Monte-Carlo and Empirical Methods for Statistical Inference, FMS091/MA$^S$M11

L4 — Markov Chain Monte-Carlo

Importance sampling

- Hard to sample from $f(x)$.
- If we want to calculate $E(\varphi(x))$, where $\varphi(x)$ and $f(x)$ are dissimilar.

The basis of importance sampling is to rewrite the integral

$$E_f(\varphi(x)) = \int \varphi(x)f(x) \, dx = \int \varphi(x)\frac{f(x)}{g(x)}g(x) \, dx = E_g \left( \frac{\varphi(x)f(x)}{g(x)} \right) = E_g (\varphi(x))$$

Algorithm:
1. Draw $N$ values $x_1, \ldots, x_N$ from $g$.
2. Approximate $\tau = E(\varphi(X))$ with

$$t_N = \frac{1}{N} \sum_{i=1}^{N} \varphi(x_i)f(x_i),$$

Monte Carlo integration

- Realize that many quantities can be written as expectations.
  - Expectations: $E(X) = \int xf(x) \, dx$
  - Probabilities: $P(X > 2) = E(\mathbb{1}(x > 2))$.
  - Standard integrals:

$$\int_0^{2\pi} \sin(2\pi/(x - \pi)^2) \, dx = 2\pi \mathbb{E}_{(0,2\pi)} \left( \sin(2\pi/(x - \pi)^2) \right).$$

- Use the Law of Large Numbers to approximate the expectation with an average.

Algorithm:
1. Draw $N$ values $x_1, \ldots, x_N$ from $f$.
2. Approximate $\tau = E(\varphi(X))$ with

$$t_N = \frac{1}{N} \sum_{i=1}^{N} \varphi(x_i)f(x_i).$$

Markov Chain Monte Carlo

This lecture and Monday

- Markov chains
- Underlying theory
- The Metropolis Hastings algorithm
- Gibbs sampling
- The slice sampler
- Hybrid methods
- Monte Carlo integration with dependent random numbers
- A worked out example
Markov Chain Monte Carlo

- Basic idea: To sample from a density $f$ we construct a Markov chain with $f$ as its stationary distribution.
- If $f$ is complex and/or high dimensional this is often easier than transformation methods or rejection sampling.
- But, the samples will not be independent.
- MCMC is currently the most common method for sampling from complex and/or high dimensional distributions.

- Dates back to the 1950's with two key papers being:
  - Equations of State Calculations by Fast Computing Machines (1953)

Markov chains: Stationarity

- $X_t$ is stationary if $P(X_t = i)$ is independent of $t$.
- $\pi = \{\pi_i\}$ is said to be a stationary distribution if $\forall i: P(X_0 = i) = \pi_i \Rightarrow P(X_1 = i) = \pi_i$
- $\pi$ is a stationary distribution of the chain if and only if it satisfies the global balance equations $\pi P = \pi$
  with $\sum_{k=1}^{K} \pi_k = 1$, $\pi_j \geq 0$.
- If $\pi$ satisfies the detailed balance equations $\pi_i p_{ij} = \pi_j p_{ji}$, $\forall i$ and $j$,
  then $\pi$ is a stationary distribution, but the converse does not have to be true.
- A Markov chain is reversible if it has a stationary distribution satisfying the detailed balance equations.

Markov chains

Let $X_t$ be a (time discrete) stochastic process that takes values $X_t \in \{1, \ldots, K\}$.

- $X_t$ is a Markov process if $P(X_t|X_{t-1}, \ldots, X_1) = P(X_t|X_{t-1})$.
- The transition probability between two states is $p_{ij} = P(X_t = j|X_{t-1} = i)$.
- The transition matrix $P$ is a matrix with elements $p_{ij}$.

Properties of Markov chains

- The chain is irreducible if we can get from all states to all other states (not necessarily in one jump).
- A state, $k$, is recurrent if $P(X_t = k$ for any $t > 0|X_0 = k) = 1$.
- If the expected time it takes to get back to $k$ is finite, the state is said to be positive recurrent.
- An irreducible Markov chain has a stationary distribution if and only if the states are positive recurrent. The stationary distribution is unique.
Properties of Markov chains

- If the chain is only able to return to some states at fixed intervals it is periodic, otherwise it is aperiodic.
- If the chain is irreducible, positive recurrent, and aperiodic, it is called ergodic.
- Ergodic chains will converge to the stationary distribution:
  \[ P(X_n = i) \to \pi_i, \text{ as } n \to \infty \]
  for all \( i \) and for all starting values of the chain.

