2 Markov chain Monte Carlo

2.1 Introduction

We now begin the next section of this course: Markov chain Monte Carlo (MCMC) simulation methods. We will only concentrate on a few of the most useful and most commonly used MCMC methods. Many, many problem specific algorithms (mostly variants of the main algorithms) have been developed to aid in convergence properties. Most of these methods were developed because the most general and widely used methods were "too slow to converge" (either N was too large for the transition kernel to converge to the stationary, or invariant, distribution, or once stationarity was reached, the chain explored the invariant distribution too slowly. That is to say the autocorrelation of the chain was too large).

Throughout we will abuse notation and let $\pi(\cdot)$ to denote a probability measure, distribution or density. We will also refer to π as a distribution regardless of whether it is a probability measure, distribution or density. The context in which π appears should make it apparent which holds. I will try to be explicit when the context is not clear. For example, if $\Phi_n = \phi$, a singleton, $\pi(\phi)$ will represent the density. In an integral, $\pi(d\phi)$ represents a distribution or measure. If ψ is a dominating measure for π (i.e. π is absolutely continuous with respect to ψ) then $\pi(A) = \int_A f(x)\psi(dx) = \int_A f(x)d\psi(x) = \int_A f d\psi$. That is, by the Radon-Nikodym theorem, $d\pi/d\psi = f(x)$ is the *Radon-Nikodym derivative* of π with respect to ψ. From time to time we may use "differential" notation: $d\pi = f d\psi$. If we are in Euclidean space, then ψ will be taken as Lebesque measure. For example, in $\mathbb{R}$, if $A = (-\infty, t)$ we write $\pi(A) = \int_{-\infty}^{t} f(x)dx$. If the dominating measure is counting measure, then the integral will be taken to be a summation: $\pi(A) = \int_A f d\psi = \sum_{x \in A} f(x)$. More often than not, we will be working in Euclidean space or a product space of Euclidean space and some countable or finite spaces.

If $A \in \mathcal{B}(\mathcal{X})$ is a set, then $\pi(A)$ obviously represents a probability measure. From the perspective of a Bayesian analysis, we will assume that we have data $\mathbf{Y} = (Y_1, \ldots, Y_n)$ that are generated from some parametrized *sampling distribution* $\pi(\mathbf{Y} \mid \Theta)$ (or likelihood), where $\Theta = (\theta_1, \cdots, \theta_p)$ is a vector of parameters with *prior distribution* $\pi(\Theta)$. The *joint distribution* of the sampling distribution and the prior is $\pi(\mathbf{Y}, \Theta) = \pi(\mathbf{Y} \mid \Theta)\pi(\Theta)$. Applying Bayes' theorem, we can invert the conditional sampling distribution to obtain the *posterior distribution* of the parameters given the data:

$$\pi(\Theta \mid \mathbf{Y}) = \frac{\pi(\mathbf{Y}, \Theta)}{\pi(\mathbf{Y})} = \frac{\pi(\mathbf{Y} \mid \Theta)\pi(\Theta)}{\int \pi(\mathbf{Y} \mid \Theta)\pi(d\Theta)} \tag{17}$$

Assume now that ψ is Lebesque measure. We will also assume that you are familiar with conjugate prior-posterior pairs. In this case we can obtain independent samples directly from

the posterior. Typically, for a Bayesian analysis we are going to be interested in expectations of some measurable function $f \in L^1$ (that is f absolutely integrable) with respect to the posterior distribution:

$$\mathbb{E}_{\pi(\Theta|\mathbf{Y})}\left[f(\Theta) \mid \mathbf{Y}\right] = \int_{\Theta \in \mathbf{\Theta}} f(\Theta)\pi(\Theta \mid \mathbf{Y})d\Theta. \tag{18}$$

In terms of Markov chain theory, $\mathbf{\Theta}$ is our state space. We will refer to the state space as the *support* of the posterior distribution.

This may seem restrictive, but for various forms of f we can estimate various posterior quantities of interest:

Posterior mean: Take $f(\Theta) = \Theta$.

