Calculating Expectations by Monte Carlo

- Parameter $\theta = (\theta_1, \theta_2, \ldots, \theta_N)$
  - $\theta_i$’s may be discrete, continuous, or mixed.
  - Dimension of the problem, $N$, is large.
  - Only $q(\theta)$, an unnormalized density/probability mass function of $\theta$, is known.
  - Many features of this distribution may be of interest:
    * $E\{\theta_1 \exp(\theta_2)\}$
    * $\Pr(\theta_1 > 0.5)$
    * quantiles of $(\sum_{i=1}^{N} \theta_i)/N$. 
• Alternative approaches:
  – approximation:
    E.g., use a mixture of normal densities to approximate $q(\theta)$
  – complete enumeration (discrete case):
    a feat to charter all configurations of a $5 \times 5 \times 5$ Ising model
  – numerical integration
• Monte Carlo methods:
  – Draw $m$ samples $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(m)}$ from the target density $q(\theta)$
  – Use empirical averages to estimate expectations:
    $$E(\log(\theta)) \approx \frac{1}{m} \sum_{j=1}^{m} \log(\theta^{(j)}); \quad \Pr(\theta > 0.5) \approx \frac{1}{m} \#\{\theta^{(j)} > 0.5\}$$
How to make the Monte Carlo draws

- Rejection sampling
- Importance sampling
- Markov chain Monte Carlo (MCMC)
Rejection Sampling

- $f(\theta)$: unnormalized target density
- $g(\theta)$: an envelop function
  - Must have $g(\theta) \geq f(\theta)$

Algorithm:
(i) Draw $\theta$ according to a density proportional to $g$
(ii) Accept $\theta$ with probability $f(\theta)/g(\theta)$; otherwise go to (i)

Properties:
- Accepted $\theta$’s have the correct distribution (i.e., with density $f(\theta)$)
- Acceptance rate: $\int f(\theta)d\theta / \int g(\theta)d\theta$
• Advantages:
  – Draws are exact (compare with the grid method)
  – Draws are independent

• Disadvantages:
  – Need to choose \( g \) carefully
    * It should be easy to draw samples according to \( g \)
    * The acceptance rate should not be too small
  – May be difficult if the dimensionality of \( \theta \) is high
Example: Tail of a Normal

- Target density $p(\theta) \propto f(\theta)$:
  - $f(\theta) = \exp(-\theta^2/2)$, $\theta > T$
  - $f(\theta) = 0$ if $\theta \leq T$
  - $T$: truncation bound

- Plan A: draw $\theta \sim N(0,1)$ and accepted if $\theta > T$
  - Acceptance rate: $1 - \Phi(T)$ ($\Phi$ denotes the standard normal cdf)
  - Does not work well if $T$ is large
Plan B: Envelop function \( g(\theta) = \exp(-c\theta + c^2/2), \ \theta > T \)
- \( g(\theta) = 0 \) if \( \theta \leq T \)
- \( c > 0 \): a constant to be determined
- Why is this a legitimate envelop? (Why \( g(\theta) \geq f(\theta) \)?)

Is it easy to sample according to \( g(\theta) \)?
- Shifted exponential: draw \( \gamma \sim \text{Exp}(1) \) and then set \( \theta = T + \gamma/c \)

Any \( c > 0 \) gives a valid sampler

But some \( c \) are more efficient
- Acceptance rate: \( \int f(\theta)d\theta / \int g(\theta)d\theta \)
  - Only the denominator depends on \( c \)
  - To minimize \( \int g(\theta)d\theta \), choose \( c = (T + \sqrt{T^2 + 4})/2 \)
Importance Sampling

• $f(\theta)$: unnormalized target density
• Goal: $E h(\theta) \equiv \int h(\theta)f(\theta)\,d\theta / \int f(\theta)\,d\theta$
• Algorithm:
  – Choose $g(\theta)$ (unnormalized density)
  – Draw $\theta^{(1)}, \ldots, \theta^{(L)}$ according to $g(\theta)$.
  – Estimate $E h(\theta)$ by
    \[
    \frac{\sum_{l=1}^{L} h(\theta^{(l)})w(\theta^{(l)})}{\sum_{l=1}^{L} w(\theta^{(l)})}
    \]
    where $w(\theta) = f(\theta)/g(\theta)$ (importance ratio)
• Why does this work? (law of large numbers)
• Works best if \( g(\theta) \) is approximately proportional to \( h(\theta)f(\theta) \)
• Not useful if importance ratios vary wildly
  – Usually one wants \( g \) to be more heavy tailed than \( f \)
• Good example
  – Goal: $E(\theta)$
  – target $f(\theta) = 1, \ 0 < \theta < 1$
  – $g(\theta) = \exp(-\theta^2/2), \ -\infty < \theta < \infty$