Sampling from a Markov chain

- Assume that we want to construct a reversible Markov chain with stationary distribution \( p(X = k) = \pi_k \).
- Let's start with a different Markov chain with transition probabilities \( q_{ij} \).
- Accept each jump from \( i \) to \( j \) with some probability \( \alpha_{ij} \).
- The transition probabilities become
  \[ p_{ij} = q_{ij} \alpha_{ij}, \quad i \neq j \]
  \[ p_{ii} = 1 - \sum_{i \neq j} p_{ij} \]
- Picking \( \alpha_{ij} \) as
  \[ \alpha_{ij} = \min \left( 1, \frac{\pi_j q_{ji}}{\pi_i q_{ij}} \right) \]
  ensures detailed balance.

Sampling from a Markov chain (cont.)

Some important properties of the Markov chain constructed above:

1. Since \( \alpha_{ij} \) is chosen to ensure detailed balance the Markov chain has a stationary distribution.
2. If the chain is irreducible the stationary distribution is unique.
3. If the chain is aperiodic our constructed Markov chain will converge to the stationary distribution.

Two and three will depend on the proposal kernel \( q_{ij} \). However three is almost never a problem in practice.

Metropolis Hastings

- The method given above is in fact the Metropolis Hastings algorithm for sampling from a Markov chain with discrete state space.
- We now need to extend the algorithm to Markov chains with continuous state space.
- Let \( q(x, y) \) be the transition kernel of a Markov chain. It defines the conditional density of the next transition given that the current state is \( x \).
- A simple example is the AR(1) process
  \[ x_t = ax_{t-1} + \epsilon_t, \quad \epsilon_t \in N \left( 0, \sigma^2 \right) \]
Metropolis Hastings

- To construct a Markov chain with stationary distribution, \( f(x) \) we start with a proposal kernel, \( q(x,y) \), and accept the proposed jumps with probability \( \alpha(x,y) \).
- The resulting combined proposal kernel is given by
  \[
  \tilde{q}(x,y) = \alpha(x,y)q(x,y) + \left(1 - \int_x \alpha(x,z)q(x,z)dz\right) \delta_x(y).
  \]
- We now want an \( \alpha(x,y) \) that gives detailed balance for the combined Markov chain,
  \[
  f(x)q(x,y) = f(y)\tilde{q}(y,x).
  \]
- As before
  \[
  \alpha(x,y) = \min\left(1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\right),
  \]
  ensures detailed balance.

Convergence

The following are sufficient requirements for convergence of the Markov chain:
1. Detailed balance (by construction).
2. Irreducible chain
   The chain should be able to reach any point \( \{x : f(x) \neq 0\} \) regardless of the starting point.
3. Aperiodic chain
   The existens of a unique stationary distribution is ensured by 1. and 2.; aperiodic chain is needed to ensure convergence.

If the above requirements are fulfilled, then for any set \( A \subseteq \mathcal{X} \),
\[
P(X(t) \in A) \to \int_A f(x)dx, \quad t \to \infty,
\]
independently of starting point, \( x(0) \).

A look at \( \alpha(x,y) \)

- First ignore the transition kernel \( q(x,y) \).
  - \( f(y)/f(x) \) says accept (keep) the proposed value, \( y \) if it is “better” than the old value \( x \).
  - If the new value is “worse” than the old accept it with a probability proportional to how much worse.
- However, some states may be easier to reach than others, \( q(x,y) \) compensates for this.
  - If it is easy to reach \( y \) from \( x \), the denominator \( q(x,y) \) will reduce the acceptance probability.
  - It it is easy to get back to \( x \) from \( y \), the nominator \( q(y,x) \) will increase the acceptance probability.

The Metropolis Hastings algorithm

**Algorithm:**

Given a density \( f(x) \) and a proposal kernel \( q(x,y) \)

Start the chain with some \( x^{(0)} \), and loop over \( t = 1, \ldots, T \).
1. Given \( x^{(t)} \), draw a proposal \( y \) from \( q(x^{(t)},y) \).
2. Calculate the acceptance probability
   \[
   \alpha(x^{(t)},y) = \min\left(1, \frac{f(y)q(y,x^{(t)})}{f(x^{(t)})q(x^{(t)},y)}\right)
   \]
3. With probability \( \alpha(x^{(t)},y) \) accept the proposal, otherwise keep the old value, \( x^{(t)} \):
   3.1 Draw \( u \in U(0,1) \).
   3.2 Take
   \[
   x^{(t+1)} = \begin{cases} y, & \text{if } u < \alpha(x^{(t)},y) \\ x^{(t)}, & \text{if } u \geq \alpha(x^{(t)},y) \end{cases}
   \]
Different proposals

