Irreducibility and Recurrence

**Proposition**

Assume $\pi$ satisfies the positivity condition, then the Gibbs sampler yields a $\pi-$irreducible and recurrent Markov chain.

**Proof.**

**Recurrence.** Will follow from irreducibility and the fact that $\pi$ is invariant. **Irreducibility.** Let $X \subset \mathbb{R}^d$, such that $\pi(X) = 1$. Write $K$ for the kernel and let $A \subset X$ such that $\pi(A) > 0$. Then for any $x \in X$

\[
K(x, A) = \int_A K(x, y) dy \\
= \int_A \pi_{X_1|X_1-1}(y_1 \mid x_2, \ldots, x_d) \times \cdots \times \pi_{X_d|X_{d-1}}(y_d \mid y_1, \ldots, y_{d-1}) dy.
\]
Proof.

Thus if for some \( x \in X \) and \( A \) with \( \pi(A) > 0 \) we have \( K(x, A) = 0 \), we must have that

\[
\pi_{X_1|X^{-1}}(y_1 \mid x_2, \ldots, x_d) \times \cdots \times \pi_{X_d|X^{-d}}(y_d \mid y_1, \ldots, y_{d-1}) = 0,
\]

for \( \pi \)-almost all \( y = (y_1, \ldots, y_d) \in A \).

Therefore we must also have that

\[
\pi(y_1, x_2, \ldots, y_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j|X^{-j}}(y_j \mid y_{1:j-1}, x_{j+1:d})}{\pi_{X_j|X^{-j}}(x_j \mid y_{1:j-1}, x_{j+1:d})} = 0,
\]

for almost all \( y = (y_1, \ldots, y_d) \in A \) and thus \( \pi(A) = 0 \) obtaining a contradiction.
Theorem

Assume the positivity condition is satisfied then we have for any integrable function $\varphi : X \rightarrow \mathbb{R}$:

$$
\lim_{t} \frac{1}{t} \sum_{i=1}^{t} \varphi \left(X^{(i)}\right) = \int_{X} \varphi \left(x\right) \pi \left(x\right) dx
$$

for $\pi$—almost all starting value $X^{(1)}$. 
Example: Bivariate Normal Distribution

- Let \( X := (X_1, X_2) \sim \mathcal{N}(\mu, \Sigma) \) where \( \mu = (\mu_1, \mu_2) \) and

\[
\Sigma = \begin{pmatrix}
\sigma_1^2 & \rho \\
\rho & \sigma_2^2
\end{pmatrix}.
\]

- The Gibbs sampler proceeds as follows in this case:

1. Sample \( X_1^{(t)} \sim \mathcal{N} \left( \mu_1 + \rho / \sigma_2^2 \left( X_2^{(t-1)} - \mu_2 \right), \sigma_1^2 - \rho^2 / \sigma_2^2 \right) \)

2. Sample \( X_2^{(t)} \sim \mathcal{N} \left( \mu_2 + \rho / \sigma_1^2 \left( X_1^{(t)} - \mu_1 \right), \sigma_2^2 - \rho^2 / \sigma_1^2 \right) \).

- By proceeding this way, we generate a Markov chain \( X^{(t)} \) whose successive samples are correlated. If successive values of \( X^{(t)} \) are strongly correlated, then we say that the Markov chain mixes slowly.
Bivariate Normal Distribution

Figure: Case where $\rho = 0.1$, first 100 steps.
Bivariate Normal Distribution

Figure: Case where $\rho = 0.99$, first 100 steps.
**Bivariate Normal Distribution**

Figure: Histogram of the first component of the chain after 1000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Figure: Histogram of the first component of the chain after 10000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Figure: Histogram of the first component of the chain after 100,000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Gibbs sampling requires sampling from $\pi_{X_j|X_{-j}}$.

In many scenarios, we can include a set of auxiliary variables $Z_1, ..., Z_p$ and have an “extended” distribution of joint density $\pi(x_1, ..., x_d, z_1, ..., z_p)$ such that

$$\int \pi(x_1, ..., x_d, z_1, ..., z_p) \, dz_1...dz_d = \pi(x_1, ..., x_d).$$

which is such that its full conditionals are easy to sample.