• BAD example
  – Goal: $E(\theta)$
  – target $f(\theta) = \exp(-\theta^2/2), \ -\infty < \theta < \infty$
  – $g(\theta) = 1, \ 0 < \theta < 1$
• In practice, if $f$ is unimodal $g$ can be chosen as a $t$ density with approximately the same center and scale
  – $t$ rather than the normal for heavier tails
  – Mixtures of $t$ distributions for multimodal $f$
  – Mode-finding algorithms: Newton-Raphson; EM
• Illustration using logistic regression.
MCMC Motivating Example

- Statistical Mechanics
  \[ \Pr(s) = \frac{1}{Z} \exp\left\{-\frac{E(s)}{kT}\right\}, \]
  \( \Pr(s) \): probability that the system is at microstate \( s \)
  \( E(s) \): energy associated with state \( s \)
  \( kT \): temperature (in suitable units) of the environment

- Macro/global characteristics of this system, e.g.,
  phase transition?
Example: 2-D Ising model

$s$: a 2-D array of binary variables

\[ E(s) = -\beta \sum_{i \sim j} s_is_j - \gamma \sum_i s_i \]

\((i \sim j \text{ means } i, j \text{ are adjacent sites in the 2-D array.})\)

```
+  +  -  +  +  -  +  +  +  
+  -  -  -  +  +  +  +  -  
+  -  +  +  +  +  +  +  -  
+  +  +  +  +  +  +  +  -  
-  +  -  +  +  -  +  +  +  
-  -  -  +  +  -  +  +  +  
+  -  -  +  +  +  +  +  +  
+  +  +  +  +  +  +  +  -  
+  +  +  +  +  +  +  +  -  
```
• magnetization: \( M(s) = |\sum_i s_i|/N \)
  Take \( \gamma = 0, n \gg 1: \)
  \[ \beta/(kT) < 0.44 \quad \beta/(kT) > 0.44 \]
  \[ M(s) \approx 0 \quad M(s) > 0 \]

• How do we calculate the magnetization \( M(s) \), averaged over \( 2^N \) possible configurations of the system?
Markov Chain Monte Carlo

Markov chain $\theta^{(1)}, \theta^{(2)}, \ldots, \theta^{(t)} \ldots$

- **Given** $\theta^{(t)}, \theta^{(t-1)}, \ldots, \theta^{(1)}$, the distribution of $\theta^{(t+1)}$ depends only on $\theta^{(t)}$

- **Transition probability:** $T(\theta^{(t+1)}|\theta^{(t)})$
  
  the probability distribution of the next iteration given the current $\theta^{(t)}$

- **Stationary distribution:** $\int q(\theta)T(\theta'|\theta)d\theta = q(\theta')$
  
  The density $q(\theta)$ is preserved through iterations

- **The Ergodic Theorem:** The Markov chain generated by $T(\cdot|\cdot)$ with specified stationary distribution $q(\theta)$ converges if this chain is irreducible and aperiodic.

- **In the long run,** $\theta^{(t)}$ is distributed according to $q(\theta)$

Need to generate a Markov chain with a pre-specified stationary distribution
Gibbs Sampling

At each iteration

- draw $\theta^{(t+1)}$ one coordinate at a time
- $\theta_i^{(t+1)}$ is drawn according to the conditional distribution $q(\theta_i|\theta_{-i})$, i.e., given most current $\theta_j$, $j \neq i$.

2-D Ising model $\beta/(kT) = 0.2$, $\gamma = 0$, number of sites $N = 2 \times 2$.