Posterior covariance: Set $f(\Theta) = (\theta_j - \mathbb{E}_\pi[\theta_j \mid D])(\theta_k - \mathbb{E}_\pi[\theta_k \mid D])$ where $\mathbb{E}_\pi(\theta_k \mid D) = \int_{\mathbf{\Theta}} \theta_j \pi(\Theta \mid \mathbf{Y})d\Theta$.

Posterior predictive density: Let $f(\Theta) = \pi(\mathbf{Y}_{\text{new}} \mid \Theta)$.

Posterior probability of $A \in \mathbf{\Theta}$: Let $f(\Theta) = \mathbb{I}_A(\Theta)$.

We end this section by a formal definition of an MCMC method.

Definition 39 *A* Markov chain Monte Carlo *method for the simulation of a distribution π is any method producing an ergodic Markov chain $\mathbf{\Phi}$ whose invariant distribution is π.*

2.2 Metropolis-Hastings algorithm

We will now construct a ψ-irreducible Markov chain $\mathbf{\Phi}$ with transition probability kernel P whose invariant probability measure is π.

Suppose that $\pi \prec \psi$, that is π is absolutely continuous with respect to ψ, the dominating measure. Furthermore suppose the $\pi(x)$ is the density of π with respect to ψ. Let Q be a transition probability kernel such that

$$Q(x, A) = \int_A Q(x, dy) = \int_A q(x, y)\psi(dy).$$

To avoid some trivial special cases, we will assume that $Q(x, \mathcal{X}^+) = 1$ for $x \notin \mathcal{X}^+$ where $\mathcal{X}^+ = \{x : \pi(x) > 0\}$. We will also assume that π is not concentrated on a singleton. Define

$$\alpha(x,y) = \begin{cases} \min\left(\frac{\pi(y)q(y,x)}{\pi(x)q(x,y)}, 1\right) & : \quad \pi(x)q(x,y) > 0, \\ 1 & : \quad \pi(x)q(x,y) = 0. \end{cases}$$

The Metropolis-Hastings algorithm

1. Given $\Phi_n = x$, draw $Y \sim Q(x, \cdot)$.
2. Set
$$\Phi_{n+1} = \begin{cases} Y & \text{with probability } \alpha(x,y) \\ x & \text{with probability } 1 - \alpha(x,y). \end{cases}$$

Note that as long as $\pi(x)q(x,y) > 0$, $\alpha(x,y) = 0$ if $\pi(y) = 0$. Thus $\mathbf{\Phi}$ stays in $\mathcal{X}^+$ almost surely once it enters $\mathcal{X}^+$. The restriction imposed on Q ensures that the chain enters $\mathcal{X}^+$ after at most one step. In practice, the initial state is always chosen to lie in $\mathcal{X}^+$.

2.2.1 Convergence properties

In order to study the properties of the Markov chain $\mathbf{\Phi}$ generated from this algorithm, we need to formally define its probability transition kernel, P, and we do so on all of $\mathcal{X}$, for completeness. Let

$$p(x,y) = \begin{cases} \alpha(x,y)q(x,y) & : \quad x \neq y \\ 0 & : \quad x = y \end{cases},$$

and

$$r(x) = 1 - \int_{\mathcal{X}} p(x,y)\psi(dy).$$

The value $r(x)$ is the probability that the algorithm remains at x. Then the Metropolis-Hastings probability kernel is

$$P(x,A) = \int_A P(x,dy) = \int_A \left[p(x,y)\psi(dy) + r(x)\delta_x(dy)\right]. \tag{19}$$

Definition 40 *We call a Markov chain $\mathbf{\Phi}$ with transition probability kernel (19) a* Metropolis-Hastings chain.

