n-1}))$.
  - Then only need to compute $\log(\pi(x))$; helps avoid overflow problems.
  - So, better to program on log scale: $\log \pi(V, W, \mu, \theta_1, \ldots, \theta_K) = \ldots$.

– Also sometimes simpler, e.g. if $\pi(x) = \exp\left(\sum_{i<j} |x_j - x_i|\right)$, then $\log(\pi(x)) = \sum_{i<j} |x_j - x_i|$. (Best to type in the log formula directly.)

GIBBS SAMPLER:

- (Special case of Componentwise Metropolis-Hastings.)
- Proposal distribution for i^{th} coordinate is equal to the full conditional distribution of that coordinate (according to π), conditional on the current values of all the other coordinates.
  - Can use either systematic or random scan, just like above.
  - Then, always accept. Why? Later!
  - (Intuition: if start in stationary distribution, then update one coordinate from its conditional stationary distribution (and always accept), then the distribution remains the same, i.e. stationarity is preserved.)

END WEEK #5

- EXAMPLE: Variance Components Model:
  - Update of μ (say) should be from conditional density of μ, conditional on current values of all the other coordinates: $\mathcal{L}(\mu \,|\, V, W, \theta_1, \ldots, \theta_K, Y_{11}, \ldots, Y_{J_K K})$.
  - This conditional density is proportional to the full joint density, but with all variables besides μ treated as constant.
  - Recall: full joint density is:

$$= C_3 e^{-b_1/V} V^{-a_1-1} e^{-b_2/W} W^{-a_2-1} e^{-(\mu-a_3)^2/2b_3} V^{-K/2} W^{-\frac{1}{2}\sum_{i=1}^K J_i} \times$$
$$\times \exp\left[-\sum_{i=1}^K (\theta_i - \mu)^2/2V\right] \exp\left[-\sum_{i=1}^K \sum_{j=1}^{J_i} (Y_{ij} - \theta_i)^2/2W\right].$$

  - So, combining "constants" (w.r.t. μ), the conditional density of μ is

$$C_4 \, e^{-(\mu-a_3)^2/2b_3} \exp\left[-\sum_{i=1}^K (\theta_i - \mu)^2/2V\right].$$

  - This equals (verify this! HW!)

$$C_5 \exp\Big(-\mu^2\big(\frac{1}{2b_3} + \frac{K}{2V}\big) + \mu\big(\frac{a_3}{b_3} + \frac{1}{V}\sum_{i=1}^K \theta_i\big)\Big).$$

  - Side calculation: if $\mu \sim N(m, v)$, then density $\propto e^{-(\mu-m)^2/2v} \propto e^{-\mu^2(1/2v) + \mu(m/v)}$.
  - Hence, here $\mu \sim N(m, v)$, 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 μ is updated, we sample it from $N(m, v)$ for this m and v (and always accept).

- Similarly (HW!), conditional distribution for V is:

$$C_6 e^{-b_1/V} V^{-a_1-1} V^{-K/2} \exp\left[-\sum_{i=1}^{K} (\theta_i - \mu)^2/2V\right], \qquad V > 0\,.$$

- 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 $IG(a_1 + K/2,\ b_1 + \frac{1}{2}\sum_{i=1}^{K}(\theta_i - \mu)^2)$.

- Can similar compute conditional distributions for W and θ_i (HW!).
- The systematic-scan Gibbs sampler then proceeds (HW!) 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.
- Or, the random-scan Gibbs sampler proceeds by choosing one of $V, W, \mu, \theta_1, \ldots, \theta_K$ uniformly at random, and then updating that coordinate from its corresponding conditional distribution.
  - Then repeat this step M times [or $M(K + 3)$ times?].
  - How well does it work? HW!

SUBSAMPLING (THINNING):

- The autocorrelations (acf) of an MCMC run usually start near 1, and then decrease until they become negligible after some lag L. (file "Rmetnorm")
  - This means that every L^{th} iteration of the chain is approximately independent.
  - But also, after the burn-in B, the chain is approximately in stationarity, i.e. has distribution approximately π.
  - So, the states $X_B, X_{B+L}, X_{B+2L}, X_{B+3L}, \ldots$ are approximately i.i.d. samples from π. Good?
- When running MCMC, some people use "subsampling" (or, "thinning") by just using every S^{th} sample, i.e. by using the Markov chain $\{X_{B+Sn}\}$.
  - If $S \geq L$, then the $\{X_{B+Sn}\}$ are approximately independent.
  - This could be useful: i.i.d. samples, classical standard error and confidence intervals, good tests for accuracy, etc.

