Simulation

Richard Lockhart

Simon Fraser University

STAT 870 — Summer 2011

Purposes of Today's Lecture

- Define Continuous Time Markov Chain.
- Prove Chapman-Kolmogorov equations.
- Deduce Kolmogorov's forward and backward equations.
- Review properties.
- Define skeleton chain.
- Discuss ergodic theorem.

Simulation

- Method of distribution theory.
- Given some random variables $X_1, \ldots, X_n$.
- Joint distribution is specified.
- Want distribution of statistic $T(X_1, \ldots, X_n)$.
- Compute, eg, $P(T > t)$ for some specific value of t.
- Use limiting relative frequency interpretation of probability $P(T > t)$ is limit of proportion of trials in long sequence in which T occurs.

Pseudo random Uniform numbers

- Details not discussed here;
- Built in to many languages: can generate sequence of Uniform$[0,1]$ rvs.
- Many methods.
- Typically generate repeating sequence with very long period.
- Actually generate discrete uniforms in binary.
- From now on: convert generated rvs with known distribution to generated rvs with desired / unknown distribution.

Monte Carlo

- Use a (pseudo) random number generator to generate a sample $X_1, \ldots, X_n$.
- Calculate statistic getting T_1.
- Generate new sample (independently of first, say).
- Calculate T_2.
- Repeat large number, N, of times.
- Count how many T_k are larger than t.
- If M such T_k exist estimate $P(T > t) = M/N$.
- M has Binomial$(N, p = P(T > t))$ distribution.

Error of computation

- Standard error of M/N is then $\sqrt{p(1-p)/N}$ which is estimated by $\sqrt{M(N-M)/N^3}$.
- Permits us to guess accuracy of our study.
- Notice standard deviation of M/N is

$$\sqrt{p(1-p)}/\sqrt{N}.$$

- To improve accuracy by factor of 2 requires 4 times as many samples.
- So Monte Carlo time consuming method.
- Tricks available to increase accuracy.
- They only change constant of proportionality; SE still inversely proportional to square root of sample size).

Generating the Sample

- Start from Uniform generator: gives

$$U \sim \text{Uniform}[0,1].$$

- Other distributions generated by transformation:
- **Exponential**: $X = -\log U$ is exponential:

$$\begin{aligned} P(X > x) &= P(-\log(U) > x) \\ &= P(U \le e^{-x}) = e^{-x} \end{aligned}$$

- Random uniforms generated on computer sometimes have only 6 or 7 digits or so of detail.
- Can make tail of distribution grainy.
- Eg: If U were actually a multiple of 10^{-6} then largest possible value of X is $6\log(10)$.

Using memoryless property

Problem ameliorated by following algorithm:

- Generate U a Uniform[0,1] variable.
- Pick a small ϵ like 10^{-3} say. If $U > \epsilon$ take $Y = -\log(U)$.
- If $U \le \epsilon$ remember conditional distribution of $Y - y$ given $Y > y$ is exponential.
- Generate new U'; compute $Y' = -\log(U')$.
- Take $Y = Y' - \log(\epsilon)$.
- Exercise: resulting Y has an exponential distribution; compute $P(Y > y)$.

Generating Normals

- **Normal**: Via inverse probability integral transformation.
- If F is a continuous cdf and U is Uniform[0,1] then $Y = F^{-1}(U)$ has cdf F because

$$\begin{aligned} P(Y \le y) &= P(F^{-1}(U) \le y) \\ &= P(U \le F(y)) = F(y) \end{aligned}$$

- Almost technique used above for exponential distribution.
- For normal distribution $F = \Phi$ (Φ is standard normal cdf) there is no closed form for F^{-1}.
- Could use numerical algorithm to compute F^{-1}

Alternative: Box Müller

- Generate U_1, U_2 two independent Uniform[0,1] variables.
- Define

$$Y_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$

and

$$Y_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2).$$

- Check using change of variables formula that Y_1 and Y_2 are independent $N(0,1)$ variables.

Acceptance Rejection

- Suppose you can't compute F^{-1} but know f.
- Find density g and constant c such that

$$f(x) \leq cg(x)$$

  for each x AND G^{-1} is computable OR can generate observations $W_1, W_2, \ldots$ independently from g.
- Generate W_1.
- Compute $p = f(W_1)/(cg(W_1)) \leq 1$.
- Generate uniform U_1 independent of all Ws.
- Let $Y = W_1$ if $U_1 \leq p$.
- Otherwise get new W and new U and repeat until $U_i \leq f(W_i)/(cg(W_i))$.
- Take Y as last W generated; Y has density f.