- Conditional distributions

\[
\begin{align*}
\Pr(s_{11}|s_{12}, s_{21}, s_{22}) &\propto \exp(0.2s_{11}(s_{12} + s_{21})) \\
\Pr(s_{12}|s_{11}, s_{21}, s_{22}) &\propto \exp(0.2s_{12}(s_{11} + s_{22})) \\
\Pr(s_{21}|s_{11}, s_{12}, s_{22}) &\propto \exp(0.2s_{21}(s_{11} + s_{22})) \\
\Pr(s_{22}|s_{11}, s_{12}, s_{21}) &\propto \exp(0.2s_{22}(s_{12} + s_{21}))
\end{align*}
\]
• Let’s do one iteration by hand :)

Starting value:  

\[
\begin{array}{c}
\text{Update } s_{11}: \\
\text{Update } s_{12}: \\
\text{Update } s_{21}: \\
\text{Update } s_{22}: \\
\end{array}
\]

\[
\begin{array}{c}
\text{? } + \\
- + \\
+ ? \\
+ + \\
\end{array}
\]

\[
\begin{array}{c}
(\Pr(s_{11} = +|\cdots) = 0.50) \rightarrow \\
- + \\
(\Pr(s_{12} = +|\cdots) = 0.69) \rightarrow \\
- + \\
(\Pr(s_{21} = +|\cdots) = 0.69) \rightarrow \\
+ + \\
(\Pr(s_{22} = +|\cdots) = 0.69) \rightarrow \\
+ - \\
\end{array}
\]
Normal Hierarchical Model

- Likelihood:
  \[ y_j | (\theta, \mu, \tau) \overset{\text{iid}}{\sim} N(\theta_j, \sigma_j^2), \ j = 1, \ldots, J \]
  - \( y_j \): estimated treatment effect in school \( j \) (see SAT coaching example in Gelman et al 2003)
  - \( \sigma_j \): known standard errors of \( y_j \)

- Priors:
  \[ \theta_j | (\mu, \tau) \overset{\text{i.i.d.}}{\sim} N(\mu, \tau^2), \ j = 1, \ldots, J \]
  - \( \theta_j \): mean treatment effect in school \( j \)

- Hyper priors:
  \[ p(\mu, \tau) \propto 1 \]
Gibbs Sampling

- Conditional distributions:
  - $\theta \mid (\mu, \tau, y)$:
    \[
    \theta_j \mid (\mu, \tau, y) \overset{\text{indep}}{\sim} N \left( \frac{\mu/\tau^2 + y_j/\sigma_j^2}{1/\tau^2 + 1/\sigma_j^2}, \frac{1}{1/\tau^2 + 1/\sigma_j^2} \right)
    \]
  - $\mu \mid (\tau, \theta, y)$:
    \[
    \mu \mid (\tau, \theta, y) \sim N(\bar{\theta}, \tau^2/J), \quad \bar{\theta} = J^{-1} \sum_j \theta_j
    \]
  - $\tau \mid (\mu, \theta, y)$:
    \[
    \tau^2 \mid (\mu, \theta, y) \sim \text{Inv-}\chi^2(\nu, s^2), \quad \nu = J - 1, \quad s^2 = \sum_j (\theta_j - \mu)^2 / \nu
    \]

- Simplification due to conditional independence
- Illustration with R
Why does Gibbs Sampling Work?

The target density is preserved by each step of conditional draws:

- Suppose \((\theta_1^0, \theta_2^0)\) is drawn according to the target \(p(\theta_1, \theta_2)\)
- Then \(\theta_2^0\) has the correct marginal density \(p(\theta_2)\)
- At iteration 1, we draw \(\theta_1^1|\theta_2^0\) according to the conditional density \(p(\theta_1|\theta_2)\)
- Hence the pair \((\theta_1^1, \theta_2^0)\) has the correct joint density \(p(\theta_1, \theta_2)\)
- Similarly for the pairs \((\theta_1^1, \theta_2^1), \ldots, (\theta_t^t, \theta_t^t)\)

By the ergodic theorem, under regularity conditions, the Markov chain converges to \(p(\theta_1, \theta_2)\) no matter where we start
Irreducibility can be an issue

• Example (draw a picture):

\[ p(\theta_1, \theta_2) = 0.5I(0 < \theta_1 < 1, 0 < \theta_2 < 1) + 0.5I(1 < \theta_1 < 2, 1 < \theta_2 < 2) \]

\( I() \): indicator function

• Conditional distributions

– \( \theta_1 | \theta_2 \sim \text{uniform}(0,1) \), if \( 0 < \theta_2 < 1 \)

– \( \theta_2 | \theta_1 \sim \text{uniform}(0,1) \), if \( 0 < \theta_1 < 1 \)

– Starting from the lower left square, the limiting density is uniform on the lower left square

– \( \theta_1 | \theta_2 \sim \text{uniform}(1,2) \), if \( 1 < \theta_2 < 2 \)

– \( \theta_2 | \theta_1 \sim \text{uniform}(1,2) \), if \( 1 < \theta_1 < 2 \)