- But does it improve the actual estimator? e.g. its variance?
  - The number of samples is reduced from $M - B$ to $(M - B)/S$, which increases the estimator variance by a factor of S. Bad.
  - But the varfact is reduced from $1 + 2\sum_{k=1}^{\infty} \rho_k$ to $1 + 2\sum_{k=1}^{\infty} \rho_{Sk}$ (good).
  - So, which is larger, $1 + 2\sum_{k=1}^{\infty} \rho_k$, or $S\,(1 + 2\sum_{k=1}^{\infty} \rho_{Sk})$?
  - Well, we can re-write $S\,(1 + 2\sum_{k=1}^{\infty} \rho_{Sk})$ as

$$(1 + \ldots + 1) + (2\rho_S + \ldots + 2\rho_S) + (2\rho_{2S} + \ldots + 2\rho_{2S}) + \ldots .$$

  - We want to compare this to

$$(1 + 2\rho_1 + \ldots + 2\rho_{S-1}) + (2\rho_S + 2\rho_{S+1} + \ldots + 2\rho_{2S-1}) + (2\rho_{2S} + \ldots) + \ldots .$$

  - Assuming the ρ_i are decreasing, and ignoring some of the "2" factors, the first (subsampled) variance is larger.
- This calculation seems to suggest that subsampling is <u>not</u> advantageous.
  - In fact, for "reversible" Markov chains (defined later), consecutive sums $\rho_{2i} + \rho_{2i+1}$ <u>are</u> decreasing, and the first (subsampled) variance is always larger: Geyer 1992
  - More generally, even if not reversible, subsampling always increases the estimator variance: Maceachern & Berliner, 1994
- Conclusion: Subsampling makes the estimator variance <u>larger</u>, i.e. subsampling does <u>not</u> improve the estimator.
  - However, if lots of extra computation is required to compute the functional values $h(X_i)$, then subsampling might help: Owen 2017

JUSTIFICATION: WHY DOES METROPOLIS ALG WORK?:

- (Uses Markov chain theory ... e.g. STA447/2006 ... already know?)
- <u>Basic fact:</u> if a Markov chain is "irreducible" and "aperiodic", with "stationarity distribution" π, then $\mathcal{L}(X_n) \to \pi$ as $n \to \infty$. More precisely:
- 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 $S \subseteq \mathcal{X}$.
- Let's figure out what this all means ...
- Notation: $P(i, j) = \mathbf{P}(X_{n+1} = j \,|\, X_n = i)$ (discrete case), or $P(x, A) = \mathbf{P}(X_{n+1} \in A \,|\, X_n = x)$ (general case). Also $\Pi(A) = \int_A \pi(x)\,dx$.

- Well, irreducible means that you have positive probability of eventually getting from anywhere to anywhere else.
  - Discrete case: for all $i, j \in \mathcal{X}$ (the state space), there is $n \in \mathbf{N}$ such that $P(X_n = j \,|\, X_0 = i) > 0$.
  - Actually, we only need to require this for states j such that $\pi(j) > 0$.
  - General case: for all $x \in \mathcal{X}$, and for all $A \subseteq \mathcal{X}$ with $\Pi(A) > 0$, there is $n \in \mathbf{N}$ such that $P(X_n \in A \,|\, X_0 = x) > 0$. ("$\pi$-irreducible")
  - (Since usually $P(X_n = y \,|\, X_0 = x) = 0$ for all y.)
  - Irreducibility is usually satisfied for MCMC.
- And, aperiodic means there are no forced cycles, i.e. there do not exist disjoint non-empty subsets $\mathcal{X}_1, \mathcal{X}_2, \ldots, \mathcal{X}_d$ for $d \geq 2$, such that $P(x, \mathcal{X}_{i+1}) = 1$ for all $x \in \mathcal{X}_i$ $(1 \leq i \leq d-1)$, and $P(x, \mathcal{X}_1) = 1$ for all $x \in \mathcal{X}_d$. [Diagram.]
  - (Discrete case: Equivalent that $gcd\{n : p^n(i,i) > 0\} = 1 \ \forall i$.)
  - This is true for virtually any Metropolis algorithm, e.g. it's true if $P(x, \{x\}) > 0$ for any one state $x \in \mathcal{X}$, e.g. if positive prob of rejection.
  - Also true if $P(x, \cdot)$ has positive density throughout S, for all $x \in S$, for some $S \subseteq \mathcal{X}$ with $\Pi(S) > 0$. (e.g. Normal proposals)
  - Not quite guaranteed, e.g. $\mathcal{X} = \{0, 1, 2, 3\}$, and π uniform on $\mathcal{X}$, and $Y_n = X_{n-1} \pm 1 \pmod 4$. [Diagram.] But almost always holds.
- What about Π being a stationary distribution?
  - This means that if we start with the probabilities Π, and then run the Markov chain for one step, that we will still have the probabilities Π.
  - Will this be true for the Metropolis algorithm?