Proposition 25 *The Metropolis-Hastings chain is reversible and π is its invariant probability measure.*

Proof: We will first show that the chain is reversible. There are two cases:

Case 1: $x \neq y$. Then $P(x, dy) = p(x, y)\psi(dy)$. Suppose $\alpha(x, y) < 1$.

$$\begin{aligned}
\pi(dx)P(x, dy) &= \pi(dx)p(x, y)\psi(dy) \\
&= \pi(dx)\alpha(x, y)q(x, y)\psi(dy) \\
&= \Big(\pi(x)\psi(dx)\Big)\left[\frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}\right]q(x, y)\psi(dy) \\
&= \Big(\pi(y)\psi(dy)\Big)\Big(\alpha(y, x)q(y, x)\Big)\psi(dx) \quad (\alpha(y, x) = 1) \\
&= \pi(dy)p(y, x)\psi(dx) = \pi(dy)P(y, dx).
\end{aligned}$$

Now suppose $\alpha(x, y) = 1$, then $\alpha(y, x) < 1$. Now start with $\pi(dy)P(y, dx)$ and follow the same line of reasoning.

Case 2: $x = y$. Then $P(x, dy) = r(x)\delta_x(dy)$. So

$$\begin{aligned}
\pi(dx)P(x, dy) &= \pi(dx)r(x)\delta_x(dy) \\
&= \pi(x)\psi(dx)r(x)\delta_x(y)\psi(dy)
\end{aligned}$$

Now, $\delta_x(y) = 1$ if and only if $x = y$ if and only if $\delta_y(x) = 1$. Therefore,

$$\begin{aligned}
&= \pi(x)\psi(dx)r(x)\delta_y(x)\psi(dy) \\
&= \pi(y)\psi(dy)r(y)\delta_y(x)\psi(dx) \\
&= \pi(dy)r(y)\delta_y(dx) = \pi(dy)P(y, dx).
\end{aligned}$$

Hence, $\mathbf{\Phi}$ is reversible.

Now we show that π is the invariant distribution (measure) of $\mathbf{\Phi}$. Let $A \in \mathcal{B}(\mathcal{X})$.

$$\begin{aligned}
\int_{\mathcal{X}} \pi(dx)P(x, A) &= \int_{\mathcal{X}} \pi(dx) \int_A P(x, dy) \\
&= \int_{\mathcal{X}} \pi(x)\psi(dx) \int_A p(x, y)\psi(dy) + \int_{\mathcal{X}} r(x)\pi(x)\psi(dx) \int_A \delta_x(dy) \\
&= \int_A \psi(dy) \int_{\mathcal{X}} \pi(x)p(x, y)\psi(dx) + \int_{\mathcal{X}} r(x)\pi(x)\delta_x(A)\psi(dx) \\
&= \int_A \pi(y)\psi(dy) \int_{\mathcal{X}} p(y, x)\psi(dx) + \int_A r(x)\pi(x)\psi(dx) \\
&= \int_A [1 - r(y)]\pi(y)\psi(dy) + \int_A r(x)\pi(x)\psi(dx) \\
&= \int_A \pi(y)\psi(dy) = \int_A \pi(dy) = \pi(A).
\end{aligned}$$

□

The Metropolis-Hastings chain is reversible with invariant measure π. We would like for the LLN and the CLT to hold for this chain. Thus, we need to find conditions such that $\mathbf{\Phi}$ is a π-irreducible, where π is a maximal irreducible measure, aperiodic, positive Harris chain. These conditions will be restrictions on the proposal distribution $Q(x, \cdot)$.

Proposition 26 *If $q(x, y) > 0$ for all $x, y \in \mathcal{X}$, then the Metropolis-Hastings chain is π-irreducible, where π is a maximal irreducible measure.*

Proof: First we note that if the chain is ψ-irreducible for any maximal irreducible measure ψ, then since π is an invariant measure for $\mathbf{\Phi}$ by the last proposition, we have that $\psi \prec \pi$ by Proposition 19 on page 45. Then by the definition of a maximal irreducible measure (Proposition 6 on page 9), π is maximal irreducible.