– Starting from the upper right square, the limiting density is uniform on the upper right square
• Gibbs sampling is convenient if conditional draws are easy to make

• Draws are dependent, hence efficiency varies
  – Illustration with bivariate normal
The Metropolis-Hastings Algorithm

- $q(\theta)$: unnormalized target density
- $J(\theta' | \theta)$: proposal density
- Algorithm: at iteration $t$
  - Generate $\theta'$ according to $J(\cdot | \theta^{(t)})$
  - Accept $\theta'$ (i.e., set $\theta^{(t+1)} = \theta'$) with probability
    \[ r = \min \left\{ 1, \frac{q(\theta') / J(\theta' | \theta^{(t)})}{q(\theta^{(t)}) / J(\theta^{(t)} | \theta')} \right\} \]
    otherwise set $\theta^{(t+1)} = \theta^{(t)}$
Special Cases

- Symmetric $J$: $J(\theta'|\theta) = J(\theta|\theta')$
  - Originally proposed by Metropolis et al. (1953)
  - Example (univariate): $J(\theta'|\theta) = (2\pi\delta^2)^{-1/2} \exp\left(-\frac{(\theta' - \theta)^2}{2\delta^2}\right)$
  - The acceptance rule becomes $r = \min\{1, q(\theta')/q(\theta)\}$
  - Interpretation: Always accept the new $\theta'$ if the density is higher
    sometimes take a step down (move to a point with lower density)
  - Illustration with the SAT coaching example
  - The effect of step-size
• Independent Hastings proposal: $J(\theta'|\theta) = J(\theta')$
  – The acceptance rule becomes
    $$r = \min\left\{1, \frac{q(\theta')/J(\theta')}{q(\theta)/J(\theta)}\right\}$$
  – Ideally $J(\theta) \propto q(\theta)$
  – Can approximate $q(\theta)$ based on the mode $\hat{\theta}$ and the curvature at $\hat{\theta}$
    * As in importance sampling, $J(\theta)$ should have heavier tails than $q(\theta)$
    * $t$ distributions rather than normals
  – Illustration with the SAT coaching example
Assessing Convergence

• Run one chain for long enough
  – Examine the trajectory of the draws
  – Examine autocorrelations
  – Illustration

• Run multiple chains from diffuse starting values until they mix well
  – Examine multiple trajectories on the same plot
  – Calculate the $\hat{R}$ statistic of Gelman and Rubin (1992)
* Run $m$ independent chains with $L$ iterations each from diffuse starting values
* $\theta_{lj}$: $l$th sample in the $j$th chain
* Between-chain variance

$$B \equiv \frac{L}{m-1} \sum_{j=1}^{m} (\bar{\theta}_j - \bar{\theta})^2$$

$$\bar{\theta}_j = \frac{1}{L} \sum_{l=1}^{L} \theta_{lj}; \quad \bar{\theta} = \frac{1}{m} \sum_{j=1}^{m} \bar{\theta}_j$$

* Within-chain variance

$$W \equiv \frac{1}{m} \sum_{j=1}^{m} s_j^2; \quad s_j^2 = \frac{1}{L-1} \sum_{l=1}^{L} (\theta_{lj} - \bar{\theta}_j)^2$$

* Consider the chain converged if $\hat{R}$ is close to one (say < 1.1):

$$\hat{R} \equiv \sqrt{\frac{T}{W}} ; \quad T = \frac{L-1}{L} W + \frac{1}{L} B$$

(As the chain converges, the within-variance $W$ approaches the total variance $T$)
General Advice

• Mode-finding is helpful
  – May suggest proposal densities in Metropolis-Hastings (M-H) schemes
  – Can serve as starting values
  – If there are multiple-modes...

• Run multiple chains

• Burn-in period: helpful to discard beginning iterations if not started from places with high posterior density

• Monitor acceptance rates in M-H schemes and adjust proposal densities in pilot runs
  – Aim at high acceptance rate (e.g., 70%) in independent Hastings schemes
  – Lower acceptance rates (e.g., 30%) are fine for random-walk-type Metropolis

• Pay attention to autocorrelations
  – Effective sample size: 1000000 highly correlated draws may be worth 10 independent ones

• Examine many quantities, not just a few parameters of interest
It Can Be Slow

Possible remedies:

• grouping
• reparameterization
• parameter expansion
• interweaving various parameterizations
• auxiliary variables
• Hamiltonian MCMC
• simulated tempering
• genetic algorithms in parallel chains