Theorem 23 *For any proposal density such that $\mathcal{X} \subseteq \cup_{x \in \mathcal{X}} \text{supp}\, Q(x, \cdot)$, the Metropolis-Hastings chain is π-irreducible, aperiodic and positive Harris. Therefore the LLN and the CLT hold.*

Proof: The first part has already been established. That the LLN of large numbers holds, we invoke Theorem 20, page 57. The CLT holds as as a result of Theorem 22, page 58. □

2.2.2 Geometric and Uniform convergence

Recall the definition of the total variation norm: For a signed measure μ on $\mathcal{B}(\mathcal{X})$

$$||\mu||_{TV} = \sup_{f:|f|\leq 1} |\mu(f)| = \sup_{A \in \mathcal{B}(\mathcal{X})} \mu(A) - \inf_{A \in \mathcal{B}(\mathcal{X})} \mu(A).$$

We now extend this definition. We will consider all functions that are dominated by a function $f : \mathcal{X} \to [1, \infty)$. That is we consider convergence in the f-norm defined by

$$||\nu||_f = \sup_{g:|g|\leq f} |\nu(g)| = \sup_{g:|g|\leq f} \left| \int_{\mathcal{X}} \nu(dx) g(x) \right|.$$

The total variation norms results when $f \equiv 1$.

Definition 42 (f-Geometric ergodicity) $\mathbf{\Phi}$ *is called* f-geometrically ergodic*, where $f \geq 1$, if $\mathbf{\Phi}$ is positive Harris with $\pi(f) = \mathbb{E}_\pi[f(\Phi_1)] < \infty$ and there exists a constant $r_f > 1$ such that*

$$\sum_{n=1}^{\infty} r_f^n \, ||P^n(x, \cdot) - \pi||_f < \infty \tag{20}$$

for all $x \in \mathcal{X}$. If (20) holds for $f \equiv 1$, then we call $\mathbf{\Phi}$ geometrically ergodic.

Note that f-geometric ergodicity means that $||P^n(x, \cdot) - \pi||_f$ is decreasing at least at a geometric rate:

$$||P^n(x, \cdot) - \pi||_f \leq M r_f^{-n}$$

where M is the l.h.s. of (20). When $f \equiv 1$ we have the total

Definition 43 (Uniform ergodicity) *A Markov chain $\mathbf{\Phi}$ is called* uniformly ergodic *if*

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

Uniform ergodicity is stronger than geometric ergodicity in the sense that the rate of geometric convergence must be uniform over all of $\mathcal{X}$.

Theorem 24 *For any Markov chain $\mathbf{\Phi}$ the following are equivalent statements:*

(i) *$\mathbf{\Phi}$ is uniformly ergodic.*

(ii) *There exists $r > 1$ and $R < \infty$ such that for all $x \in \mathcal{X}$*

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

(iii) *The state space $\mathcal{X}$ is ν_m-small for some m.*

Proof: Meyn & Tweedie, Chapter 16.

Next we will consider two special cases of the Metropolis-Hastings algorithm: the independent MH algorithm and the random walk MH algorithm.

2.2.3 The independent MH algorithm

The independent MH algorithm is the Metropolis-Hastings algorithm with $Q(x, A) = Q(x', A)$ for all $x, x' \in \mathcal{X}$. That is to say the proposition distribution $Q(x, A)$ is independent of x. In this case we write $Q(x, A) = Q(A)$ and it's associated density with respect to ψ as $q(x, y) = q(y)$.

Note that although the Y's are generated independently, the elements in the resulting chain $\mathbf{\Phi}$ are not independent as the probability of accepting a Y at time $n + 1$ depends on the value of $\mathbf{\Phi}_n$ through the ratio $\pi(y)/\pi(x)$, except in the trivial case when $\pi \equiv q$.