There are a number of different ways of constructing the proposal kernel, \( q(x, y) \):
- Independent proposals
- Symmetric proposals
- Multiplicative proposals
- and others

Independent proposal

- Draw the new points from \( q(y) \), independent of the current state \( x \).
- The acceptance probability reduces to
  \[
  \alpha(x, y) = \min \left( 1, \frac{f(y)q(x)}{f(x)q(y)} \right).
  \]
- \( \{ x : f(x) \neq 0 \} \subseteq \{ x : q(x) \neq 0 \} \) is needed to ensure convergence.
- If we take \( q(x) = f(x) \), the acceptance probability reduces to 1 and we get independent samples from \( f(x) \).

Symmetric proposal

- Select the proposal kernel so that \( q(x, y) = q(y, x) \).
- The acceptance probability reduces to
  \[
  \alpha(x, y) = \min \left( 1, \frac{f(y)}{f(x)} \right).
  \]
- Commonly this is \( y = x + \varepsilon \) with \( \varepsilon \) as
  - \( \varepsilon \in N(0, \Sigma) \), random walk proposal.
  - \( \varepsilon \in U(-R, R) \).

Multiplicative proposals

- Typically used for skewed distributions, such as variance parameters.
- Want an asymmetric proposal where the size of the jump depends on the current state \( x \).
- Take \( y = x\varepsilon \), with \( \varepsilon \) from som density \( g \).
- The proposal kernel now becomes \( q(x, y) = g(y/x)/x \).
- The acceptance probability reduces to
  \[
  \alpha(x, y) = \min \left( 1, \frac{f(y)g(x/y)/y}{f(x)g(y/x)/x} \right).
  \]
- The somewhat strange density
  \[
  g(f) \propto 1 + 1/f, \quad \text{for } f \in [1/F, F]
  \]
  has the nice property that \( g(x/y)/y = g(y/x)/x \) and we obtain a symmetric proposal.
Sampling from a Gumbel

- As a simple example we’ll start by sampling from a Gumbel distribution, with density function
  \[ f(x) = \frac{1}{\beta} \exp\left(-\frac{x-\mu}{\beta}\right) \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right), \]
  and distribution function
  \[ F(x) = \exp\left(-\exp\left(-\frac{x-\mu}{\beta}\right)\right). \]
- We take \( \mu = 0 \) and \( \beta = 2 \), and note that we could sample using the inversion method:
  \[ X = \mu - \beta \ln(-\ln(U)), \quad U \in (0,1). \]

Sampling from a Gumbel (cont.)

- A second alternative we’ll use an asymmetric proposal distribution. A triangular distribution with
  \[ q(x,y) = \frac{y-x+R}{2R^2}, \quad x-R \leq y \leq x+R \]
- We now have \( q(x,y) \neq q(y,x) \) and
  \[ \alpha(x,y) = \min\left(1, \frac{f(y)q(y,x)}{f(x)q(x,y)}\right) = \min\left(1, \frac{\exp\left(-\frac{y}{\beta}\right)\exp\left(-\exp\left(-\frac{y}{\beta}\right)\right)}{\exp\left(-\frac{x}{\beta}\right)\exp\left(-\exp\left(-\frac{x}{\beta}\right)\right)} (x-y+R)(y-x+R)\right). \]
- Code for both versions is available from the homepage.

Random walk proposal

- First we use a symmetric Normal proposal, \( y = x + \varepsilon \), where \( \varepsilon \sim N(0,\sigma^2) \).
- We now have \( q(x,y) = q(y,x) \) and
  \[ \alpha(x,y) = \min\left(1, \frac{f(y)}{f(x)}\right) \]
  \[ = \min\left(1, \frac{\exp\left(-\frac{y}{\beta}\right)\exp\left(-\exp\left(-\frac{y}{\beta}\right)\right)}{\exp\left(-\frac{x}{\beta}\right)\exp\left(-\exp\left(-\frac{x}{\beta}\right)\right)} \right). \]
- To avoid dividing two large numbers calculate \( \alpha(x,y) \) in log scale.
Triangular proposal

\[ \alpha: 0.52 \]

Results