Chapter 10

- 10.1 Numerical integration (overview)
- 10.2 Distributional approximations (overview, more in Chapter 4 and 13)
- 10.3 Direct simulation and rejection sampling (overview)
- 10.4 Importance sampling (used in PSIS-LOO discussed later)
- 10.5 How many simulation draws are needed? (Ex 10.1 and 10.2)
- 10.6 Software (can be skipped)
- 10.7 Debugging (can be skipped)

Notation

- $p(\theta)$ vs. $p(\theta|y)$
- unnormalized $q(\cdot)$
- proposal $g(\cdot)$

Computational issues

- Log densities
  - use log densities to avoid over- and underflows in floating point presentation
  - compute exp as late as possible
  - e.g. in Metropolis-algorithm compute the log of ratio of densities using the identity $log(a/b) = log(a) - log(b)$

Monte Carlo - history

- Used already before computers
  - Buffon (18th century; needles)
  - De Forest, Darwin, Galton (19th century)
  - Pearson (19th century; roulette)
  - Gosset (Student, 1908; hat)

Monte Carlo - history

- Used already before computers
  - Buffon (18th century; needles)
  - De Forest, Darwin, Galton (19th century)
  - Pearson (19th century; roulette)
  - Gosset (Student, 1908; hat)
- "Monte Carlo method" term was proposed by Metropolis, von Neumann or Ulam in the end of 1940s
  - they worked together in atomic bomb project
  - Metropolis and Ulam, "The Monte Carlo Method", 1949

Monte Carlo - history

- Used already before computers
  - Buffon (18th century; needles)
  - De Forest, Darwin, Galton (19th century)
  - Pearson (19th century; roulette)
  - Gosset (Student, 1908; hat)
- "Monte Carlo method" term was proposed by Metropolis, von Neumann or Ulam in the end of 1940s
  - they worked together in atomic bomb project
  - Metropolis and Ulam, "The Monte Carlo Method", 1949
- Bayesians started to have enough cheap computation time in 1990s
  - BUGS project started 1989
  - Gelfand & Smith, 1990

Monte Carlo

- Simulate samples from the target distribution
  - these samples can be treated as any observations
- Use these samples, for example,
  - to compute means, deviations, quantiles
  - to draw histograms
  - to marginalize
  - etc.

Monte Carlo vs. deterministic

- Monte Carlo = simulation methods
  - evaluation points are selected stochastically (randomly)
- Deterministic methods (e.g. grid)
  - evaluation points are selected by some deterministic rule

How many simulation samples are needed?

- If samples are independent
  - usual methods to estimate the uncertainty due to a finite number of observations
- Markov chain Monte Carlo produces dependent samples
  - requires additional work to estimate the effective number of samples

How many simulation samples are needed?

- Expectation of unknown quantity

$$\mathrm{E}(\theta) \approx \frac{1}{L} \sum_{l} \theta^{(l)}$$

  if L is big and $\theta^{(l)}$ are independent, way may assume that the distrbution of the expectation approaches normal distribution (see Ch 4) with variance σ_θ^2/L (asymptotic normality)

  - this variance is independent on dimensionality of θ

How many simulation samples are needed?

- Expectation of unknown quantity

$$\mathrm{E}(\theta) \approx \frac{1}{L}\sum_{l} \theta^{(l)}$$

if L is big and $\theta^{(l)}$ are independent, way may assume that the distrbution of the expectation approaches normal distribution (see Ch 4) with variance σ_θ^2/L (asymptotic normality)

  - this variance is independent on dimensionality of θ
  - total variance is sum of the epistemic unecrtainty in the posterior and the uncertainty due to using finite number of Monte Carlo samples

$$\sigma_\theta^2 + \sigma_\theta^2/L$$

How many simulation samples are needed?

- Expectation of unknown quantity

$$E(\theta) \approx \frac{1}{L} \sum_{l} \theta^{(l)}$$

if L is big and $\theta^{(l)}$ are independent, way may assume that the distrbution of the expectation approaches normal distribution (see Ch 4) with variance σ_θ^2/L (asymptotic normality)

  - this variance is independent on dimensionality of θ
  - total variance is sum of the epistemic unecrtainty in the posterior and the uncertainty due to using finite number of Monte Carlo samples

$$\sigma_\theta^2 + \sigma_\theta^2/L = \sigma_\theta^2(1 + 1/L)$$

How many simulation samples are needed?