Theorem 25 *The independent MH algorithm produces a uniformly ergodic chain if there exists a constant M such that*

$$\pi(x) \leq Mq(x), \quad \forall\, x \in \mathcal{X}.$$

Then,

$$||P^n(x,\cdot) - \pi||_{TV} \leq 2\left(1 - \frac{1}{M}\right)^n.$$

If for every M there exists a set, A, of positive measure such that

$$\pi(x) > Mq(x), \quad \forall\, x \in A,$$

then $\mathbf{\Phi}$ *is not even geometrically ergodic.*

Proof: We will prove the first part (the second part relies on results that we have not covered). Now, for all $x \in \mathcal{X}$ and $A \in \mathcal{B}(\mathcal{X})$

$$\begin{aligned} P(x,A) &= \int_A \alpha(x,y)q(y)\psi(dy) + r(x)\delta_x(dy) \\ &\geq \int_A q(y)\min\left(\frac{\pi(y)q(x)}{\pi(x)q(y)}, 1\right)\psi(dy) \\ &= \int_A \min\left(\pi(y)\frac{q(x)}{\pi(x)}, q(y)\right)\psi(dy) \\ &\geq \frac{1}{M}\int_A \pi(y)\psi(dy). \end{aligned}$$

Therefore, by definition, $\mathcal{X}$ is a ν_1-small set with $\nu_1(A) = \int_A f(y)\psi(dy)$, By Theorem 24, page 66, $\mathbf{\Phi}$ is uniformly ergodic.

The bound on $||P^n(x,\cdot) - \pi||_{TV}$ is established via induction. For $n = 1$:

$$\begin{aligned} ||P(x,\cdot) - \pi||_{TV} &= 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\pi(A) - P(x,A)| \\ &= 2\sup_A \left| \int_A \pi(y)\psi(dy) - P(x,dy) \right| \\ &= 2\int_{y:\pi(y) \geq P(x,y)} \pi(y)\psi(dy) - P(x,dy) \\ &\leq 2\int_{y:\pi(y) \geq P(x,y)} \pi(y)\psi(dy) - \frac{1}{M}\pi(y)\psi(dy) \\ &= 2\left(1 - \frac{1}{M}\right)\int_{y:\pi(y) \geq P(x,y)} \pi(y)\psi(dy) \\ &\leq 2\left(1 - \frac{1}{M}\right). \end{aligned}$$

Now assume $||P^{n-1}(x,\cdot)-\pi||_{TV} \leq 2(1-1/M)^{n-1}$. Then for n,

$$\begin{aligned}
||P^{n}(x,\cdot)-\pi||_{TV} &= 2\sup_A |P^n(x,A)-\pi(A)| \\
&= 2\sup_A \left| \int_A \left[P^n(x,z)-\pi(z)\right]\psi(dz) \right| \\
&= 2\sup_A \left| \int_A \left\{ \int_{\mathcal{X}} \left[P^{n-1}(u,z)-\pi(z)\right]\left[P(x,u)-\pi(u)\right]\psi(du) \right\} \psi(dz) \right| \\
&= \sup_A \left| \int_{\mathcal{X}} \left[2\int_A \left(P^{n-1}(u,z)-\pi(z)\right)\psi(dz) \right] \left[P(x,u)-\pi(u)\right]\psi(du) \right| \\
&\leq \sup_A \left| \int_{\mathcal{X}} 2(1-1/M)^{n-1}\left[P(x,u)-\pi(u)\right]\psi(du) \right| \\
&= (1-1/M)^{n-1} 2\sup_A \left| \int_{\mathcal{X}} \left[P(x,u)-\pi(u)\right]\psi(du) \right| \\
&\leq (1-1/M)^{n-1} 2\sup_A \left| \int_{A} \left[P(x,u)-\pi(u)\right]\psi(du) \right| \qquad [\text{since } \mathcal{X} \in \mathcal{B}(\mathcal{X})] \\
&\leq 2(1-1/M)^n.
\end{aligned}$$

□

Proposition 28 *If there exists an M such that $\pi(x) \leq Mq(x)$ for all $x \in \mathcal{X}$, then the expected acceptance probability associated with the independent MH algorithm is at least M^{-1} when the chain is stationary.*