END WEEK #6

- Begin with DISCRETE CASE (e.g. www.probability.ca/metropolis).
  - State space is $\mathcal{X}$, e.g. $\mathcal{X} \equiv \{1, 2, 3, 4, 5, 6\}$.
- Here stationary means that if $X_0 \sim \pi$, i.e. $\mathbf{P}(X_0 = i) = \pi(i)$ for all i, then also $X_1 \sim \pi$, i.e. $\mathbf{P}(X_1 = j) = \pi(j)$ for all j.
  - But $\mathbf{P}(X_1 = j) = \sum_{i \in S} \mathbf{P}(X_0 = i, X_1 = j) = \sum_{i \in S} \mathbf{P}(X_0 = i)\, P(i, j)$.
  - So, π is stationary if $\sum_{i \in S} \pi(i)\, P(i, j) = \pi(j)$ for all j.
- Let $q(x, y) = \mathbf{P}(Y_n = y \,|\, X_{n-1} = x)$ be the proposal distribution, e.g. perhaps $q(x, x+1) = q(x, x-1) = 1/2$.
  - Assume that q is symmetric, i.e. $q(x, y) = q(y, x)$ for all $x, y \in \mathcal{X}$.
  - Then if $\alpha(x, y)$ is the probability of accepting a proposed move from

x to y, then

$$\alpha(x,y) = \mathbf{P}(U_n < A_n \,|\, X_{n-1} = x,\, Y_n = y)$$

$$= \mathbf{P}(U_n < \frac{\pi(y)}{\pi(x)}) \;=\; \min[1,\; \frac{\pi(y)}{\pi(x)}]\,.$$

– (Assume for simplicity that $\pi(x) > 0$.)

- Then we compute that for $i, j \in \mathcal{X}$ with $i \neq j$,

$$P(i,j) \;=\; q(i,j)\;\alpha(i,j) \;=\; q(i,j)\;\min(1,\; \frac{\pi(j)}{\pi(i)})\,.$$

– Hence, using the symmetry of q,

$$\pi(i)\,P(i,j) \;=\; q(i,j)\;\min(\pi(i),\;\pi(j))$$

$$=\; q(j,i)\;\min(\pi(i),\;\pi(j)) \;=\; \pi(j)\,P(j,i)\,.$$

– This (obviously) still holds if $i = j$.

– It follows that chain is "(time) reversible", i.e. $\pi(i)\,P(i,j) \;=\; \pi(j)\,P(j,i)\;\forall i, j \in \mathcal{X}$.

– (Intuition: if $X_0 \sim \pi$, i.e. $\mathbf{P}(X_0 = i) = \pi(i)$ for all $i \in \mathcal{X}$, then $\mathbf{P}(X_0 = i,\; X_1 = j) = \pi(i)\,P(i,j) = \mathbf{P}(X_0 = j,\; X_1 = i)\;\ldots\;)$

- Suppose now that $X_0 \sim \pi$, i.e. that $\mathbf{P}(X_0 = i) = \pi(i)$ for all $i \in \mathcal{X}$.

– Then using reversibility, we have

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

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

i.e. $X_1 \sim \pi$ too!

– So, the Markov chain "preserves" π, i.e. π is a stationary distribution.

– This is true for any Metropolis algorithm!

- It then follows from the Theorem (i.e., "Basic Fact") that as $n \to \infty$, $\mathcal{L}(X_n) \to \pi$, i.e. $\lim_{n\to\infty} P(X_n = i) = \pi(i)$ for all $i \in \mathcal{X}$.

