Markov chains - general state space

We give here a brief overview on how the theory of Markov chains generalizes to a continuous state space $\mathcal{X}$, typically a subset of $\mathbb{R}^d$ with non-zero Lebesgue measure.

Definition: a Markov Transition kernel on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, where $\mathcal{B}(\mathcal{X})$ is the Borel σ-algebra on $\mathcal{X}$, is a function $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$, s.t.

1) $\forall x \in \mathcal{X}$, $P(x, \cdot)$ is a probability measure on $\mathcal{X}$
2) $\forall A \in \mathcal{B}(\mathcal{X})$, $P(\cdot, A)$ is measurable

Whenever $P(x, \cdot)$ admits a density with respect to the Lebesgue measure, we denote it by $p : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, i.e.

$$\forall x \in \mathcal{X},\ A \in \mathcal{B}(\mathcal{X}), \quad P(x, A) = \int_A p(x, y)\, dy$$

Definition: Given a Markov Transition kernel P and a measure λ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, a sequence of random variables $\{X_n \in \mathcal{X},\ n \geq 0\}$ is a Markov chain with transition kernel P and initial distribution λ, in short $\{X_n\} \sim \text{Markov}(\lambda, P)$ if

- $X_0 \sim \lambda$
- $\mathbb{P}(X_{n+1} \in A \,/\, X_n = x_n, \dots, X_0 = x_0) = \mathbb{P}(X_{n+1} \in A \,/\, X_n = x_n) = P(x_n, A)$

Then, the Markov chain $\{X_n\} \sim \text{Markov}(\lambda, P)$ satisfies a strong Markov property: let τ be a stopping time, conditional on $\tau < +\infty$,

$$\mathbb{E}_\lambda[h(X_{\tau+1}, X_{\tau+2}, \dots)] = \mathbb{E}_{X_\tau}[h(X_1, X_2, \dots)]$$

for any bounded function $h : \mathcal{X}^{\mathbb{N}} \to \mathbb{R}$

The n-step transition kernel $P^{(n)}(x, A) = \mathbb{P}(X_n \in A \,/\, X_0 = x)$ is

given by the recursion

$$P^{(n)}(x,A) = \int P^{(n-1)}(y,A)\, P(x,dy) \qquad P^{(1)}(x,A) = P(x,A)$$

In terms of densities, this corresponds to a convolution operation:

$p^{(2)}(x,z) = \int p(y,z)\, p(x,y)\, dy = p(x,\cdot) * p(\cdot,z)$ and iteratively, $p^{(n)} = p^{*n}$.

To each Markov transition kernel P we can associate transition operator $\mathcal{P} : M_1(\mathcal{X}) \to M_1(\mathcal{X})$ with $M_1(\mathcal{X})$ the set of probability measures on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ as

$$\mu = \lambda \mathcal{P} \Rightarrow \forall A \in \mathcal{B}(\mathcal{X}), \quad \mu(A) = \int P(y,A)\, \lambda(dy)$$

Notice that $\lambda \mathcal{P}^2 = (\lambda \mathcal{P})\mathcal{P} = \iint P(x,\cdot) P(y,dx) \lambda(dy) = \int P^{(2)}(y,\cdot)\, \lambda(dy)$ so $\mathcal{P}^2$ is the operator associated to $P^{(2)}$ and more generally, $\mathcal{P}^n$ to $P^{(n)}$.

If $\pi^{n,\lambda}$ denotes the measure associated to X_n, i.e. $\pi^{n,\lambda}(A) = \mathbb{P}_\lambda(X_n \in A)$, it follows that $\pi^{n,\lambda} = \lambda \mathcal{P}^n = \int P^{(n)}(y,\cdot)\, \lambda(dy)$

Definition: a measure π on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ is called invariant (or stationary) if $\pi = \pi \mathcal{P} = \int P(y,\cdot)\, \pi(dy)$. If the measure has a density $\pi(A) = \int_A f(y)\, dy$, then $f(x) = \int P(y,x) f(y)\, dy$.

We now extend the concepts of irreducibility, recurrence and aperiodicity. In the discrete setting, we have said that i communicate with j if $\exists n > 0 : P^n_{ij} > 0$ and a chain is irreducible if any state communicates with every other state. In the continuous case, the definition is slightly more cumbersome as, in general, for continuous random variables $P^{(n)}(x,y) = \mathbb{P}_x(X_n = y) = 0 \ \forall n$.

Definition (φ-irreducibility): Given a measure φ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, we say that $\{X_n\} \sim \text{Markov}(\lambda, P)$ is φ-irreducible if $\forall x \in \mathcal{X}$ and $\forall A \in \mathcal{B}(\mathcal{X})$ with $\varphi(A) > 0$, $\exists n > 0$ s.t. $P^{(n)}(x,A) > 0$.

The notion of irreducibility does not really depend on the measure φ as shown by the next result:

Theorem [Meyn-Tweedie, Prop. 4.2.2.] If $\{X_n\} \sim \text{Markov}(\lambda, P)$ is φ-irreducible for some measure φ on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$, then there exists a a probability measure ψ on $\mathcal{B}(\mathcal{X})$, called maximal irreducibility measure such that

- $\{X_n\}$ is ψ-irreducible
- for any other measure ψ' on $\mathcal{B}(\mathcal{X})$ for which $\{X_n\}$ is ψ'-irreducible, one has ψ' is absolutely continuous with respect to ψ (i.e. $\forall A \in \mathcal{B}(\mathcal{X}), \quad \psi(A) > 0 \Rightarrow \psi'(A) > 0$)

The maximal irreducibility measure is in general not unique, but all maximal irreducibility measures have the same null sets (i.e. they are equivalent). If $\{X_n\}$ has an invariant distribution π, then π is a maximal irreducibility measure.

In the context of Markov Chain Monte Carlo, where the target distribution π is given, we have to check that the chain is π-irreducible.

A similar construction applies for the notion of recurrence. In general, requiring that a single state is recurrent, i.e. $\mathbb{P}(X_n = y \text{ i.o.}) = 1$, will never be satisfied for continuous random variable and we have to relax it:

Definition (Harris recurrence) Let ψ be any maximal irreducibility measure. A chain $\{X_n\} \sim \text{Markov}(\lambda, P)$ is Harris recurrent if it is ψ-irreducible and $\forall A \in \mathcal{B}(\mathcal{X})$, $\psi(A) > 0$ and $\underline{\forall x \in \mathcal{X}}$, $\mathbb{P}_x(X_n \in A \text{ i.o.}) = 1$.