- Expectation of unknown quantity

$$\mathrm{E}(\theta) \approx \frac{1}{L}\sum_{l} \theta^{(l)}$$

if L is big and $\theta^{(l)}$ are independent, way may assume that the distrbution of the expectation approaches normal distribution (see Ch 4) with variance σ_θ^2/L (asymptotic normality)

  - this variance is independent on dimensionality of θ
  - total variance is sum of the epistemic unecrtainty in the posterior and the uncertainty due to using finite number of Monte Carlo samples

$$\sigma_\theta^2 + \sigma_\theta^2/L = \sigma_\theta^2(1 + 1/L)$$

  - e.g. if $L = 100$, deviation increases by $\sqrt{1 + 1/L} = 1.005$ ie. Monte Carlo error is very small (for the expectation)

How many simulation samples are needed?

- Expectation of unknown quantity

$$\mathrm{E}(\theta) \approx \frac{1}{L} \sum_{l} \theta^{(l)}$$

if L is big and $\theta^{(l)}$ are independent, way may assume that the distrbution of the expectation approaches normal distribution (see Ch 4) with variance σ_θ^2/L (asymptotic normality)

  - this variance is independent on dimensionality of θ
  - total variance is sum of the epistemic unecrtainty in the posterior and the uncertainty due to using finite number of Monte Carlo samples

$$\sigma_\theta^2 + \sigma_\theta^2/L = \sigma_\theta^2(1 + 1/L)$$

  - e.g. if $L = 100$, deviation increases by $\sqrt{1 + 1/L} = 1.005$ ie. Monte Carlo error is very small (for the expectation)
  - See Ch 4 for counter-examples for asymptotic normality

How many simulation samples are needed?

- Posterior probability

$$p(\theta \in A) \approx \frac{1}{L} \sum_{l} I(\theta^{(l)} \in A)$$

where $I(\theta^{(l)} \in A) = 1$ jos $\theta^{(l)} \in A$

  - $I(\cdot)$ is binomially distributed as $p(\theta \in A)$

    $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)$ (Appendix A, p. 579)

    $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/L}$

How many simulation samples are needed?

- Posterior probability

$$p(\theta \in A) \approx \frac{1}{L} \sum_l I(\theta^{(l)} \in A)$$

where $I(\theta^{(l)} \in A) = 1$ jos $\theta^{(l)} \in A$

  - $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
    - $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)$ (Appendix A, p. 579)
    - $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/L}$
  - if $L = 100$ and $p \approx 0.5$, $\sqrt{p(1-p)/L} = 0.05$
    ie. accuracy is about 5% units

How many simulation samples are needed?

- Posterior probability

$$p(\theta \in A) \approx \frac{1}{L}\sum_{l} I(\theta^{(l)} \in A)$$

where $I(\theta^{(l)} \in A) = 1$ jos $\theta^{(l)} \in A$

  - $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
    - $\rightarrow$ $\mathrm{var}(I(\cdot)) = p(1-p)$ (Appendix A, p. 579)
    - $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/L}$
  - if $L = 100$ and $p \approx 0.5$, $\sqrt{p(1-p)/L} = 0.05$
    ie. accuracy is about 5% units
  - $L = 2500$ simulation samples needed for 1% unit accuracy

How many simulation samples are needed?

- Posterior probability

$$p(\theta \in A) \approx \frac{1}{L} \sum_{l} I(\theta^{(l)} \in A)$$

where $I(\theta^{(l)} \in A) = 1$ jos $\theta^{(l)} \in A$

  - $I(\cdot)$ is binomially distributed as $p(\theta \in A)$
    - $\rightarrow$ var$(I(\cdot)) = p(1 - p)$ (Appendix A, p. 579)
    - $\rightarrow$ standard deviation of p is $\sqrt{p(1-p)/L}$
  - if $L = 100$ and $p \approx 0.5$, $\sqrt{p(1-p)/L} = 0.05$
    ie. accuracy is about 5% units
  - $L = 2500$ simulation samples needed for 1% unit accuracy
- To estimate small probabilities, a large number of samples is needed
  - to be able to estimate p, need to get samples with $\theta^{(l)} \in A$, which in expectation requires $L \gg 1/p$

How many simulation samples are needed?