– Also follows that if $\mathbf{E}_\pi(|h|) < \infty$, then $\lim\limits_{n\to\infty} \frac{1}{n}\sum_{i=1}^n h(X_i) = \mathbf{E}_\pi(h) \equiv \int h(x)\,\pi(x)\,dx$. ("LLN")

JUSTIFICATION: 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. $\mathcal{X} = \mathbf{R}^2$, or $\mathcal{X} = (0,\infty) \times (0,1) \times \mathbf{R}^5$, or $\ldots$)

- Let $\pi(x)$ be the target density function.
- Let $q(x, y)$ be the proposal density for y given x. (e.g. $q(x, y) = (2\pi\sigma)^{-d/2} \exp\big(-\sum_{i=1}^{d}(y_i - x_i)^2/2\sigma^2\big)$.)
- Assume again that q is symmetric: $q(x, y) = q(y, x)$.
- Let $\alpha(x, y) = \min[1, \frac{\pi(y)}{\pi(x)}]$ be probability of accepting a proposed move from x to y.
- Let $P(x, S) = \mathbf{P}(X_1 \in S \,|\, X_0 = x)$ be the transition probabilities.
- (Don't use $P(x, y)$ since that is usually 0.)

- Then if $x \notin S$, then

$$P(x, S) \;=\; \mathbf{P}(Y_1 \in S,\; U_1 < A_1 \,|\, X_0 = x)$$

$$= \int_S q(x, y)\; \alpha(x, y)\; dy \;=\; \int_S q(x, y)\; \min[1,\, \pi(y)/\pi(x)]\; dy\,.$$

- Shorthand: write "$P(x, dy)$" for the transition measure, i.e. a quantity whose integral over any subset S is equal to the transition probability to S, meaning that $P(x, S) = \int_{y \in S} P(x, dy)$.
  - Then, for $x \neq y$, $P(x, dy) = q(x, y)\; \min[1, \pi(y)/\pi(x)]\, dy$.
  - Hence, for $x \neq y$,

$$\pi(x)\, dx\, P(x, dy) \;=\; \pi(x)\, dx\; q(x, y)\; \min[1, \pi(y)/\pi(x)]\, dy$$

$$=\; q(x, y)\; \min[\pi(x),\, \pi(y)]\, dy\, dx\,.$$

  - This is symmetric in x and y, i.e. $\pi(x)\, dx\, P(x, dy) = \pi(y)\, dy\, P(y, dx)$ for all $x, y \in \mathcal{X}$.
  - Shorthand: $\Pi(dx)\, P(x, dy) = \Pi(dy)\, P(y, dx)$. ("reversible")

- How does "reversible" help? Just like for discrete chains!

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

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

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

$$= \int_{y \in S} \Pi(dy) \;=\; \int_{y \in S} \pi(y)\, dy \;\equiv\; \Pi(S)\,,$$

  so also $X_1 \sim \Pi$. So, chain "preserves" Π, i.e. Π is stationary distribution.

- And, again, almost always irreducible and aperiodic.

- So, again, the Theorem applies.

EXAMPLES RE WHY DOES MCMC WORK:

- 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 it's Metropolis, and $q(x, y)$ still symmetric.
  - π stationary? Yes since reversible!
  - Irreducible? Yes, since the n-step transitions $P^n(x, dy)$ have positive density whenever $|y - x| < n$.
  - Aperiodic? Yes since if periodic, then if e.g. $\mathcal{X}_1 \cap [0, 1]$ 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 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! (Not even symmetrically zero, for the Metropolis-Hastings algorithm below, e.g. have positive density $3 \to 9$ but not $9 \to 3$.)