In the discrete case, irreducibility + recurrence implies the existence of an invariant measure. The proof relies on the fact that the times $\tau_k^{(r)}$ of visits to a given state x_k are renewal times, i.e. the chains $\{X_n, \tau_k^{(r-1)} < n \le \tau_k^{(r)}\}$ are iid and the invariant measure can be constructed as $\pi_i = \mathbb{E}_k\left[\sum_{n=0}^{\tau_k^{(1)}-1} \mathbb{1}_{\{X_n = x_n\}}\right]$.

In the continuous case the construction is more difficult as a single state $x \in \mathcal{X}$ is not, in general, recurrent. Therefore, we should look for recurrent sets $A \in \mathcal{B}(\mathcal{X})$ and arrival times $\tau_A^{(r)} = \inf\{n > \tau_A^{(r-1)} : X_n \in A\}$. However, these are generally not renewal times as the chain X_n, $n > \tau_A^{(r)}$ depends on the specific state $X_{\tau_A^{(r)}}$ visited in A and not just on A itself. A useful concept to construct renewal times is the following:

Definition (small set): $A \in \mathcal{B}(\mathcal{X})$ is a small set if there exist $\varepsilon > 0$ and a non zero measure on $(\mathcal{X}, \mathcal{B}(\mathcal{X}))$ such that

$$\forall x \in A \qquad \mathbb{P}(X_{n+1} \in B / X_n = x) = P(x, B) \ge \varepsilon \nu(B) \quad \forall B \in \mathcal{B}(\mathcal{X}).$$

(or more generally $P^{(n)}(x, B) \ge \varepsilon \nu(B)$ $\forall B \in \mathcal{B}(\mathcal{X}), x \in A$ and some $n \ge 1$).

If a ψ-irreducible Markov chain $\{X_n\} \sim \text{Markov}(\lambda, P)$ has a recurrent small set A, with $\psi(A) > 0$, (i.e. $\mathbb{P}(\tau_A < +\infty) = 1$), then it is possible to construct a modified chain $\{\tilde{X}_n\}$ with the same law as $\{X_n\}$ and a sequence of renewal times. More precisely:

if $\tilde{X}_n \notin A$, $\quad \mathbb{P}(\tilde{X}_{n+1} \in B / \tilde{X}_n) = P(\tilde{X}_n, B)$ same as for X_n.

if $\tilde{X}_n \in A$, $\quad \mathbb{P}(\tilde{X}_{n+1} \in B / \tilde{X}_n) = \begin{cases} \nu(B) & \text{with probab. } \varepsilon \\ \dfrac{P(\tilde{X}_n, B) - \varepsilon \nu(B)}{1-\varepsilon} & \text{with prob. } 1-\varepsilon \end{cases}$

i.e. when $\tilde{X}_n \in A$, $\tilde{X}_{n+1}$ is drawn from the mixture distribution $\tilde{P}(\tilde{X}_n, B) = \varepsilon \nu(B) + (1-\varepsilon) \frac{P(\tilde{X}_n, B) - \varepsilon \nu(B)}{1-\varepsilon}$ which obviously coincides with $P(\tilde{X}_n, B)$ so the distribution of the chain is unchanged.

However, we can now define the stopping times

$\tilde{\tau}_A^{(r)} = \inf \{ n > \tilde{\tau}_A^{(r-1)} : \tilde{X}_n \in A \text{ and } \tilde{X}_{n+1} \sim \nu \}$ $\quad \tilde{\tau}_A^{(0)} = 0$.

and since the distribution of $\tilde{X}_{\tilde{\tau}_A^{(r)}+1}$ does not depend on $\tilde{X}_{\tilde{\tau}_A^{(r)}}$ (but only on $\tilde{X}_{\tilde{\tau}_A^{(r)}}$ being in A), $\tilde{\tau}_A^{(r)}$ are renewal times and, if A is recurrent, the invariant measure can be constructed as

$$\hat{\pi}(C) = E_\nu \Big[\sum_{n=0}^{\tilde{\tau}_A^{(1)}-1} \mathbb{1}_{\{\tilde{X}_n \in C\}} \Big] = \sum_{n=0}^{\infty} P_\nu (\tilde{X}_n \in C, \tilde{\tau}_A^{(1)} > n). \quad (*)$$

the following result holds, which shows the equivalence between existence of small sets and φ-recurrence.

Lemma: If $\{X_n\}$ is φ-recurrent for some $\varphi \neq 0$, then every set $E \subset \mathcal{B}(X)$ with $\varphi(E) > 0$ contains a small set $A \subset E$, $\varphi(A) > 0$, and the measure $\nu(\cdot)$ is given by $\nu(B) = \varphi(A \cap B)/\varphi(A)$.

Conversely, if $\{X_n\}$ is φ-irreducible and has a small set A, with $\varphi(A) > 0$, such that $P_x(\tau_A < +\infty) = 1 \ \forall x \in X$, with minorizing measure ν, then $\{X_n\}$ is ν-recurrent.

From the previous arguments it follows that

Theorem: if $\{X_n\}$ is φ-recurrent, then there exists a unique invariant measure $\hat{\pi}$ (which can be constructed as in $(*)$), up to a multiplicative factor, and $\varphi \ll \mu$.

If the measure $\hat{\pi}$ can be normalized to make it a probability measure on $(X, \mathcal{B}(X))$, which happens if and only if $E_\nu[\tilde{\tau}_A^{(1)}] < \infty$

where A is a small set with minorizing measure ν, then the invariant probability measure π satisfies $\pi(C) = \frac{\hat{\pi}(C)}{\mathbb{E}_\nu[\tilde{\tau}_A^{(1)}]}$.
In this case, the chain is called positive ψ-recurrent. It can be shown that a positive ψ-recurrent chain is also ψ-recurrent.

<u>Theorem</u>: If $\{X_n\}$ is positive ψ-recurrent then for any λ and any f: $\mathbb{E}_\pi[|f|] < \infty$, $\mathbb{P}_\lambda\left(\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} f(X_i) = \mathbb{E}_\pi[f]\right) = 1$.