Now let $A \in \mathcal{B}(\mathcal{X})$ such that $\psi(A) > 0$. By uniqueness of $\mathcal{B}^+(\mathcal{X})$ and the equivalence of π and ψ, it follows that $\psi(A) > 0$ if and only if $\pi(A) > 0$. For all $x \in \mathcal{X}$,

$$P(x, A) = \int_A P(x, dy) = \int_A \alpha(x, y) q(x, y) \psi(dy) + r(x)\delta_x(dy) \geq \int_A \alpha(x, y) Q(x, dy) > 0,$$

since $q(x, y) > 0$ for all $x, y \in \mathcal{X}$, then for $A \in \mathcal{B}^+(\mathcal{X})$, we have $Q(x, A) > 0$. Hence the chain is π-irreducible. □

We will also define $P(x, y)$ via

$$\int_A P(x, dy) = \int_A P(x, y)\psi(dy) = \int_A [\alpha(x, y)q(x, y) + r(x)\delta_x(y)]\psi(dy).$$

A sufficient condition for the chain to be aperiodic is that $P_x(\Phi_1 = x) > 0$. And this true by construction of the transition kernel. Note that this implies, by the definition of $\alpha(x, y)$, that

$$\Pr[\pi(x)q(x, y) \leq \pi(y)q(y, x)] < 1.$$

To show the Metropolis-Hastings chain is Harris requires the notion of a harmonic function.

Definition 41 (Harmonic function) *A measurable function h is* harmonic *for $\mathbf{\Phi}$ if*

$$\mathbb{E}[h(\Phi_{n+1}) \mid \Phi_n = x] = h(x).$$

Note that if $h(x) = x$, then this is the definition of a martingale.

Lemma 11 *If $\mathbf{\Phi}$ is positive recurrent and aperiodic with invariant measure π and if h is a harmonic function, then*

$$h(\Phi_n) = \mathbb{E}_\pi[h(\Phi_1)]$$

π-almost everywhere (h is π-almost everywhere constant).

Lemma 12 *If $\mathbf{\Phi}$ is positive, then $\mathbf{\Phi}$ is Harris if and only if the only bounded harmonic functions are the constant functions.*

Proof: Meyn & Tweedie, p. 425ff.

Proposition 27 *If the Metropolis-Hastings chain $\mathbf{\Phi}$ is π-irreducible, then it is Harris.*

Proof: Suppose the Metropolis-Hastings chain $\mathbf{\Phi}$ is π-irreducible (e.g. $q(x, y) > 0$ for all $x, y \in \mathcal{X}$). Also $\mathbf{\Phi}$ is aperiodic by construction of its transition kernel and is positive with invariant measure π. Hence Lemma 11 tells us that any harmonic function is constant π-almost everywhere.

$$\begin{aligned}
h(x) = \mathbb{E}_x[h(\Phi_1)] &= \int_{\mathcal{X}} h(y)\alpha(x,y)q(x,y)\psi(dy) + r(y)h(y)\delta_x(dy) \\
&= \int_{\mathcal{X}} h(y)\alpha(x,y)q(x,y)\psi(dy) + r(x)h(x) \\
&= \int_{\mathcal{X}} \mathbb{E}_\pi[h(\Phi_1)]\alpha(x,y)q(x,y)\psi(dy) + r(x)h(x) \quad \text{by Lemma 11} \\
&= \mathbb{E}_\pi[h(\Phi_1)][1 - r(x)] + r(x)h(x). \\
\Longrightarrow & \\
h(x) &= \mathbb{E}_\pi[h(\Phi_1)], \quad \forall\, x \in \mathcal{X},
\end{aligned}$$

since $1 - r(x) > 0$ by virtue of π-irreducibility. Therefore h is constant, and by Lemma 12, $\mathbf{\Phi}$ is Harris. □

Now we have shown that the Metropolis-Hastings chain is π-irreducible, positive Harris with invariant measure π and aperiodic if the proposal density $q(x, y) > 0$ for all $x, y \in \mathcal{X}$. This last condition is equivalent to saying that the support of Q is at least as large as the support of π. That is to say

$$\mathcal{X} \subseteq \bigcup_{x \in \mathcal{X}} \operatorname{supp} Q(x, \cdot).$$