END WEEK #7

JUSTIFICATION OF METROPOLIS-HASTINGS:

- Can we modify the above proof to work for Metropolis-Hastings, too?
- For Metropolis, key was that the Markov chain is reversible, i.e. $\pi(x)\, P(x,y) = \pi(y)\, P(y,x)$, i.e. $q(x,y)\, \alpha(x,y)\, \pi(x)$ is symmetric in x and y.
  - If instead $A_n = \frac{\pi(Y_n)\, q(Y_n, X_{n-1})}{\pi(X_{n-1})\, q(X_{n-1}, Y_n)}$, i.e. acceptance prob. $\equiv \alpha(x,y) = \min\left[1, \frac{\pi(y)\, q(y,x)}{\pi(x)\, q(x,y)}\right]$, 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, $\pi(x)\, P(x,y)$ is still symmetric, even if q wasn't.
  - So, still reversible. So, still have stationary distribution Π.
  - So, if irreducible and aperiodic (nearly always true), then can again apply usual Theorem, and again conclude that it converges to Π.
- Conclusion: For the Metropolis-Hastings algorithm, if we replace "$A_n = \pi(Y_n) \,/\, \pi(X_{n-1})$" by $A_n = \frac{\pi(Y_n)\, q(Y_n, X_{n-1})}{\pi(X_{n-1})\, q(X_{n-1}, Y_n)}$, and keep everything else the same (i.e., still accept if $U_n < A_n$, otherwise reject), then it is still valid, and still converges to the correct distribution Π.

JUSTIFICATION FOR VARIABLE-AT-A-TIME:

- The exact same justification works just like for the "regular" (full-dimensional) Metropolis and Metropolis-Hastings algorithms:
  - If we update the variables one-at-a-time (e.g. Metropolis-within-Gibbs, Metropolis-Hastings-within-Gibbs, etc.), then each individual step is still reversible (for the same reason), so π is still stationary.

– So, like any irreducible, aperiodic Markov chain with stationary distribution π, it will eventually converge to π.

JUSTIFICATION OF 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 except the i^{th} one.

– (And $q_i(x, y) = 0$ if $x^{(-i)} \neq y^{(-i)}$.)

– Here $C[x^{(-i)}]$ is the appropriate normalising constant (which depends on $x^{(-i)}$). So, will always have $C[x^{(-i)}] = C[y^{(-i)}]$.

– 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 (i.e., can ignore the accept-reject step).

INITIAL DISTRIBUTION CONDITION:

- Why does the above Theorem say "π-a.e." $X_0 = x$?
- Example: $\mathcal{X} = \{1, 2, 3, 4, \ldots\}$, and $P(1, \{1\}) = 1$, and for $x \geq 2$, $P(x, \{x+1\}) = 1 - (1/x^2)$ and $P(x, \{1\}) = 1/x^2$.
  – Stationary distribution?
  $\Pi(\cdot) = \delta_1(\cdot)$, i.e. $\Pi(S) = \mathbf{1}_{1\in S}$ for $S \subseteq \mathcal{X}$.
  – Irreducible?
  Yes, 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?
  Yes, since $P(1, \{1\}) > 0$.
  – Converges?
  Yes, by Theorem, for π-a.e. X_0, have $\lim_{n\to\infty} \mathbf{P}(X_n \in S) = \Pi(S)$, i.e. $\lim_{n\to\infty} \mathbf{P}(X_n = 1) = 1$.
  – From everywhere?
  No! If $X_0 = x \geq 2$, then $\mathbf{P}[X_n = x+n \text{ for all } n] = \prod_{j=x}^{\infty}(1-(1/j^2)) > 0$ (since $\sum_{j=x}^{\infty}(1/j^2) < \infty$), 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 usually holds from any $x \in \mathcal{X}$. ("Harris recurrent", see e.g. http://probability.ca/jeff/ftpdir/harris.pdf)

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"). (e.g. $\Pi_\tau = \frac{1}{2}\,N(0,\tau^2) + \frac{1}{2}\,N(20,\tau^2)$; file "Rtempered")
- In the end, only "count" those samples where $\tau = 1$.
- Proceed by defining a joint Markov chain (x, τ) 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 $\frac{1}{m}$.)
- The Markov chain should have 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 $\overline{\Pi}$.
- Then, as above, only "count" those samples where $\tau = 1$. (red)
- 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 and temp=1; 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 better for larger τ! (file "Rtempered" with dotempering=FALSE and temp=1,2,3,4,...,10)
  - (Compare graphs of π_1 and π_8: plot commands at bottom of "Rtempered" ...)
  - So, use above "(a)–(b)" algorithm; converges fairly 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 $\mathbf{E}_\pi(h)$.

FINDING THE TEMPERED DENSITIES:

- 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 π_τ flatter 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:

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

  - But this depends on the c_τ, which are usually unknown – bad.
  - e.g. in above example, could try $\pi_\tau(x) = \left(\frac{1}{2}N(0,1;x)+\frac{1}{2}N(20,1;x)\right)^{1/\tau}$, but don't know normalising constants c_τ so not valid! ("Rtempered2")
- What to do?