<u>Definition</u>: a ψ-irreducible chain $\{X_n\}$ is called aperiodic if there exists a small set C, $\psi(C) > 0$, and $\bar{n} \in \mathbb{N}$ such that
$\mathbb{P}(\tilde{X}_{n+m} \in C, \tilde{X}_{n+m+1} \sim \nu \,/\, \tilde{X}_n \in C, \tilde{X}_{n+1} \sim \nu) > 0 \quad \forall\, m > \bar{n}$

<u>Theorem</u>: If $\{X_n\}$ is positive ψ-recurrent and aperiodic then
$\lim_{n\to\infty} \| \pi^{n,\lambda} - \pi \|_{TV} = \lim_{n\to\infty} \| \int P^{(n)}(x,\cdot)\,\lambda(dx) - \pi(\cdot) \|_{TV} = 0$ for any λ.

<u>Definition</u>: A chain $\{X_n\}_{n=0}^{N} \sim \text{Markov}(\lambda, P)$ is reversible if the chain $\{Y_n = X_{N-n}\}_{n=0}^{N} \sim \text{Markov}(\lambda, P)$.

$\{X_n\}$ is reversible if and only if (λ, P) are in detailed balance, i.e.

$$\int_A P(x,B)\,\lambda(dx) = \int_B P(y,A)\,\lambda(dy) \quad \forall A,B \in \mathcal{B}(\mathcal{X}),\ \lambda(A), \lambda(B) > 0$$

If this happens, then λ is an invariant distribution. Indeed

$$\int_{\mathcal{X}} P(x,B)\,\lambda(dx) = \int_B \underbrace{P(y,\mathcal{X})}_{=1}\,\lambda(dy) = \lambda(B).$$

Markov chain Monte Carlo / Metropolis-Hastings

Suppose we want to simulate a sample from a given probability density function $f: x \in \mathbb{R}^d \to \mathbb{R}_+$ with complicated form, so that direct methods as those discussed in chapter 2 are not viable. Moreover, it is not uncommon in applications that the available density f is not normalized, that is the probability density is $\tilde{f} = f / \int f$ but the normalization constant is not easily computable. Typical examples are:

Bayesian statistics: let $\vec{x} = (x_1, \dots, x_n)$ be an iid sample from a parametric density $g(x/\theta)$. Then the joint density of $\vec{x}$ given θ is $g(\vec{x}/\theta) = \prod_{i=1}^{n} g(x_i/\theta)$ and we want to estimate θ from the data $\vec{x}$. In the Bayesian paradigm, θ is thought as a random variable itself, with prior density $\pi(\theta)$ that summarizes any prior information on θ in the absence of data. Then, the posterior density of θ given the data is

$$f(\theta) = \frac{1}{z(\vec{x})} g(\vec{x}/\theta)\, \pi(\theta) \qquad \text{with } z(\vec{x}) = \int g(\vec{x}/\theta)\, \pi(\theta)\, d\theta$$

which is often unknown and difficult to compute.

Statistical physics. Let $x \in X$ be a configuration of a physical system on X the configuration space. Let $H: X \to \mathbb{R}$ be an energy function and T the temperature. Then the probability density of finding the system in a given state x is

$$f(x) = \frac{1}{z} \exp\{ -H(x)/kT \}$$

where k is the Boltzmann constant and $z = \int e^{-H(x)/kT} dx$ the partition function, often difficult to compute.

The idea of Markov chain Monte Carlo (MCMC) is to generate a Markov chain $\{X_n\}$ that has invariant distribution f. Hence, if the chain is recurrent and aperiodic, as $n \to \infty$ $X_n \sim f$ and expectations w.r.t. f can be computed using the ergodic theorem $\mathbb{E}_f[\varphi] = \int \varphi(x) f(x) dx \approx \frac{1}{n} \sum_{j=1}^{n} \varphi(X_j)$.

Metropolis-Hastings algorithm

A very general strategy to achieve the goal above is offered by the Metropolis-Hastings (MH) algorithm: let $f : \mathcal{X} \to \mathbb{R}_+$ be the target density and take a transition kernel Q with density $q : \mathcal{X} \times \mathcal{X} \to \mathbb{R}_+$, i.e. $Q(x, A) = \int_A q(x,y) dy \quad \forall x \in \mathcal{X}, A \in \mathcal{B}(\mathcal{X})$. A Markov chain with transition kernel Q will not have invariant distribution f, in general, so at each step we have to correct the new state by using an Acceptance Rejection condition so that at the end the chain has the desired invariant distribution. Let us define the function $\alpha : \mathcal{X} \times \mathcal{X} \to [0,1]$

$$\alpha(x,y) = \min \left\{ \frac{f(y)}{f(x)} \cdot \frac{q(y,x)}{q(x,y)}, 1 \right\}.$$ Given X_n, the algorithm

- generates $Y_{n+1} \sim q(X_n, \cdot)$ proposal state
- with probability $\alpha(X_n, Y_{n+1})$, accepts the proposal $\to X_{n+1} = Y_{n+1}$
- " $1 - \alpha(X_n, Y_{n+1})$, rejects the proposal $\to X_{n+1} = X_n$.

the transition density is often called the proposal or instrumental density. If q is symmetric, the formula for α simplifies as $\alpha(x,y) = \min \left\{ \frac{f(y)}{f(x)}, 1 \right\}$ which shows that if the proposal state Y_{n+1} is more likely than X_n, it will always be accepted, whereas

if it is less likely, it will only be accepted with probability $\alpha(X_n, Y_{n+1}) < 1$. If Y_{n+1} is rejected, the chain does not move, i.e. $X_{n+1} = X_n$. the probability $\alpha(x,y)$ is called the acceptance probability.