Proof: If the distribution of $\pi(X)/q(X)$ is absolutely continuous then,

$$\begin{aligned}
\mathbb{E}\left[\alpha(x,y)\right] &= \int_{\mathcal{X}}\int_{\mathcal{X}} \alpha(x,y)\mathbb{I}\left(\frac{\pi(y)q(x)}{\pi(x)q(y)} > 1\right)\pi(dx)Q(dy) \\
&\qquad + \int_{\mathcal{X}}\int_{\mathcal{X}} \alpha(x,y)\mathbb{I}\left(\frac{\pi(y)q(x)}{\pi(x)q(y)} \leq 1\right)\pi(dx)Q(dy) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \mathbb{I}\left(\frac{\pi(y)q(x)}{\pi(x)q(y)} > 1\right)\pi(x)q(y)\psi(dx)\psi(dy) \\
&\qquad + \int_{\mathcal{X}}\int_{\mathcal{X}} \frac{\pi(y)q(x)}{\pi(x)q(y)}\mathbb{I}\left(\frac{\pi(y)q(x)}{\pi(x)q(y)} \leq 1\right)\pi(x)q(y)\psi(dx)\psi(dy) \\
&= \int_{\mathcal{X}}\int_{\mathcal{X}} \pi(x)q(y)\psi(dx)\psi(dy) \\
&\geq \int_{\mathcal{X}}\int_{\mathcal{X}} \pi(x)\frac{\pi(y)}{M}\psi(dx)\psi(dy) = \frac{1}{M}\int_{\mathcal{X}}\pi(dx)\int_{\mathcal{X}}\pi(dy) = \frac{1}{M}
\end{aligned}$$

The probability above is equal to 1/2 because Y and X are independent and identically distributed according to π due to the fact that the chain is in stationarity. □

2.2.4 Random walk Metropolis-Hastings algorithm

If in the general MH algorithm, we choose a proposal distribution that is some random perturbation of the current value:

$$Y = \Phi_n + \epsilon,$$

where epsilon is a random perturbation independent of Φ_n, then we can take $q(x, y) = x+q(y)$, and the resulting algorithm is called a *random walk MH algorithm.* Typical examples are $\epsilon \sim N(0, \sigma^2)$ or ϵ distributed uniformly on a sphere centered at 0. if q is a symmetric function such that $g(-x) = g(x)$, then the resulting algorithm is that originally proposed by Metropolis and his colleagues in 1953. It typically is given the shortened name: Metropolis algorithm. For the Metropolis algorithm, $\alpha(x, y)$, the MH acceptance probability, take a simplified form:

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

This is because $Y \sim q(|y - x|)$ and so the ratio of the proposal densities cancel.

Unlike the independent MH algorithm, the random MH walk algorithm does not produce a uniformly ergodic Markov chain. However, Mengersen and Tweedie (1996, *Annals of Statistics*) show that (I believe in the case $\mathcal{X} = \mathbb{R}$) if there exists $\alpha > 0$ and x_1 such that

$$\ln \pi(x) - \ln \pi(y) \geq \alpha|y - x|$$

for $y < x < -x_1$ or $x_1 < x < y$, and π is a symmetric density and q is positive and symmetric, then the Markov chain produced by the Metropolis algorithm is geometrically ergodic. If π is asymmetric then a sufficient condition for geometric ergodicity is that $q(x)$ be bounded by $b \exp[-\alpha|x|]$ for a sufficiently large constant b.

Typically, it is much easier to use a random walk MH algorithm. It may be quite difficult to find a good proposal distribution q such that $\pi(x) \leq Mq(x)$ where M is reasonably small so that the expected acceptance rate $1/M$ is reasonably large.

A difficulty with a random walk MH algorithm is that of determining the scale of the symmetric density q. If the scale is too large, the acceptance rate is going to be too small and the chain will spend a long time ”stuck” at one particular value. If the scale is too small, the acceptance rate my be reasonable, but it will take the chain a long time to explore the entire distribution π. Furthermore, if the chain is started in a location not near the center of π, it may take the chain a long time to reach stationarity. Once there, it may take an extremely long time for the LLN to kick in.

One solution, is to “tune” the algorithm by sampling for a short amount of time and estimating the acceptance rate. If it is too low, decrease the proposal variance. If it is too high, increase the proposal variance. I second method, one that I use often, is an adaptive procedure. This is done by monitoring the acceptance rate and modifying the proposal variance on the fly. However, after some time, the adaptation must cease. Otherwise, you are not producing a time-homogeneous Markov chain. Furthermore, the entire chain during the adaptation phase must be discarded as we have no assurance that any of the ergodic theory is valid.