PARALLEL TEMPERING:

- (a.k.a. replica exchange: Swendsen and Wang, 1986)
- (a.k.a. Metropolis-Coupled MCMC, or MCMCMC: Geyer, 1991)
- Alternative to tempered MCMC.
- Again have a sequence $\Pi_1, \Pi_2, \ldots, \Pi_m$ where $\Pi_1 = \Pi$ ("cold"), and Π_τ is flatter for larger τ ("hot").
  - e.g. $\pi_\tau(x) = c_\tau\,(\pi(x))^{1/\tau}$, where τ ranges over $\tau_1 = 1, \tau_2, \tau_3, \ldots, \tau_m$.
- Use state space $\mathcal{X}^m$, with m chains, i.e. one chain for each temperature.
  - So, state at time n is $X_n = (X_{n1}, X_{n2}, \ldots, X_{nm})$, 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 $X_{n-1,\tau}$ at temperature τ (for each $1 \le \tau \le m$), by proposing e.g. $Y_{n,\tau} \sim N(X_{n-1,\tau},\ 1)$, and accepting with probability $\min\left(1,\ \frac{\pi_\tau(Y_{n,\tau})}{\pi_\tau(X_{n-1,\tau})}\right)$.
- Or, 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 $\pi_\tau(x) = c_\tau\,(\pi(x))^{1/\tau}$, 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.

- Hence, can set $g_\tau(x) = \pi(x)^{1/\tau}$, no problem.

- EXAMPLE: again $\Pi = \frac{1}{2}\, N(0,1) + \frac{1}{2}\, N(20,1)$.
  - Now can set $g_\tau(x) = \pi(x)^{1/\tau}$, and ignore c_τ.
  - Then run parallel tempering ... works pretty well. (file "Rpara")

END WEEK #8

MONTE CARLO OPTIMISATION – Simulated Annealing:

- General method to find highest mode of π.
- Idea: mode of π is same as mode of a flatter or a more peaked version π_τ, for any $\tau > 0$.
  - e.g. $\pi_\tau \equiv \pi^{1/\tau}$. Flatter if $\tau > 1$, more peaked if $\tau < 1$. ("tempered")
  - 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 concentrated 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 $\tau_M = \tau_0 - dM \geq 0$).
  - or logarithmic ($\tau_n = \tau_0 / \log(1+n)$).
  - or ...
  - Theorem:: if $c \geq \sup \pi$, then simulated annealing with $\tau_n = c/\log(1+n)$ will converge to the global maximum as $n \to \infty$. (But very slow.)
- 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 geometric and linear (better?) cooling ... (file "Rsimann")

- EXAMPLE with real density powers:
  - Set $\pi_\tau(x) = \Big(0.3\,N(0,1) + 0.7\,N(20,1)\Big)^{1/\tau}$.
  - Need longer run, and smaller final τ.
  - Then it works pretty well. (file "Rsimann2")

DIGRESSION – CODE BREAKING:

- Try it out: "decipherdemo". [uses file "decipher.c"]
- 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 (for now) of $\mathcal{A}$, i.e. one-to-one onto mappings $f : \mathcal{A} \to \mathcal{A}$, subject to $f(\text{space}) = \text{space}$.
  - ["substitution cipher"]
- Use a reference text (e.g. "War and Peace") to get matrix $M(x, y) = 1+$ 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 the choice of f which maximises $\pi(f)$.
- 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 $\min\Big(1, \frac{\pi(g)}{\pi(f)}\Big)$.
- Easily seen to be an irreducible, aperiodic, reversible Markov chain.
- So, converges (quickly!) to correct answer, breaking the code.
- References: S. Conner (2003), "Simulation and solving substitution codes". P. Diaconis (2008), "The Markov Chain Monte Carlo Revolution".
- We later extended this, to transposition ciphers and more: J. Chen and J.S. Rosenthal (2010), "Decrypting Classical Cipher Text Using Markov Chain Monte Carlo" (*Statistics and Computing* **22(2)**, 397–413, 2011).

DIGRESSION – PATTERN DETECTION:

- Data is an image, given in terms of a grid of pixels (each "on" or "off").
- Want to "find" the face in the image.