Markov Chain Monte Carlo

- Popular for Bayes, for multivariate simulation.
- Suppose $W_1, W_2, \ldots$ (ergodic) Markov chain with stationary transitions.
- Suppose stationary initial distribution of W has density f.
- Then get random variables which have marginal density f by starting off the Markov chain and letting it run for a long time.
- Marginal distribution of W_i converges to f.
- So you can estimate things like $\int_A f(x)dx$ by computing the fraction of the W_i which land in A.
- Uses ergodic theorem.

Other versions of MCMC

- Now many versions of technique including
- Gibbs Sampling
- Metropolis-Hastings algorithm.
- Metropolis Hastings invented in 1950s by physicists: Metropolis et al.
- One authors of paper was Edward Teller "father of the hydrogen bomb".
- Hastings was a student of Don Fraser at Toronto; had career at U Vic.

Importance Sampling

- Want to compute

$$\theta \equiv E(T(X)) = \int T(x)f(x)dx.$$

- Can generate observations from different density g.
- Then compute

$$\hat{\theta} = n^{-1} \sum T(X_i)f(X_i)/g(X_i)$$

- Then

$$\begin{aligned} E(\hat{\theta}) &= n^{-1} \sum E(T(X_i)f(X_i)/g(X_i) \\ &= \int [T(x)f(x)/g(x)]g(x)dx \\ &= \int T(x)f(x)dx \\ &= \theta \end{aligned}$$

Variance reduction

- Example problem: estimate distribution of sample mean for a Cauchy random sample.
- The Cauchy density is

$$f(x) = \frac{1}{\pi(1+x^2)}$$

- Generate $U_1, \ldots, U_n$ uniforms.
- Cauchy: $X_i = \tan\{\pi(U_i - 1/2)\}$.
- Compute $T = \bar{X}$.
- To estimate $p = P(T > t)$ would use

$$\hat{p} = \sum_{i=1}^{N} 1(T_i > t)/N$$

  after generating N samples of size n.
- Estimate is unbiased with standard error

$$\sqrt{p(1-p)/N}.$$

Antithetic Variables

- Can improve this estimate by remembering that $-X_i$ also has Cauchy distribution.
- Take $S_i = -T_i$.
- Remember S_i has same distribution as T_i.
- Try (for $t > 0$)

$$\tilde{p} = [\sum_{i=1}^{N} 1(T_i > t) + \sum_{i=1}^{N} 1(S_i > t)]/(2N)$$

which is average of two estimates like $\hat{p}$.

- Variance of $\tilde{p}$ is

$$(4N)^{-1}\mathrm{Var}(1(T_i > t) + 1(S_i > t)) = (4N)^{-1}\mathrm{Var}(1(|T| > t))$$

which is

$$\frac{2p(1-2p)}{4N} = \frac{p(1-2p)}{2N}$$

Variance Reduction

- Notice variance has extra 2 in denominator.
- Notice numerator is also smaller – particularly for p near $1/2$.
- Variance reduction has resulted in need for smaller sample size to get same accuracy.
- Jargon: antithetic variables.

Regression estimates

- Want to compute
$$\theta = E(|Z|)$$
where Z is standard normal.
- Generate N iid $N(0,1)$ variables $Z_1, \ldots, Z_N$.
- Compute $\hat\theta = \sum |Z_i|/N$.
- But we know that $E(Z_i^2) = 1$.
- Also $\hat\theta$ is positively correlated with $\sum Z_i^2/N$.
- So consider using
$$\tilde\theta = \hat\theta - c(\sum Z_i^2/N - 1)$$

Regression estimation continued

- Notice that $E(\tilde\theta) = \theta$ and

$$\mathrm{Var}(\tilde\theta) = \mathrm{Var}(\hat\theta) - 2c\mathrm{Cov}(\hat\theta, \sum Z_i^2/N) + c^2\mathrm{Var}(\sum Z_i^2/N)$$

- The value of c which minimizes this is

$$c = \frac{Cov(\hat\theta, \sum Z_i^2/N)}{Var(\sum Z_i^2/N)}$$

- This value can be estimated by regressing $|Z_i|$ on Z_i^2!
- Reduces variability by factor of $\sqrt{1-\rho^2}$:

$$\rho = \mathrm{Corr}(|Z_i|, Z_i^2).$$