Algorithm (Metropolis-Hastings)

Given an initial distribution λ and a proposal q

- generate $X_0 \sim \lambda$
- for $n = 0, 1, \dots$
  - generate $Y_{n+1} \sim q(X_n, \cdot)$ proposal state
  - generate $U \sim \mathcal{U}(0,1)$

    if $U \leq \alpha(X_n, Y_{n+1})$ set $X_{n+1} = Y_{n+1}$ (accept proposal)

    otherwise set $X_{n+1} = X_n$ (reject proposal)

For the algorithm to work, the chain has to be able to explore the whole density f. Let us denote $D_f = \text{supp}(f) = \{x \in \mathcal{X} : f(x) > 0\}$ the support of f. Minimum requirements are:

- $X_0 \in D_f$ (otherwise $\alpha(X_0, \cdot)$ is not defined), so that $X_n \in D_f \ \forall n$
- $\bigcup_{x \in D_f} \text{supp}(q(x, \cdot)) \supset D_f$ (otherwise the chain fails to visit some parts of D_f).

We present here two common strategies to build a proposal density q:

- Independent samples

  let $g : \mathcal{X} \to \mathbb{R}_+$ be a probability density function and $g(x) > 0$ whenever $f(x) > 0$ (i.e. $f \ll g$)

  We choose simply $q(x,y) = g(y)$ independently of the current state x.

Algorithm (Independent Metropolis-Hastings)

Given $X_0 \sim \lambda$, $\text{supp}(\lambda) \subset D_f$

for $n = 0, 1, \dots$

- generate $Y_{n+1} \sim g$
- set $X_{n+1} = \begin{cases} Y_{n+1} & \text{with prob. } \alpha(X_n, Y_{n+1}) = \min\left\{ \frac{f(Y_{n+1})}{f(X_n)} \cdot \frac{g(X_n)}{g(Y_{n+1})}, 1 \right\} \\ X_n & \text{otherwise} \end{cases}$

- Random walk Metropolis

let g_σ be a probability density (σ being a scaling parameter)
typically $g_\sigma = N(0, \sigma^2)$
then, we choose $q(x,y) = g_\sigma(y-x)$ i.e. the proposal distribution is g_σ centered in the current state x.
the choice of σ is rather delicate. Small σ implies small steps from the current state, hence high correlation of the chain. Large steps might lead to high rejection rate, hence the chain will stay for long time in a given state which also leads to high correlation in the chain.

We study now the transition kernel P (resp. density p) of the Markov chain generated by the Metropolis-Hastings algorithm.
There is a non zero probability that $X_{n+1} = X_n$, so $P(X_n, \cdot)$ has a part mass in X_n:

$$P(X_{n+1} = x \,/\, X_n = x) = \int_x q(x,y)\,(1 - \alpha(x,y))\,dy = 1 - \int_x \alpha(x,y) q(x,y)\,dy$$

so the Transition density p is

$$p(x,y) = \alpha(x,y) q(x,y) + (1 - \alpha^*(x)) \delta_x(y) \qquad \alpha^*(x) = \int \alpha(x,y) q(x,y)\,dy.$$

where $\delta_x(y)$ is a Dirac mass in x.

The key result is that P and f satisfy the detailed balance equation. We recall that for a kernel $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ and a probability distribution $\pi : \mathcal{B}(\mathcal{X}) \to [0,1]$, we say that (P, π) are in detailed balance if $\forall A, B \in \mathcal{B}(\mathcal{X})$, $\pi(A), \pi(B) > 0$

$$\int_A P(x, B)\, \pi(dx) = \int_B P(y, A)\, \pi(dy).$$

If this happens, then π is an invariant distribution. Indeed

$$\int_{\mathcal{X}} P(x, B)\, \pi(dx) = \int_B \underbrace{P(y, \mathcal{X})}_{=1}\, \pi(dy) = \pi(B).$$

<u>Lemma</u> The transition kernel P of the Metropolis-Hastings algorithm, with density $p(x,y) = \alpha(x,y) q(x,y) + (1 - \alpha^*(x)) \delta_x(y)$ is in detailed balance with the measure π with density f. Hence f is an invariant probability density for P.

<u>Proof</u> : Observe first that

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

$$= \min\{ f(y) q(y,x),\ f(x) q(x,y) \} = f(y) q(y,x) \alpha(y,x)$$

Hence,

$$\int_A P(x, B) f(x) dx = \int_A \left(\int_B (\alpha(x,y) q(x,y) + (1 - \alpha^*(x)) \delta_x(y)) dy \right) f(x) dx$$

$$= \int_A \int_B f(y) \alpha(y,x) q(y,x) dy\, dx + \int_{A \cap B} (1 - \alpha^*(x)) f(x) dx$$

$$= \int_B \left(\int_A (\alpha(y,x) q(y,x) + (1 - \alpha^*(y)) \delta_y(x)) dx \right) f(y) dy = \int_B P(y, A) f(y) dy.$$

□

To assess the convergence to equilibrium of the chain, we should further check ψ-irreducibility, aperiodicity and recurrence. These properties should actually be checked w.r.t. the density f.

- f-irreducibility is something that should be checked every time depending on the choice of the proposal density.
- Concerning f-recurrence, the following result holds:

<u>Theorem</u>: If the Metropolis-Hastings chain is f-irreducible, then it is Harris positive recurrent and $\forall \varphi: \mathcal{X} \to \mathbb{R}$, $\mathbb{E}_f[|\varphi|] < +\infty$

$$\lim_{n\to\infty} \frac{1}{n} \sum_{j=1}^{n} \varphi(X_j) = \mathbb{E}_f[\varphi] = \int_{\mathcal{X}} \varphi(x) f(x) dx$$

- Concerning aperiodicity, observe that in general $\mathbb{P}(X_{n+1} = x \,/\, X_n = x) > 0$ as long as $\alpha^*(x) < 1$, since the transition kernel $P(x,\cdot)$ has an atom at x. Consider the set $C = \{x : \alpha(x) < 1\}$. This is a f-zero measure set, i.e. $\int_C f(x)dx = 0$ if and only if (verify as exercise)

$$f(x)\, q(x,y) = f(y)\, q(y,x) \quad \text{for } f\text{-almost every } x, y \in \mathcal{D} \qquad (*)$$

which corresponds to the case in which the proposal q is in detailed balance with f (hence the acceptance rejection step is useless and we should check the aperiodicity of q) Therefore, if (q, f) are <u>not</u> in detailed balance and the chain is f-irreducible (and therefore Harris recurrent), there exists a small set $E \subset \mathcal{D}$, $\int_E f(x)dx > 0$, on which $P(x,E) > 0$ $\forall x \in E$ and the chain is <u>aperiodic</u> and for any initial distribution $\lambda \ll f$ we have (denoting π_f the measure associated to f)

$$\lim_{n\to\infty} \| \pi^{n,\lambda} - \pi \|_{TV} = \lim_{n\to\infty} \sup_{B \in \mathcal{B}(\mathcal{X})} \left| \int_{\mathcal{X}} P^{(n)}(x,B)\, \lambda(dx) - \int_B f(x)dx \right| = 0.$$

