Part 2: Markov chain Monte Carlo (MCMC) methods

- Indirect sampling methods
- Markov chains
- MCMC methods - Metropolis-Hastings and Gibbs Sampler
- The multivariate normal model
Introduction

- We have seen that Monte Carlo sampling is a useful tool for sampling from distributions.
- In simple models, especially with conjugate prior distributions, it is often easy to derive the posterior distribution in closed form. MC sampling is convenient (though not really necessary).
- When the posterior density does not have a recognizable form, it might be possible to factor the distribution analytically and simulate in parts, as we did for the Normal model.
- For more complicated problems, it may not possible to directly generate samples from the target distribution (e.g., non-conjugate settings and, in general, situations where we cannot sample from the joint posterior distribution).
- We will see indirect sampling schemes first and then introduce Markov chain Monte Carlo (MCMC) methods.
If we cannot directly sample from the joint posterior then we cannot use a plain Monte Carlo algorithm. We need to search for ways to sample from the joint posterior indirectly. Several methods have been proposed in the literature for indirect sampling. We will discuss two of them.

These are

1. Acceptance-Rejection Sampling (not covered)
2. Importance Sampling
Importance sampling

- Suppose that $f(x)$ can be approximated by some density $g(x)$, from which we can easily sample.
- The importance sampling approach is based on the principle that

\[
E(h(x)) = \mu = \int h(x) f(x) \, dx
\]

\[
= \int h(x) \frac{f(x)}{g(x)} g(x) \, dx
\]

\[
= \frac{\int h(x) \frac{f(x)}{g(x)} g(x) \, dx}{\int \frac{f(x)}{g(x)} g(x) \, dx} = \frac{E_g[h(x)w(x)]}{E_g[w(x)]}
\]

where $g$ is called **importance sampling distribution** or **envelope** and $w(x) = f(x)/g(x)$ is the importance weight function.
Importance sampling - unstandardized weights

The previous equations suggest the following two ways to estimate $E(h(X))$ by means of a MC approach:

1. Draw $X_1, \ldots, X_n \overset{i.i.d.}{\sim} g$ and use the estimator

$$\hat{\mu}_{IS}^{*} = \frac{1}{n} \sum_{i=1}^{n} h(X_i) w^*(X_i),$$

where $w^*(X_i) = f(X_i)/g(X_i)$ are unstandardized weights (importance ratios).

Of course, it must be easy to sample from $g$ and evaluate $f$, even when it is not easy to sample from $f$. 
Importance sampling - standardized weights

2. Otherwise, we can draw $X_1, \ldots, X_n \overset{i.i.d.}{\sim} g$ and compute

$$\hat{\mu}_{IS} = \sum_{i=1}^{n} h(X_i) w(X_i),$$

where $w(X_i) = \frac{w^*(X_i)}{\sum_{i=1}^{n} w^*(X_i)}$ are standardized weights. This approach is particularly appealing when $f$ is known up to a proportionality constant (e.g. posterior in Bayesian inference).
Importance sampling

\[ E(h(x)) \approx \frac{1}{n} \sum_{i=1}^{n} h(X_i)w^*(X_i), \]

- The *importance weighting* allows us to over sample a portion of the domain of the function that receives lower probability under the target distribution.

- Also, the same sample generated from \( g \) can be used repeatedly, not only for different functions \( h(\cdot) \), but also for different densities \( f(x) \), which is quite attractive for robustness and Bayesian sensitivity analysis.
**$f$ and $g$**

(a) the support of $g$ must include all the support of $f$

(b) the ratio $f(x)/g(x)$ must be bounded

(c) $g$ must have heavier tails than $f$

(d) Since

$$E\left(\frac{f(X)}{g(X)}\right) = \int \frac{f(x)}{g(x)} g(x) \, dx = 1$$

and $g(x)$ has heavier tails than $f$, the variance of $\frac{f(X)}{g(X)}$ must be large.