Mixture models, Capture-recapture models, Tobit models, Probit models etc.
Independent data $y_1, ..., y_n$

$$Y_i|\theta \sim \sum_{k=1}^{K} p_k \mathcal{N}(\mu_k, \sigma_k^2)$$

where $\theta = (p_1, ..., p_K, \mu_1, ..., \mu_K, \sigma_1^2, ..., \sigma_K^2)$.
Bayesian Model

- Likelihood function

\[
p(y_1, \ldots, y_n | \theta) = \prod_{i=1}^{n} p(y_i | \theta) = \prod_{i=1}^{n} \left( \sum_{k=1}^{K} \frac{p_k}{\sqrt{2\pi\sigma_k^2}} \exp \left(-\frac{(y_i - \mu_k)^2}{2\sigma_k^2} \right) \right)
\]

Let’s fix \( K = 2, \sigma_k^2 = 1 \) and \( p_k = 1/K \) for all \( k \).

- Prior model

\[
p(\theta) = \prod_{k=1}^{K} p(\mu_k)
\]

where

\[
\mu_k \sim \mathcal{N}(\alpha_k, \beta_k).
\]

Let us fix \( \alpha_k = 0, \beta_k = 1 \) for all \( k \).

- Not obvious how to sample \( p(\mu_1 | \mu_2, y_1, \ldots, y_n) \).
Auxiliary Variables for Mixture Models

- Associate to each $Y_i$ an auxiliary variable $Z_i \in \{1, \ldots, K\}$ such that

  $$\mathbb{P}(Z_i = k | \theta) = p_k$$

  and

  $$Y_i | Z_i = k, \theta \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

  so that

  $$p(y_i | \theta) = \sum_{k=1}^{K} \mathbb{P}(Z_i = k) \mathcal{N}(y_i; \mu_k, \sigma_k^2)$$

- The extended posterior is given by

  $$p(\theta, z_1, ..., z_n | y_1, ..., y_n) \propto p(\theta) \prod_{i=1}^{n} \mathbb{P}(z_i | \theta) p(y_i | z_i, \theta).$$

- Gibbs samples alternately

  $$\mathbb{P}(z_{1:n} | y_{1:n}, \mu_{1:K})$$

  $$p(\mu_{1:K} | y_{1:n}, z_{1:n}).$$
We have

\[ P(z_{1:n} \mid y_{1:n}, \theta) = \prod_{i=1}^{n} P(z_i \mid y_i, \theta) \]

where

\[
P(z_i \mid y_i, \theta) = \frac{P(z_i \mid \theta) p(y_i \mid z_i, \theta)}{\sum_{k=1}^{K} P(z_i = k \mid \theta) p(y_i \mid z_i = k, \theta)}
\]

Let \( n_k = \sum_{i=1}^{n} 1_{\{k\}} (z_i) \), \( n_k \bar{y}_k = \sum_{i=1}^{n} y_i 1_{\{k\}} (z_i) \) then

\[
\mu_k \mid z_{1:n}, y_{1:n} \sim \mathcal{N} \left( \frac{n_k \bar{y}_k}{1 + n_k}, \frac{1}{1 + n_k} \right).
\]
Mixtures of Normals

Figure: 200 points sampled from $\frac{1}{2} \mathcal{N}(-2, 1) + \frac{1}{2} \mathcal{N}(2, 1)$. 

Mixtures of Normals

Figure: Histogram of the parameters obtained by 10,000 iterations of Gibbs sampling.
Mixtures of Normals

Figure: Traceplot of the parameters obtained by 10,000 iterations of Gibbs sampling.
Many posterior distributions can be automatically decomposed into conditional distributions by computer programs.

This is the idea behind BUGS (Bayesian inference Using Gibbs Sampling), JAGS (Just another Gibbs Sampler).
Given a target \( \pi(x) = \pi(x_1, x_2, ..., x_d) \), Gibbs sampling works by sampling from \( \pi_{X_j|X_{-j}}(x_j| x_{-j}) \) for \( j = 1, ..., d \).

Sampling exactly from one of these full conditionals might be a hard problem itself.

Even if it is possible, the Gibbs sampler might converge slowly if components are highly correlated.