  - (Harder for computers than for humans!)
- Java applet: faces.html (See before and after images.)
- 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 the 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 file "faces.java", or the paper J.S. Rosenthal, Optimising Monte Carlo Search Strategies for Automated Pattern Detection. F. E. J. Math. Sci. 2009.
- Here, again, we want to maximise (i.e., optimise) π, not sample from π.

MCMC CONVERGENCE RATES THEORY:

- $\{X_n\}$: Markov chain on $\mathcal{X}$, with stationary distribution $\Pi(\cdot)$.
- Let $P^n(x, S) = \mathbf{P}[X_n \in S \,|\, X_0 = x]$ be the probabilties for the Markov chain after n steps, when started at 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. $x \in \mathcal{X}$.
- Theorem (mentioned earlier): if chain is irreducible and aperiodic and $\Pi(\cdot)$ stationary, then chain is ergodic, i.e. converges asymptotically to Π.
- 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 $D(x, n) \le M(x)\, \rho^n$ for all $x \in \mathcal{X}$ and $n \in \mathbf{N}$.
- Fact (mentioned earlier): CLT holds for $\frac{1}{n} \sum_{i=1}^n h(X_i)$ 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.)
  - If CLT holds, then (as before) have 95% confidence interval $(e - 1.96\sqrt{v},\ e + 1.96\sqrt{v})$, where $v \approx \frac{1}{M-B} \text{Var}_\pi(h)$(varfact).

- DEFN: chain is uniformly ergodic if there is $\rho < 1$, and $M < \infty$ such that $D(x, n) \leq M\,\rho^n$ for all $x \in \mathcal{X}$ and $n \in \mathbf{N}$.
- DEFN: a quantitative bound on convergence is an actual number n^* such that $D(x, n^*) < 0.01$ (say). [Then sometimes say chain "converges in n^* iterations".]
- Quantitative bounds often difficult (though I've worked on them a lot, see e.g. Rosenthal, "Quantitative convergence rates of Markov chains: A simple account", Elec Comm Prob 2002 and the references therein), but "geometric ergodicity" is often easier to verify.
- But what about convergence rates, i.e. bounds on $D(x, n)$?
- Useful Theorem: If $P(x, dy) \geq \delta\,\pi(dy)$ for all $x, y \in \mathcal{X}$, then $D(x, n) \leq (1 - \delta)^n$. ("minorisation condition" ... proof by "coupling" ...)
  - If true, shows uniform ergodicity and can give a quantitative bound.
- Special Case: INDEPENDENCE SAMPLER (mentioned earlier):
  - Proposals $\{Y_n\}$ i.i.d. from some fixed distribution (say, $Y_n \sim MVN(0,\, I)$).
  - Another special case of Metropolis-Hastings algorithm, where $q(x, y) = q(y)$ depends only on y. So, Π is a stationary distribution.
  - By above Theorem, independence sampler is ergodic provided $q(x) > 0$ whenever $\pi(x) > 0$ (since then it must be irreducible and aperiodic).
  - But does that guarantee that it will work well?
  - No, e.g. previous "Rind" example with $k = 5$: ergodic (of course), but performs terribly.
  - FACT: independence sampler is geometrically ergodic IF AND ONLY IF there is $\delta > 0$ such that $q(x) \geq \delta\pi(x)$ for π-a.e. $x \in \mathcal{X}$.
  - If so, then we compute that

$$p(x, y) \;\geq\; q(y)\,\min[1,\, \frac{\pi(y)\,q(x)}{\pi(x)\,q(y)}] \;=\; \min[q(y),\, \pi(y)\,(q(x)\,/\,\pi(x))]$$

$$\geq\; \min[\delta\,\pi(y),\; \delta\,\pi(y)] \;=\; \delta\,\pi(y)\,.$$

    Hence, from the above, $D(x, n) \leq (1 - \delta)^n$ for π-a.e. $x \in \mathcal{X}$.
- PREVIOUS EXAMPLE: Independence sampler with $\pi(x) = e^{-x}$ and $q(x) = ke^{-kx}$ for $x > 0$. (file "Rind")
  - If $0 < k \leq 1$, then setting $\delta = k$, we have that $q(x) = ke^{-kx} \geq ke^{-x} = k\pi(x) = \delta\pi(x)$ for all $x > 0$, so it's geometrically ergodic, and furthermore $D(x, n) \leq (1 - k)^n$.
  - 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 cannot find any $\delta > 0$ such that $q(x) \geq \delta\pi(x)$ for all x, so it is not geometrically ergodic.
- If $k > 2$, then no CLT (Roberts, J. Appl. Prob. **36**, 1210–1217, 1999).
- So, if $k = 5$ (as in "Rind"), then not geometrically ergodic, and CLT does not hold. Indeed, confidence intervals often miss 1. (file "Rind2")
- 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 2011.]