If we desire to use the importance sampling approach to reduce the variance of the MC estimate, then it is better to choose $g(x)$ so that $\frac{f(X)}{g(X)}$ is large when $h(x)$ is small, and viceversa. If $g(X)$ is a poor approximation, many of the weights will be close to zero, and thus few $X_i$’s will dominate the sums, producing an inaccurate approximation.
Effective Sample Size

Effective sample size to measure the efficiency of an importance sampling strategy using an envelope $g$. For the unstardized weights, the ESS $N(f, g)$ is

$$N(f, g) = \frac{n}{1 + \hat{V}\{w^*(x)\}},$$

where $\hat{V}\{w^*(x)\}$ is the sample variance of the $w^*(x_i)$. If the standardized weights are used, then

$$N(f, g) = \frac{n}{1 + \hat{cv}^2\{\tilde{w}(x)\}},$$

where $\hat{cv}\{\tilde{w}(x)\}$ is the sample standard deviation of the standardized importance weights divided by their sample mean. The ESS indicates that $n$ weighted samples used in IS are worth $N(f, g)$ unweighted i.i.d. samples drawn exactly from $f$ and used in a simple Monte Carlo estimate.
Example

Suppose $X \sim \text{Exp}(\lambda)$, $\lambda = 1$. We want to compute

$$E[\sqrt{X}] = \int \sqrt{x} f(x) \, dx$$

We consider three different importance density functions.

```r
set.seed(1)
samp.size=10000
lambda=1

## storage
x <- matrix(NA, samp.size, 3)
wts <- matrix(NA, samp.size, 3)

## Envelopes
# Case 1: abs(Normal(0,1))
x[,1]=abs(rnorm(samp.size))
wts[,1]=dexp(x[,1], rate=lambda)/(dnorm(x[,1])*2)
```
# Case 2: \text{uniform}(0,1000)
\begin{align*}
x[,2] &= \text{runif}(\text{samp.size},0,1000) \\
wts[,2] &= \text{dexp}(x[,2], \text{rate}=\lambda)/(\text{dunif}(x[,2],0,1000))
\end{align*}

# Case 3: \text{abs} (\text{Cauchy}(0,1))
\begin{align*}
x[,3] &= \text{abs} (\text{rcauchy}(\text{samp.size})) \\
wts[,3] &= \text{dexp}(x[,3], \text{rate}=\lambda)/(\text{dcauchy}(x[,3]) \times 2)
\end{align*}

## Importance sampling estimate of \(\mathbb{E}[\sqrt{X}]\)
\begin{align*}
\text{mu.is} &= \text{rep}(\text{NA}, 0) \\
\text{for} (i \in 1:3) \\
\text{mu.is}[i] &= \text{sum}(\sqrt{x[,i]} \times wts[,i]) / \text{sum}(wts[,i])
\end{align*}
Summary so far

- We have seen that Monte Carlo sampling is a useful tool for sampling from distributions.
- For more complicated problems, it may not possible to directly generate samples from the target distribution (e.g., non-conjugate settings and, in general, situations where we cannot sample from the joint posterior distribution).
- We have seen indirect sampling schemes.
- We now introduce Markov chain Monte Carlo (MCMC) methods.
Lecture 2.2: Markov Chains (discrete case - brief review)

A Markov chain is a sequence of random variables $X_1, X_2, X_3, \ldots$ with the Markov property, namely that, given the present state, the future and past states are independent. The possible values of $X_i$ form a countable set $S$ called the state space of the chain.

$$Pr(X_{n+1} = x_j | X_n = x_n, \ldots, X_1 = x_1) = Pr(X_{n+1} = x_j | X_n = x_n).$$

Properties of a Markov Chain:

1. Reducibility
2. Periodicity
3. Recurrence
4. Ergodicity
Markov Chains properties:

- **Reducibility** A state \( j \) is said to be accessible from a state \( i \) if a system started in state \( i \) has a non-zero probability of transitioning into state \( j \) at some point.
  \[
  Pr(X_n = j | X_0 = i) = p_{ij}^{(n)} > 0
  \]

- **Irreducibility** A chain is irreducible if all states communicate (there’s a positive probability to visit all states in a finite number of steps).