Independent sampler

We recall that in the independent sampler case, $q(x,y) = g(y)$ independent of x, with $f \ll g$ (i.e. $\int_B f(y)dy > 0 \Rightarrow \int_B g(y)dy > 0$)

Algorithm

- generate $X_0 \sim \lambda$
- for $n = 0, 1, \dots$
  - generate $Y_{n+1} \sim g$ and compute $\alpha(X_n, Y_{n+1}) = \min\left\{\frac{f(Y_{n+1})\,g(X_n)}{f(X_n)\,g(Y_{n+1})}, 1\right\}$
  - generate $U \sim \mathcal{U}(0,1)$ and set

$$X_{n+1} = \begin{cases} Y_{n+1} & \text{if } U \le \alpha(X_n, Y_{n+1}) \\ X_n & \text{otherwise} \end{cases}$$

Concerning the convergence to equilibrium, we recall first a useful result for general state space Markov chains.

Lemma: Let $P : \mathcal{X} \times \mathcal{B}(\mathcal{X}) \to [0,1]$ be a Markov transition kernel with π invariant distribution. If there exist $\varepsilon \in (0,1)$ and a probability measure ν on $\mathcal{X}$ such that $P(x,A) \ge \varepsilon \nu(A)$ $\forall x \in \mathcal{X}$, $\forall A \in \mathcal{B}(\mathcal{X})$ (uniform minorizing condition), then $\| \pi^{n,\lambda} - \pi \|_{TV} \le 2(1-\varepsilon)^n$ where $\pi^{n,\lambda}(A) = \int_A P^{(n)}(x,A)\,\lambda(dx)$.

More generally, if there exist $k_0 \in \mathbb{N}^*$ such that $P^{(k_0)}(x,A) \ge \varepsilon \nu(A)$ $\forall x \in \mathcal{X}$, $\forall A \in \mathcal{B}(\mathcal{X})$, then $\| \pi^{n,\lambda} - \pi \|_{TV} \le 2(1-\varepsilon)^{\lfloor n/k_0 \rfloor}$.

Idea of the proof for $k_0 = 1$: we build the following two coupled chains

- let $X_0 \sim \lambda$, $Y_0 \sim \pi$
- for $n = 0, 1, \dots$
  - draw $Z_n \sim \text{Bernoulli}(\varepsilon)$ $\quad P(Z_n = 1) = \varepsilon$, $P(Z_n = 0) = 1 - \varepsilon$
    if $Z_n = 1$ draw $W \sim \nu$ and set $X_{n+1} = Y_{n+1} = W$
    otherwise, draw $X_{n+1} \sim \frac{P(X_n, \cdot) - \varepsilon \nu(\cdot)}{1-\varepsilon}$ $\quad Y_{n+1} \sim \frac{P(Y_n, \cdot) - \varepsilon \nu(\cdot)}{1-\varepsilon}$ independently

Clearly $\{X_n\} \sim \text{Markov}(\lambda, P)$ and $\{Y_n\} \sim \text{Markov}(\pi, P)$.

Let $T = \inf\{n \geq 0 : Z_n = 1\}$. It is also clear that after T, the two chains have the same distribution $X_n \sim Y_n$ $n \geq T$.

Moreover, $\mathbb{P}(T \geq n) = (1-\varepsilon)^n$. Now

$$\begin{aligned}
\|\pi^{n,\lambda} - \pi\|_{TV} &= 2 \sup_{A \in \mathcal{B}(\mathcal{X})} |\mathbb{P}(X_n \in A) - \mathbb{P}(Y_n \in A)| \\
&= 2 \sup_A |\mathbb{P}(X_n \in A, T < n) + \mathbb{P}(X_n \in A, T \geq n) - \mathbb{P}(Y_n \in A, T < n) - \mathbb{P}(Y_n \in A, T \geq n)| \\
&= 2 \sup_A |\mathbb{P}(X_n \in A, T \geq n) - \mathbb{P}(Y_n \in A, T \geq n)| \\
&\leq 2 \sup_A \mathbb{P}(X_n \in A, Y_n \in A, T \geq n) \leq 2\mathbb{P}(T \geq n) \leq 2(1-\varepsilon)^n.
\end{aligned}$$

□

In the case of independent sampler, the following result holds:

<u>Theorem</u>: If $\exists M < +\infty$ such that $f(x) \leq M g(x)$ $\forall x \in \mathcal{X}$, then the chain generated by the independence sampler algorithm is geometrically ergodic and

$$\|\pi^{n,\lambda} - \pi\|_{TV} \leq \left(1 - \frac{\int f(x)dx}{M}\right)^n \quad \text{for any } \lambda.$$

<u>Proof</u>: If f is not normalized, let $\tilde{f} = f/c$, $c = \int f$. Notice that

$$\alpha(x,y) q(x,y) = g(y) \min\left\{\frac{f(y)}{f(x)}\frac{g(x)}{g(y)}, 1\right\} = \min\left\{f(y)\frac{g(x)}{f(x)}, g(y)\right\} \geq \frac{1}{M} f(y)$$

($\frac{g(x)}{f(x)} \geq 1/M$, $g(y) \geq \frac{1}{M} f(y)$)

It follows that $\forall A \in \mathcal{B}(\mathcal{X})$

$$P(x,A) = \int_A p(x,y)\,dy = \int_A \left(\alpha(x,y) q(x,y) + (1 - \alpha^+(x))\delta_x(y)\right) dy \geq \frac{1}{M}\int_A f(y)\,dy \geq \frac{c}{M}\pi(A)$$

and the result follows from the previous Lemme. □

Moreover, under the same condition $f(x) \leq M g(x)$ $\forall x \in \mathcal{X}$, it can be shown that the expected acceptance probability satisfies $\mathbb{E}[\alpha(X_n, Y_{n+1})] \geq \frac{c}{M}$ (exercise). This result has to be compared with a pure acceptance-rejection sampling strategy, for which the expected acceptance probability $= c/M$. Hence, independent MH sampler accepts more often than a pure acceptance-rejection sampler.