If the components are not highly correlated then Gibbs sampling performs well, even when \( d \to \infty \), e.g. with an error increasing “only” polynomially with \( d \).

Metropolis–Hastings algorithm (1953, 1970) is a more general algorithm that can bypass these problems.

Additionally Gibbs can be recovered as a special case.
Metropolis–Hastings algorithm

- Target distribution on $\mathbb{X} = \mathbb{R}^d$ of density $\pi(x)$.
- Proposal distribution: for any $x, x' \in \mathbb{X}$, we have $q(x'|x) \geq 0$ and $\int_{\mathbb{X}} q(x'|x) \, dx' = 1$.
- Starting with $X^{(1)}$, for $t = 2, 3, ...$

1. Sample $X^* \sim q(\cdot | X^{(t-1)})$.

2. Compute

$$\alpha(X^* | X^{(t-1)}) = \min \left(1, \frac{\pi(X^*) q(X^{(t-1)} | X^*)}{\pi(X^{(t-1)}) q(X^* | X^{(t-1)})} \right).$$

3. Sample $U \sim \mathcal{U}_{[0,1]}$. If $U \leq \alpha(X^* | X^{(t-1)})$, set $X^{(t)} = X^*$, otherwise set $X^{(t)} = X^{(t-1)}$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

Figure: Metropolis–Hastings on a bivariate Gaussian target.
Metropolis–Hastings algorithm

- Metropolis–Hastings only requires point-wise evaluations of $\pi (x)$ up to a normalizing constant; indeed if $\tilde{\pi} (x) \propto \pi (x)$ then

$$\frac{\pi (x^*) q \left( x^{(t-1)} \mid x^* \right)}{\pi (x^{(t-1)}) q \left( x^* \mid x^{(t-1)} \right)} = \frac{\tilde{\pi} (x^*) q \left( x^{(t-1)} \mid x^* \right)}{\tilde{\pi} (x^{(t-1)}) q \left( x^* \mid x^{(t-1)} \right)}.$$ 

- At each iteration $t$, a candidate is proposed. The probability of a candidate being accepted is given by

$$a \left( x^{(t-1)} \right) = \int_X \alpha \left( x \mid x^{(t-1)} \right) q \left( x \mid x^{(t-1)} \right) dx$$

in which case $X^{(t)} = X$, otherwise $X^{(t)} = X^{(t-1)}$.

- This algorithm clearly defines a Markov chain $\left( X^{(t)} \right)_{t \geq 1}$. 
The kernel of the Metropolis–Hastings algorithm is given by

$$K(y \mid x) \equiv K(x, y) = \alpha(y \mid x)q(y \mid x) + (1 - a(x))\delta_x(y).$$

Proof.

We have

$$K(x, y)$$

$$= \int q(x^* \mid x)\{\alpha(x^* \mid x)\delta_{x^*}(y) + (1 - \alpha(x^* \mid x))\delta_x(y)\}dx^*$$

$$= q(y \mid x)\alpha(y \mid x) + \left\{ \int q(x^* \mid x)(1 - \alpha(x^* \mid x))dx^* \right\} \delta_x(y)$$

$$= q(y \mid x)\alpha(y \mid x) + \left\{ 1 - \int q(x^* \mid x)\alpha(x^* \mid x)dx^* \right\} \delta_x(y)$$

$$= q(y \mid x)\alpha(y \mid x) + \{1 - a(x)\} \delta_x(y).$$
The Metropolis–Hastings kernel $K$ is $\pi$–reversible and thus admit $\pi$ as invariant distribution.

Proof.

For any $x, y \in X$, with $x \neq y$

$$\pi(x)K(x, y) = \pi(x)q(y \mid x)\alpha(y \mid x)$$

$$= \pi(x)q(y \mid x) \left( 1 \land \frac{\pi(y)q(x \mid y)}{\pi(x)q(y \mid x)} \right)$$

$$= \left( \pi(x)q(y \mid x) \land \pi(y)q(x \mid y) \right)$$

$$= \pi(y)q(x \mid y) \left( \frac{\pi(x)q(y \mid x)}{\pi(y)q(x \mid y)} \land 1 \right) = \pi(y)K(y, x).$$

If $x = y$, then obviously $\pi(x)K(x, y) = \pi(y)K(y, x)$. 