2.3 Gibbs sampler

As opposed to the quite general applicability of the MH algorithm, the Gibbs sampler utilizes properties of the stationary distribution π. We suppose that $X \in \mathcal{X}$ can be written as $X = (X_1, \ldots, X_p)$, for $p > 1$ where each X_i is either a singleton or multidimensional. Suppose that we can simulate from the corresponding conditional densities $\pi_1, \ldots, \pi_p$, i.e.

$$X_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p \sim \pi_i(\cdot | x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_p).$$

The associated Gibbs algorithm (or Gibbs sampler) is

The Gibbs sampler

Given $\Phi_n = x = (x_1, \ldots, x_p)$, draw Φ_{n+1} element-wise

$$\begin{array}{lllll}
1. & \Phi_{n+1}(1) & \sim & \pi_1(\cdot \mid x_2, \ldots, x_p), & \text{call it } x_1(n+1) \\
2. & \Phi_{n+1}(2) & \sim & \pi_2(\cdot \mid x_1(n+1), x_3, \ldots, x_p), & \text{call it } x_2(n+1) \\
 & \vdots & & & \\
p. & \Phi_{n+1}(p) & \sim & \pi_2(\cdot \mid x_1(n+1), x_2(n+1), \ldots, x_{p-1}(n+1)), & \text{call it } x_p(n+1)
\end{array}$$

We could spend an entire lecture or more discussing the convergence properties of the Gibbs sampler based on its own merits. But since it can thought of a special case of the MH algorithm (it is the composition of p MH algorithms all with $\alpha(x, y) \equiv 1$—the proposal distributions are just the conditional densities), we will simply state that the convergence properties (and ergodic properties) of the MH algorithm carry over to the Gibbs sampler.

There is another important sampler called the *Metropolis-within-Gibbs algorithm.* Suppose that one of the full conditionals in the Gibbs algorithm does not have an analytic form from which we can easily sample. We simply replace that step with a Metropolis-Hastings algorithm for that particular element of Φ, say it is the kth element. It is important to note that only one step of the MH algorithm is performed. We either except the proposed Y and set $\Phi_{n+1}(k) = y$ or reject it and set $\Phi_{n+1}(k) = \Phi_n(k) = x_k$. Then proceed to the $(k+1)$th element of the algorithm.

What remains to be shown is that the joint distribution π is uniquely determined, up to a constant multiple, by the full conditionals. We will show that this is the case in a special case when the positivity condition holds:

Definition 44 (Positivity condition) *Let $(Y_1, \ldots, Y_n) \sim F(Y_1, \ldots, Y_n)$ with density f. Let $F^{(i)}$ denote the marginal distribution of Y_i with density $f^{(i)}$. If $f^{(i)}(y_i) > 0$, for all i, implies $f(y_1, \ldots, y_n) > 0$, then F is said to satisfy the* positivity condition.

Lemma 13 (Brook's lemma) *Suppose F satisfies the positivity condition. Then the joint density is uniquely determined (up to a multiplicative constant) by the full conditional densities.*

Proof: Suppose $(x_1, x_2, \ldots, x_n)$ is given, then

$$\begin{aligned}
f(y_1, y_2, \ldots, y_n) &= f(y_n \mid y_1, , \ldots, y_{n-1}) f(y_1, \ldots, y_{n-1}) \\
&= \frac{f(y_n \mid y_1, , \ldots, y_{n-1})}{f(x_n \mid y_1, \ldots, y_{n-1})} f(y_1, \ldots, y_{n-1}, x_n) \\
&= \frac{f(y_n \mid y_1, \ldots, y_{n-1})}{f(x_n \mid y_1, \ldots, y_{n-1})} f(y_{n-1} \mid, y_1, \ldots, y_{n-2}, x_n) f(y_1, \ldots, y_{n-2}, x_n) \\
&= \frac{f(y_n \mid y_1, \ldots, y_{n-1})}{f(x_n \mid y_1, \ldots, y_{n-1})} \frac{f(y_{n-1} \mid y_1, \ldots, y_{n-2}, x_n)}{f(x_{n-1} \mid y_1, \ldots, y_{n-2}, x_n)} f(y_1, \ldots, x_{n-1}, x_n) \\
&\;\;\vdots \\
&= \prod_{j=1}^{n} \frac{f(y_j \mid y_1, \ldots, y_{j-1}, x_{j+1}, \ldots, x_n)}{f(x_j \mid y_1, \ldots, y_{j-1}, x_{j+1}, \ldots, x_n)} f(x_1, \ldots, x_n).
\end{aligned}$$

The positivity condition ensures that all of the factors in the denominator are non-zero. □

It is usually the case that the positivity condition holds.