Random walk Metropolis Hastings

We recall that in this case $q(x,y) = g_\sigma(y-x)$. If we assume $g_\sigma(\cdot)$ symmetric, the acceptance probability takes the simplified form $\alpha(x,y) = \min\left\{\frac{f(y)}{f(x)}, 1\right\}$.

Concerning convergence of this algorithm, one could try to verify a uniform minorizing condition

$$g_\sigma(y-x) \geq \varepsilon f(y) \qquad \forall x, y \in \mathcal{D}_f.$$

By the same arguments as for independent sampler, this would imply $P(x,A) \geq \varepsilon \pi(A)$ hence uniform geometric convergence $\|\pi^{n,\lambda} - \pi\|_{TV} \leq 2(1-\varepsilon)^n \quad \forall \lambda$.

However, such uniform minorizing condition does not hold for unbounded / non-compact $\mathcal{X}$.

We mention a result by Mengersen & Tweedie ('96) showing geometric ergodicity for tail-log-concave f for $\mathcal{X} = \mathbb{R}$.

Definition a probability density f on $\mathbb{R}$ is log-concave in the tails if $\exists \alpha, M > 0$ s.t. $\log f(x) - \log f(y) \geq \alpha(|y| - |x|) \quad \forall |y| \geq |x| \geq M$.

Theorem: If the invariant density f on $\mathbb{R}$ is log concave in tails for some $\alpha, M > 0$ and $\inf_{|x| \leq R} f(x) > 0 \; \forall R > 0$, then the Markov chain generated by the random walk Metropolis-Hastings algorithm with symmetric proposal $g_\sigma(\cdot)$ is geometrically ergodic.

One variable at the time Metropolis-Hastings

Suppose that a state $x \in \mathcal{X}$ has several components, $x = (x^{(1)}, \dots, x^{(d)})$, $x^{(i)} \in \mathcal{X}^{(i)}$.

One can thus construct a Metropolis-Hasting algorithm by updating one component at the time, either chosen randomly or by performing a systematic scan over the components.

Say that the i-th component has been chosen. We use the notation $x = (x^{(i)}, x^{(-i)})$ with $x^{(-i)} = (x^{(1)}, \dots, x^{(i-1)}, x^{(i+1)}, \dots, x^{(d)})$.

Let $q_i : \mathcal{X}^{(i)} \times \mathcal{X}^{(i)} \to \mathbb{R}$ be a proposal density function on $\mathcal{X}^{(i)}$. Then the one variable at the time MH algorithm reads

Algorithm (random scan)

- generate $X_0 \sim \lambda$
- for $n = 0, 1, \dots$
  - draw index $i_n \sim \beta$ (prob. mass function on $\{1, \dots, d\}$)
  - draw $y^{(i_n)} \sim q_{i_n}(X_n^{(i_n)}, \cdot)$ and set $Y_{n+1} = (y^{(i_n)}, X_n^{(-i_n)})$
  - compute $\alpha_{i_n}(X_n, Y_{n+1}) = \min\left\{ \frac{f(Y_{n+1})\, q_{i_n}(Y_{n+1}^{(i_n)}, X_n^{(i_n)})}{f(X_n)\, q_{i_n}(X_n^{(i_n)}, Y_{n+1}^{(i_n)})}, 1 \right\}$

  set $X_{n+1} = \begin{cases} Y_{n+1} & \text{with prob. } \alpha_{i_n}(X_n, Y_{n+1}) \\ X_n & \text{otherwise} \end{cases}$

  end

Algorithm (fixed scan)

- generate $X_0 \sim \lambda$
- for $n = 0, 1, \dots$

  set $Y_{n+1,0} = X_n$

  for $i = 1, \dots, d$

  - draw $y^{(i)} \sim q_i(X_n^{(i)}, \cdot)$ and set $\tilde{Y} = (y^{(i)}, Y_{n+1,i-1}^{(-i)})$

- set $Y_{n+1,i} = \begin{cases} \tilde{Y} & \text{with prob. } \alpha_i(Y_{n+1,i-1}, \tilde{Y}) \\ Y_{n+1,i-1} & \text{otherwise} \end{cases}$

end

$X_{n+1} = Y_{n+1,d}$

end

Gibbs sampler

The gibbs sampler is a "one variable at the time" MH algorithm in which the "componentwise" proposal density is the conditional density $q_i(x,\cdot) = f(\cdot / x^{(-i)})$, i.e. the i-th component is drawn independently from $f(\cdot / x^{(-i)})$.

Observe that, in this case, the hasting ratio for $X = (X^{(i)}, X^{(-i)})$ and $Y = (Y^{(i)}, X^{(-i)})$ looks

$$\alpha_i(X,Y) = \min\left\{ \frac{f(Y)\, f(X^{(i)} / X^{(-i)})}{f(X)\, f(Y^{(i)} / X^{(-i)})}, 1 \right\} = 1$$

i.e. the move is always accepted or, in other words, the transition kernel Q_i which samples independently the i-th component from the conditional density $f(\cdot / x^{(-i)})$ preserves the density f.

Algorithm (Gibbs with random scan)

- generate $X_0 \sim \lambda$
- for $n = 0, 1, \ldots$
  - draw i_n from a pmf β on $\{1, \ldots, d\}$
  - generate $y^{(i_n)} \sim f(\cdot / X_n^{(-i_n)})$
  - set $X_{n+1} = (y^{(i_n)}, X_n^{(-i_n)})$

end

Convergence diagnostics

Let us consider an f-irreducible Metropolis-Hastings Markov chain, which is in particular Harris recurrent and ergodic. Given any function $\varphi : \mathbb{E}_f[\varphi] < +\infty$, by the ergodic theorem

$$\lim_{n \to \infty} \frac{1}{n} \sum_{j=1}^{n} \varphi(X_j) = \mathbb{E}_f[\varphi].$$

Hence we can consider the estimator $\hat{\mu}_N^{MH}$ for $\mu = \mathbb{E}_f[\varphi]$:

$$\hat{\mu}_N^{MH} = \frac{1}{N} \sum_{j=1}^{N} \varphi(X_j).$$

The question is how to monitor properly the convergence of $\hat{\mu}^{MH}$ to μ and how to choose N.