- Less samples needed with
  - deterministic methods
  - marginalization (Rao-Blackwellization)
  - variance reduction methods, such, control variates

Direct simulation

- Produces independent samples
  - efficient methods for standard distributions (see, e.g., appendix A)
  - inverse-cdf
  - factorization

Random number generators

- Good pseudo random number generators are sufficient for Bayesian inference
  - modern software used for statistical analysis have good pseudo RNGs

Direct simulation

- Drwas directly from the posterior distribution
  - Using transformations of unifrom random numbers (eg. appendix A)
  - Factorization of multidimensional distributions (eg. normal distribution with unknown mean and variance)
  - 1–3 dimensional cases discrete grid approximaton
- Problem: restricted to only some models

Direct simulation: Example

- Box-Muller -method:
  If U_1 and U_2 are independent draws from distribution $\mathrm{U}(0, 1)$, and

$$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$
$$X_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  then X_1 and X_2 are independent samples from the distribution $\mathrm{N}(0, 1)$

Direct simulation: Example

- Box-Muller -method:
  If U_1 and U_2 are independent draws from distribution U(0, 1), and

$$X_1 = \sqrt{-2\log(U_1)}\cos(2\pi U_2)$$
$$X_2 = \sqrt{-2\log(U_1)}\sin(2\pi U_2)$$

  then X_1 and X_2 are independent samples from the distribution N(0, 1)
  - not the fastest method due to trigonometric computations
  - for normal distrbution more than ten different methods
  - Matlab uses fast Ziggurat method

Grid sampling

- Generalization of inverse-cdf method
  - discretize the parameter space in a grid and compute the normalization term
  - easy to sample from a discrete distrbution
- Problem: number of grid points required grows exponentially with respect to number of dimensions

Grid sampling and curse of dimensionality

- Example: SAT
  - 10 parameters
  - if we don't know beforehand where the posterior mass is
    - need to choose wide box for the grid
    - need to have enough grid points to get some of them where essential mass is
  - e.g. 1000 grid points per dimension

    $\rightarrow$ 1000^{10} = 1e30 grid points
  - Matlab and basic PC in 2013 can compute density of normal distribution about 60 million times per second

    $\rightarrow$ evaluation in all grid points would take about n. 500 billion years

Indirect sampling

- Rejection sampling
  - draw directly from a proposal distribution, reject some draws, remaining draws are independent draws from the target distribution

Indirect sampling

- Rejection sampling
  - draw directly from a proposal distribution, reject some draws, remaining draws are independent draws from the target distribution
- Importance sampling
  - draw directly from a proposal distribution, weight the draws

Indirect sampling

- Rejection sampling
  - draw directly from a proposal distribution, reject some draws, remaining draws are independent draws from the target distribution
- Importance sampling
  - draw directly from a proposal distribution, weight the draws
- Markov chain Monte Carlo
  - draw directly from a transition distribution forming a Markov chain, draws are dependent draws from the target distribution

Rejection sampling

- Proposal forms envelope over the target distribution

$$\frac{q(\theta|y)}{Mg(\theta)} \leq 1$$

  - selection of good proposal gets very difficult when the number of dimensions increase
- demo10_1.m: Rejection sampling

Rejection sampling

- The number of accepted samples is the effective sample size
    - with bad proposal distribution may require a lot of trials

Importance sampling

- Proposal does not need to have a higher value everywhere

$$E[f(\theta)] \approx \frac{\sum_s w_s f(\theta^{(s)})}{\sum_s w_s}, \qquad \text{where} \quad w_s = \frac{q(\theta^{(s)})}{g(\theta^{(s)})}$$

  - selection of good proposal gets more difficult when the number of dimensions increase
  - often used to correct distributional approximations
- demo10_2.m: Importance sampling

- Variation of the weights affect the effective sample size
  - if single weight dominates, we have effectively one sample
  - if weights are equal, we have effectively S samples
- Central limit theorem holds only if variance of the weight distribution is finite

Importance sampling leave-one-out cross-validation

- Later in the course you will learn how $p(\theta|y)$ can be used as a proposal distribution for $p(\theta|y_{-i})$
  - which allows fast computation of

$$p(y_i|y_{-i}) = \int p(y_i|\theta)p(\theta|y_{-i})d\theta$$