6 Markov Chain Monte Carlo (Roberts and Rosenthal 2004)

The Metropolis Hastings algorithm

The aim is to sample, approximately, from the posterior distribution π that follows from doing some Bayesian statistics.

The Metropolis-Hastings algorithm uses a proposal distribution $Q(x, dy)$ to create a Markov chain kernel P_{MH} such that the Markov chain is reversible with respect to π,

$$\pi(x)P_{MH}(x, dy) = \pi(y).$$

By stationarity, π is stationary with respect to P_{MH},

$$\int \pi(x)P_{MH}(x, dy)dx = \pi(dy).$$

Starting at x, the Metropolis-Hastings algorithm is to:

- generate a sample $Y \sim Q(x, dy)$,
- if $Y = y$, jump to y with probability

$$\alpha = \min\left\{1, \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right\}.$$

Common proposal distributions are:

- The independence sampler ($Q(x, dy)$ does not depend on x, so successive proposals are independent).
- The random walk sampler ($Q(x, dy)$ is a jump of a random walk in x-space, i.e. $N(x, \sigma^2)$).
- The Gibbs sampler is a special case of the Metropolis-Hastings algorithm. Assume π is a distribution on $\mathbb{R}^d$. Pick one of the d dimensions (lets say uniformly at random). Then sample a new value for that coordinate, from the distribution πconditioned on all the other $d - 1$ coordinates staying the same. For the Gibbs sampler α is always 1. So you don't need to calculate α; all moves are accepted.

Markov chain convergence

The total-variation distance between two probability measures μ and ν on $\mathbb{R}^d$ is

- $\|\mu - \nu\|_{TV} = \sup_{A \subset \mathbb{R}^d} |\mu(A) - \nu(A)| = \frac{1}{2} \|\mu - \nu\|_{L_1}$.
- The probability $X = Y$ when $X \sim \mu$,$Y \sim \nu$, and (X, Y) lives on some joint probability space chosen to maximize the probability $X = Y$, i.e. the optimal coupling of μ and ν.

Given a Markov chain kernel P with equilibrium distribution π, we want to find $n = n(\epsilon)$ such that for some/all starting points $x \in \mathcal{X}$ ($\mathcal{X}$ is the state space, i.e. $\mathbb{R}^d$),

$$\|xP^n - \pi\|_{TV} \leq \epsilon.$$

Definition: A Markov chain is ϕ-irreducible if there exists a non-zero σ-finite measure ϕ such that for all $A \subset \mathcal{X}$ with $\phi(A) > 0$, and for all $x \in \mathcal{X}$, there is a positive integer $n = n(x, A)$ such that $P^n(x, A) > 0$.

Definition: A Markov chain is aperiodic if you cannot partition $\mathcal{X}$ into $k \geq 2$ subsets $A_1, \ldots, A_k$ such that the chain always jumps

$$A_1 \to A_2, A_2 \to A_3, \ldots, A_{k-1} \to A_k, A_k \to A_1.$$

Theorem: If a Markov chain on a state space with a countably generated σ-algebra is ϕ-irreducible and aperiodic, and has a stationary distribution π, then for π-almost every $x \in \mathcal{X}$,

$$\lim_{n\to\infty} \|P^n(x, \cdot) - \pi(\cdot)\|_{TV} = 0.$$

Definition (Small sets): A set C is *small* if there exists a positive integer n_0, and a positive ϵ, and a probability measure $\nu(\cdot)$ on $\mathcal{X}$ such that the following *minorisation condition* holds:

$$P^{n_0}(x, A) \geq \epsilon\nu(A), \qquad \forall x \in C, \forall \text{ measurable } A \subset \mathcal{X}.$$

Theorem (Uniform ergodicity): Suppose a Markov chain state space $\mathcal{X}$ is a small set (i.e. $C = \mathcal{X}$ in the definition above). Then the chain is uniformly ergodic, i.e. for some finite M and some $\rho < 1$, for every $x \in \mathcal{X}$,

$$\|P^n(x, \cdot) - \pi(\cdot)\|_{TV} \leq M\rho^n.$$

Definition (Drift Condition): A Markov chain with kernel P satisfies a drift condition (with respect to a set C) if there is a constant $b > 0$, a constant $\lambda \in (0, 1)$, and a function $V : \mathcal{X} \to [1, \infty]$ such that

$$PV \leq \lambda V + b1_C$$

i.e.

$$\forall x \in \mathcal{X}, \qquad \int V(y)P(x, dy)dx \leq \lambda V(x) + b1_c(x).$$

Theorem (Geometric Ergodicity): Suppose a ϕ-irreducible, aperiodic Markov chain with stationary distribution π has a small set C, and a drift condition for C. Then the chain is geometrically ergodic, i.e. there is a $\rho < 1$, for π-almost every $x \in \mathcal{X}$,

$$\|xP^n - \pi\|_{TV} \leq M(x)\rho^n, \qquad M(x) < \infty.$$

Questions

The (*) questions from sheets 5 to 8 form homework 2.