The estimator $\hat{\mu}_N^{MH}$ is biased, in general, since $X_n \sim f$ only asymptotically as $n \to \infty$. The bias is generally of order $1/N$.

Lemma: Let $\{X_n\} \sim \text{Markov}(\delta_x, P)$ with P a Metropolis-Hastings transition kernel with invariant distribution π (and density f). If $\{X_n\}$ is geometrically ergodic, i.e. $\exists\, \gamma > 0$ and $h : \mathcal{X} \to \mathbb{R}_+$ s.t. $\| \pi^{n,\delta_x} - \pi \|_{TV} \leq h(x) e^{-\gamma n}$, then for any bounded $\varphi : \mathcal{X} \to \mathbb{R}$ $\exists\, C_\varphi > 0$ s.t. $\left| \mathbb{E}[\hat{\mu}_N^{MH}] - \mu \right| \leq \frac{C}{N}$.

Proof:

$$\left| \mathbb{E}[\hat{\mu}_N^{MH}] - \mu \right| = \left| \frac{1}{N} \sum_{j=1}^{N} \mathbb{E}[\varphi(X_j) - \mu] \right| \leq \frac{1}{N} \sum_{j=1}^{N} \left| \int_{\mathcal{X}} \varphi(y) (\pi^{n,x}(dy) - \pi(dy)) \right|$$

$$\leq \frac{1}{N} \sum_{j=1}^{N} \sup_{x \in \mathcal{X}} |\varphi(x)| \, \| \pi^{n,x} - \pi \|_{TV}$$

$$\leq \frac{1}{N} \sup_{x \in \mathcal{X}} |\varphi(x)| \, h(x) \frac{1}{1 - e^{-\gamma}}$$

∎

Such bias can be further reduced by considering the estimator $\hat{\mu}_{N,B}^{MH} = \frac{1}{N} \sum_{j=B+1}^{N+B} \varphi(X_j)$, i.e. by disregarding the first B terms. The lag B is often called the "burn-in" or "warm-up" period.

Under the assumptions of the previous Lemma, the bias of the estimator $\hat{\mu}^{MH}_{N,B}$ is bounded by $\left|\mathbb{E}[\hat{\mu}^{MH}_{N,B}] - \mu\right| \leq \frac{e^{-\gamma B}}{N} \sup_{x \in X} |\varphi(x)| \frac{h(x)}{1-e^{-\gamma}}$

and is thus reduced by a factor $e^{-\gamma B}$ w.r.t. the base estimator $\hat{\mu}^{MH}_N = \hat{\mu}^{MH}_{N,0}$.

The exponential decay rate $1/\gamma$ is often called the "relaxation time" and choosing $B = m/\gamma$ with moderate m makes the bias negligible.

The variance of $\hat{\mu}^{MH}_N$, on the other hand, is of the order $1/N$, hence of lower order than $(\text{bias})^2$.

Consider the ideal case of $X_0 \sim \pi$ so that $X_n \sim \pi$ $\forall n$ and the chain is at equilibrium. In such case, $\hat{\mu}^{MH}_N$ is unbiased.

Let us denote $c(k) = \mathrm{Cov}_\pi(\varphi(X_0), \varphi(X_k)) = \mathrm{Cov}_\pi(\varphi(X_j), \varphi(X_{j+k}))$ $\forall j$ thanks to the stationarity of the chain.

Lemma: Let $\{X_n\} \sim \text{Markov}(\pi, P)$ and $\sum_{k=0}^{\infty} |c(k)| < +\infty$. Then

$$\lim_{N \to \infty} N \,\mathrm{Var}\left(\hat{\mu}^{MH}_N\right) = \sigma^2 \qquad \text{with } \sigma^2 = c(0) + 2\sum_{k=1}^{\infty} c(k).$$

Proof:

$$\mathrm{Var}\left(\hat{\mu}^{MH}_N\right) = \mathbb{E}\left[\frac{1}{N}\sum_{j=1}^{N}(\varphi(X_j)-\mu)\right] = \frac{1}{N^2}\sum_{j=1}^{N}\sum_{k=1}^{N}\mathbb{E}[(\varphi(X_j)-\mu)(\varphi(X_k)-\mu)]$$

$$= \frac{1}{N^2}\left[\sum_{j=1}^{N}\underbrace{\mathrm{Var}(\varphi(X_j))}_{c(0)} + 2\sum_{j=1}^{N-1}\sum_{k=j+1}^{N}\underbrace{\mathrm{Cov}(\varphi(X_j),\varphi(X_k))}_{c(k-j)}\right]$$

$$= \frac{c(0)}{N} + \frac{2}{N^2}\sum_{j=1}^{N-1}\sum_{\ell=1}^{N-j} c(\ell)$$

$$= \frac{c(0)}{N} + \frac{2}{N}\sum_{\ell=1}^{N-1}\frac{N-\ell}{N}c(\ell) = \frac{1}{N}\left[c(0) + 2\sum_{\ell=1}^{N-1}\left(1-\frac{\ell}{N}\right)c(\ell)\right]$$

Under the assumption $\sum_{\ell=0}^{\infty}|c(\ell)|<+\infty$, it follows $\lim_{N\to\infty} N\,\mathrm{Var}\left(\hat{\mu}^{MH}_N\right) = \sigma^2$. ∎

The quantity σ^2 is called "time average variance constant" (TAVC)
If $\{X_n\}$ were iid $\sim \pi$, then the variance of a Crude Monte Carlo estimator $\hat{\mu}_N^{MC} = \frac{1}{N}\sum_{j=1}^{N} \varphi(X_j)$ would be $\mathrm{Var}(\hat{\mu}_N^{MC}) = \frac{c(0)}{N}$.

From this we see that $\lim_{N\to\infty} \frac{\mathrm{Var}(\hat{\mu}_N^{MH})}{\mathrm{Var}(\hat{\mu}_N^{MC})} = \frac{\sigma^2}{c(0)} = 1 + 2\sum_{k=1}^{\infty} \frac{c(k)}{c(0)}$

Hence $\hat{\mu}^{MH}$ is less effective than a pure iid sampling from π, due to the correlation in the chain.
For the estimator $\hat{\mu}_N^{MH}$ a CLT is also available (and more generally for aperiodic, irreducible and reversible chains with invariant distribution π).