- What about other MCMC algorithms (besides independence sampler)?
- FACT: if state space is finite, and chain is irreducible and aperiodic, then always ergodic (of course) and also geometrically ergodic. (See e.g. J.S. Rosenthal, SIAM Review 37:387-405, 1995.)
- What about for the "random-walk Metropolis algorithm" (RWM), i.e. where $\{Y_n - X_{n-1}\} \sim q$ (i.i.d.) 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 exponentially light 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]

END WEEK #9

- [11 STUDENT PRESENTATIONS!]
- EXAMPLES: 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. $\sigma = 16$)
- Geometrically ergodic? Yes! (By above.)
- Does CLT hold? Yes! (geometrically ergodic, and $\mathbf{E}_\pi(|h|^p) < \infty$ for all p.)
- Indeed, confidence intervals "usually" contain 41. (file "Rnorm2")

→ CASE #2: $\pi(y) = c\,\frac{1}{(1+y^4)}$, and functional $h(y) = y^2$, so

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

- Not exponentially light tails, so not geometrically ergodic; estimates less stable, confidence intervals often miss 1. (file "Rheavy")

→ CASE #3: $\pi(y) = \frac{1}{\pi(1+y^2)}$ (Cauchy), and functional $h(y) = \mathbf{1}_{-10<y<10}$.

- Recall that for Cauchy, $\Pi(0 < X < y) = \arctan(y)/\pi$.
- So, $\mathbf{E}_\pi(h) = \Pi(|X| < 10) = 2\arctan(10)/\pi = 0.93655$.
- Again, not exponentially light tails, so not geometrically ergodic.
- Confidence intervals often miss 0.93655. (file "Rcauchy")

→ CASE #4: $\pi(y) = \frac{1}{\pi(1+y^2)}$ (Cauchy), and functional $h(y) = \min(y^2, 100)$.

- [Numerical integration: $\mathbf{E}_\pi(h) \doteq 11.77$]
- Once again, not exponentially light tails, so not geometrically ergodic.
- And, 95% CI often miss 11.77, though iid MC does better. ("Rcauchy2")

- NOTE: Even when CLT holds, it can be rather unstable, e.g. it requires that chain has converged to Π, so it might underestimate v.
  - Estimate of v is very important! And "varfact" is not always reliable!
  - Repeated runs?
  - Another approach is "batch means", whereby chain is broken into m large "batches", which are assumed to be approximately i.i.d.

OPTIMAL RWM PROPOSAL SHAPE:

- 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", with file "Rtarg20"):
  - $Y_n \sim MVN(X_{n-1}, 0.025\, I_d)$. ($acc \approx 0.234$; $varfact \approx 400$ 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.
  - Indeed, best is when proposal covariance $= ((2.38)^2/d)\Sigma_0$.
  - 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).
  - And it's "approximately" true for most unimodal targets ...
- Problem: Σ_0 would usually be unknown; then what?
  - Can perhaps "adapt"!

ADAPTIVE MCMC:

- Recall: RWM optimal proposal covariance is $((2.38)^2/d)\Sigma_0$.
- What if target covariance Σ_0 is unknown??
- Can estimate Σ_0 based on run so far, to get empirical covariance Σ_n.
- Then update proposal covariance "on the fly".
- "Learn as you go": see e.g. the Javascript simulation.
- For Adaptive MCMC, could use proposal $Y_n \sim MVN(X_{n-1}, ((2.38)^2/d)\Sigma_n)$.
  - 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. $\epsilon = 0.05$.)
- 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 $s_n = d(d/d^2) \equiv 1$.] See Roberts and Rosenthal, Examples of Adaptive MCMC, JCGS 2009.
- Higher dimensions: figure "RplotAMx200.png" (dim=200). (beautiful!)
  - Works well, but it takes many iterations before the adaption is helpful.

END WEEK #10