- **Periodicity** A state \( i \) has period \( k \) if any return to state \( i \) must occur in multiples of \( k \) time steps. If \( k = 1 \), then the state is said to be aperiodic i.e. returns to state \( i \) can occur at irregular times.
Markov Chains properties:

- **Recurrence** A state $i$ is said to be **transient** if, given that we start in state $i$, there is a non-zero probability that we will never return to $i$. A state $i$ is called **absorbing** if it is impossible to leave this state. A state $i$ is **recurrent** if the expected number of visit to $i$ is equal to infinity.

- **Ergodicity: to what is the chain converging?**

\[
\lim_{n \to \infty} \| P^n - \pi \| = 0
\]

that is the distribution $P^n$ of $X_n$ converges to a target invariant distribution $\pi$ irrespective of the initial conditions.

A chain is said to be ergodic if it is aperiodic and positive recurrent. (positive = the chain has an invariant probability measure)
Stationary distribution

- A chain is **reversible** if and only if \( \pi_j p_{ji} = \pi_i p_{ij} \) for all \( i \neq j \)

\[ p_{ij} = Pr(X_t = j | X_{t-1} = i) \] - transition probability from state \( i \) to state \( j \)

\( \pi_i \) - equilibrium probability of being in state \( i \).

When \( Pr(X_{t-1} = i) = \pi_i \) for all \( i \), condition above is equivalent to saying that the joint probability matrix, \( Pr(X_{t-1} = i, X_t = j) \) is symmetric in \( i \) and \( j \).

This is also known as the **detailed balance condition**.

- Also, \( \pi_j p_{ji} = \pi_i p_{ij} \) implies that \( \pi = \pi P \), since

\[ (\pi P)_j = \sum_i \pi_i p_{ij} = \sum_i \pi_j p_{ji} = \pi_j \sum_i p_{ji} = \pi_j. \]

**Important result:** A Markov chain with transition matrix \( P \) will have an equilibrium distribution \( \pi \) if and only if \( \pi = \pi P \).

- In this way, to sample from the limiting distribution \( \pi \), we run a Markov Chain with transition matrix \( P \) satisfying the detailed balance condition until the chain appears to have settle down to equilibrium.
Monte Carlo methods for Markov Chains

Classical LLN’s and CLT’s not directly applicable due to:

- Markovian dependence structure between the obs $X_i$,
- Non-stationarity of the sequence.

Theorem (Ergodic Theorem)
If the Markov chain $(X_n)$ is Harris recurrent (irreducible, aperiodic and positive recurrent) then for any function $h$ with $E|h| < \infty$,

$$\lim_{n \to \infty} \frac{1}{n} \sum h(X_i) = \int h(x)\pi(x)dx$$

If the chain is also reversible ($X_{n+1}|X_{n+2} = x \sim X_{n+1}|X_n = x$) then a “modified” version of the CLT holds.
The MCMC principle - general idea

- A Markov Chain Monte Carlo (MCMC) method for simulation from a distribution $\pi$ is any method producing an ergodic Markov chain $(X_n)$ whose stationary distribution is $\pi$.

- For an arbitrary starting value $x^{(0)}$, a chain $(X_n)$ is generated using a transition kernel with stationary distribution $\pi$, which ensures the convergence in distribution of $X_n$ to a random variable from $\pi$.

- In order for the Markov chain to converge to the target (stationary or equilibrium) distribution $\pi$, it must be irreducible, aperiodic and positive recurrent.

- The main problem is how to construct a suitable chain; we will see some of the most used algorithms (namely, Metropolis-Hastings and Gibbs sampling) and also look into methods to assess whether the chain has reached its stationary distribution.

- MCMC was invented in 1953 but was not widely applied until statisticians became aware of it around 1990. Then problems that had previously been intractable suddenly became tractable.