1. (*) Consider the Metropolis Hastings algorithm for target distribution $\pi = Uniform(0, 100)$ and proposal distribution (when the chain is at a point x) of $Uniform(x-1, x+1)$. Prove, with much hand waving, that the chain is uniformly ergodic.

2. Consider the Metropolis Hastings algorithm for target distribution $\pi = Normal(0, 100^2)$ and proposal distribution(when the chain is at a point x) of $Uniform(x-1, x+1)$. Prove, with much hand waving, that the chain is geometrically ergodic.

3. Please download lab6.zip from the website. The examples in lab6examples.R show how to use the code to do MCMC.

   - summary(chain) and plot(chain) take an mcmc object and show you useful things.
   - raftery.diag looks at a MCMC run and tells you if it is long enough, assuming you want to calculate quantiles of the posterior distribution with a fair degree of accuracy. It will also tell you if your input was too short to produce an answer.
   - geweke.diag looks at the first 10% and last 50% of the input chain. It output a Z statistic (i.e. Normal(0,1) under the assumption the chain is in equilibrium). Big values indicate an insufficient burn-in period (or some other problem).
   - gelman.diag (use run.gelman.diagnostic in lab6.R) produces a number. Values much larger than 1 (i.e.larger than 1.2 in particular) indicate lack of convergence

   (a) Try modifying some of the samples and see what happens.

   (b) The Possion change points model: You observe data $X_1, \ldots, X_N$ sampled from a Poisson distribution with changing parameter:

   $$X_1, \ldots, X_M \sim Poisson(\lambda_1), X_{M+1}, \ldots, X_N \sim Poisson(\lambda_2)$$

   Your prior beliefs about λ_1, λ_2, M are

   $$\lambda_1 \sim Gamma(\alpha_1, \beta_1), \quad \lambda \sim Gamma(\alpha_2, \beta_2), \quad M \sim Uniform\{1, 2, \ldots, N-1\}.$$

   Your posterior belief is π and

   $$\pi(\lambda_1 \mid x, \lambda_2, M) = Gamma\left(\alpha + \sum_{i=1}^{M} x_i, \beta_i + M\right),$$

   $$\pi(\lambda_2 \mid x, \lambda_1, M) = Gamma\left(\alpha + \sum_{i=M+1}^{N} x_i, \beta_i + N - M\right),$$

$$\pi(M \mid x, \lambda_1, \lambda_2) \propto \exp(M(\lambda_2 - \lambda_1))\lambda_1^{\sum_{i=1}^{M} x_i}\lambda_2^{\sum_{i=M+1}^{N} x_i}.$$

Use a Gibbs sampler to estimate λ_1, λ_2 and M.

(c) Define a random walk in $(\lambda_1, \lambda_2, M)$-space and use that instead of the Gibbs sampler.

4. (*) Please download lab6.zip from the website. Please include any R commands/source code used to generate your answers.
The University of Warwick—Centre for Complexity Science Curling Club (UWCCSCC) have been challenged to a game by their arch rivals, the University of Wyoming—Center for Complexity Science Curling Association (UWCCSCA). UWCSCCC has eight members, and you have been asked to pick the best three player team to represent the club for the game.
In preparation, the club players have played a number of games. The teams matrix describes the different groupings in the different games. The first row

```
7 6 1 4 8 5
```

indicates that players 7, 6 and 1 (team 1) played against players 4, 8 and 5 (team 2), with players 2 and 3 sitting out. The first player on each team has a role of giving the stones the initial push; the other two players are in charge of scrubbing. The entry in outcomes is 1, indicating that team 1 (players 7, 6 and 1) won the game.
You believe that each player has a skill level $(\text{skill}(i) : i = 1, \ldots, 6)$. If players a,b and c compete with players d, e and f then the probability that a, b and c win is

$$\sigma(2\text{skill}(a)+\text{skill}(b)++\text{skill}(c)-2\text{skill}(d)-\text{skill}(e)-\text{skill}(f)), \qquad \sigma(x) = \frac{1}{1+\exp(-x)}.$$

The function σ is called a sigmoid function (it looks s-shaped). Note that $\sigma(x) + \sigma(-x) = 1$; the situation is symmetric. The skill of the first player on each team team matters more, because (I guess) their role is more important.

(a) Your prior belief is that the players skills are independent and distributed according to the normal distribution with mean 0 and variance 2^2. Calculate the posterior distribution as a function of teams and outcomes.

(b) Use random-walk Metropolis Hastings MCMC to sample from the posterior distribution. Plot the output and check that your algorithm is performing reasonably. Use the diagnostics to check nothing is going wrong.
The full dataset has 10^6*samples*. You can use a subset of the data (i.e. the first n samples) to see how much data is needed to get a concentrated posterior distribution.