Theorem (CLT for Metropolis-Hastings Markov chains)
Let $\{X_n\}$ be an f-irreducible, aperiodic Metropolis-Hastings chain, with invariant dist. π (resp. density f), and $\varphi: \mathcal{X} \to \mathbb{R}$ such that

$$\sigma^2 := \mathrm{Var}_\pi(\varphi(X_0)) + 2\sum_{\ell=1}^{\infty} \mathrm{Cov}_\pi(\varphi(X_0), \varphi(X_\ell)) < +\infty .$$

Then, $\sqrt{N}\,(\hat{\mu}_N^{MH} - \mu) \xrightarrow[N\to\infty]{d} N(0, \sigma^2)$.

From the CLT, asymptotic confidence intervals can be derived.
The practical question, however, is how to estimate σ^2.

Estimation of the asymptotic variance
We recall the formula $\sigma^2 = c(0) + 2\sum_{k=1}^{\infty} c(k)$. Given a path $\{X_n\}_{n=0}^{N}$, if we discard a sufficient burn-in lag B, we can reasonably assume that $\{X_n\}_{n=B+1}^{N}$ is (nearly) stationary, so that a sample estimator for $c(k)$ is

$$\hat{c}(k) = \frac{1}{N-B-k-1} \sum_{j=B+1}^{N-k} \left(\varphi(X_j) - \hat{\mu}^{MH}_{N,B}\right)\left(\varphi(X_{j+k}) - \hat{\mu}^{MH}_{N,B}\right)$$

and an estimator for σ^2 is

$$\hat{\sigma}^2 = \hat{c}(0) + 2 \sum_{k=1}^{N-B-2} \hat{c}(k).$$

However, the last terms in the sum are very unstable since they are sample averages of very few terms. It is often wiser to truncate the sum much earlier

$$\hat{\sigma}^2_M = \hat{c}(0) + 2 \sum_{k=1}^{M} \hat{c}(k) \qquad M < N-B-2$$

It has been shown [Geyer '92] that the sequence $\Gamma_k = c(2k) + c(2k+1)$ is strictly positive, decreasing and convex for a reversible Markov chain. Hence a good choice is

$$M = \max\{k : \hat{\Gamma}_j > 0 \ \ \forall j \le k\} = \min\{k : \hat{\Gamma}_k < 0\} - 1$$

Batch means

An alternative idea to estimate σ^2 is to split the sequence $\{X_n\}_{n=B+1}^{N}$ in M blocks of size $T = (N-B)/M$ (assumed to be an integer). Then we can build M different sample averages

$$Y_i = \frac{1}{T} \sum_{j=(i-1)T+B+1}^{iT+B} \varphi(X_j) \qquad \text{and} \quad \hat{\mu}^{MH}_{N,B} = \frac{1}{M} \sum_{i=1}^{M} Y_i$$

If T is sufficiently large (larger than the relaxation time), the M blocks are nearly independent, so $\mathrm{Var}(\hat{\mu}^{MH}_{N,B}) \simeq \frac{\mathrm{Var}(Y_i)}{M}$ and an estimator for $\sigma^2 \simeq (N-B)\,\mathrm{Var}(\hat{\mu}^{MH}_{N,B})$ is

$$\hat{\sigma}^2_{batch} = \frac{N-B}{M} \hat{\sigma}^2_Y \qquad \hat{\sigma}^2_Y = \frac{1}{M-1} \sum_{i=1}^{M} \left(Y_i - \hat{\mu}^{MH}_{N,B}\right)^2.$$

Rigenerative method

We have seen that the proof of convergence for a Markov chain $\{X_n\} \sim \text{Markov}(\lambda, P)$ relies on the existence of renewal times $\tau^{(r)}$, $r = 1, 2, \dots$ This idea can also be used to monitor the convergence. One could indeed generate a path $\{X_n\}_{n=0}^{\tau^{(N+1)}-1}$ up to the $N+1$ renewal time. Such path will contain then N iid blocks $\{X_n\}_{n=\tau^{(i)}}^{\tau^{(i+1)}-1}$ $\quad i = 1, \dots, N$.

Setting $T_i = \tau^{(i+1)} - \tau^{(i)}$ and $Y_i = \sum_{j=\tau^{(i)}}^{\tau^{(i+1)}-1} \varphi(X_j)$

then $\mathbb{E}[Y_i] = \mu\, \mathbb{E}[T_i]$ and an estimator for μ is

$$\hat{\mu}_{Reg} = \frac{\bar{Y}}{\bar{T}} \qquad \bar{Y} = \frac{1}{N}\sum_{i=1}^{N} Y_i, \quad \bar{T} = \frac{1}{N}\sum_{i=1}^{N} T_i$$

Since (Y_i, T_i) are iid, by the delta method it holds

$$\sqrt{N}\left(\hat{\mu}_{Reg} - \mu\right) \xrightarrow[N\to\infty]{d} N(0, S^2), \qquad S^2 = \frac{\text{Var}(Y) - 2\mu\,\text{Cov}(Y,T) + \mu^2 \text{Var}(T)}{\mathbb{E}[T]^2}$$

and S^2 can be estimated by

$$\hat{S}^2 = \frac{\hat{\sigma}^2_Y - 2\hat{\mu}_{Reg}\hat{\sigma}^2_{YT} + \hat{\mu}^2_{Reg}\hat{\sigma}^2_T}{\bar{T}^2}$$

$$\hat{\sigma}^2_Y = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})^2$$

$$\hat{\sigma}^2_{YT} = \frac{1}{N-1}\sum_{i=1}^{N}(Y_i - \bar{Y})(T_i - \bar{T})$$

Observe that in this approach there is no need to disregard a burn-in lag (in a sense, the lag $\tau^{(1)}$ is the burn-in period) the main difficulty in this approach is to find good renewal times. In the discrete state case $\mathcal{X} = \{x_1, x_2, \dots\}$ one can take simply the visit time to a given state x_i:

$\tau^{(k)} = \inf\{n > \tau^{(k-1)} : X_n = x_i\}$ $\quad \tau^{(0)} = 0$.

In the general state space, one should identify a small set A and modify the chain as described earlier in this chapter.