\qed
Consider the target distribution

\[ \pi(x) = \left( \mathcal{U}_{[0,1]}(x) + \mathcal{U}_{[2,3]}(x) \right) / 2 \]

and the proposal distribution

\[ q(x^* | x) = \mathcal{U}_{(x-\delta, x+\delta)}(x^*) . \]

The MH chain is reducible if \( \delta \leq 1 \): the chain stays either in \([0, 1]\) or \([2, 3]\).

Note that the MH chain is aperiodic if it always has a non-zero chance of staying where it is.
Some results

**Proposition**

If $q(x^*|x) > 0$ for any $x, x^* \in \text{supp}(\pi)$ then the Metropolis-Hastings chain is *irreducible*, in fact every state can be reached in a single step (strongly irreducible).

Less strict conditions in (Roberts & Rosenthal, 2004).

**Proposition**

*If the MH chain is irreducible then it is also Harris recurrent* (see Tierney, 1994).
Theorem

If the Markov chain generated by the Metropolis–Hastings sampler is $\pi$–irreducible, then we have for any integrable function $\varphi : \mathbb{X} \to \mathbb{R}$:

$$\lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} \varphi \left( X^{(i)} \right) = \int_{\mathbb{X}} \varphi (x) \pi (x) \, dx$$

for every starting value $X^{(1)}$. 
Random Walk Metropolis–Hastings

- In the Metropolis–Hastings, pick \( q(x^* \mid x) = g(x^* - x) \) with \( g \) being a \textit{symmetric} distribution, thus

\[
X^* = X + \varepsilon, \quad \varepsilon \sim g;
\]

e.g. \( g \) is a zero-mean multivariate normal or t-student.

- Acceptance probability becomes

\[
\alpha(x^* \mid x) = \min \left( 1, \frac{\pi(x^*)}{\pi(x)} \right).
\]

- We accept...
  - a move to a more probable state with probability 1;
  - a move to a less probable state with probability

\[
\pi(x^*) / \pi(x) < 1.
\]
Independent proposal: a proposal distribution \( q(x^* \mid x) \) which does not depend on \( x \).

Acceptance probability becomes

\[
\alpha(x^* \mid x) = \min \left( 1, \frac{\pi(x^*) q(x)}{\pi(x) q(x^*)} \right).
\]

For instance, multivariate normal or t-student distribution.

If \( \pi(x)/q(x) < M \) for all \( x \) and some \( M < \infty \), then the chain is uniformly ergodic.

The acceptance probability at stationarity is at least \( 1/M \) (Lemma 7.9 of Robert & Casella).

On the other hand, if such an \( M \) does not exist, the chain is not even geometrically ergodic!
Choosing a good proposal distribution

- **Goal:** design a Markov chain with small correlation \( \rho \left( X^{(t-1)}, X^{(t)} \right) \) between subsequent values (why?).

- Two sources of correlation:
  - between the current state \( X^{(t-1)} \) and proposed value \( X \sim q \left( \cdot \mid X^{(t-1)} \right) \),
  - correlation induced if \( X^{(t)} = X^{(t-1)} \), if proposal is rejected.

- Trade-off: there is a compromise between
  - proposing large moves,
  - obtaining a decent acceptance probability.

- For multivariate distributions: covariance of proposal should reflect the covariance structure of the target.
Choice of proposal

- Target distribution, we want to sample from
  \[\pi(x) = \mathcal{N}(x; \begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.5 \\ 0.5 & 1 \end{pmatrix})\].

- We use a random walk Metropolis—Hastings algorithm with
  \[g(\varepsilon) = \mathcal{N}(\varepsilon; 0, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix})\].

- What is the optimal choice of \(\sigma^2\)?
- We consider three choices: \(\sigma^2 = 0.1^2, 1, 10^2\).
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 0.1^2$, the acceptance rate is $\approx 94\%$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 0.1^2$, the acceptance rate is $\approx 94\%$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 1$, the acceptance rate is $\approx 52\%$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 1$, the acceptance rate is $\approx 52\%$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 10$, the acceptance rate is $\approx 1.5\%$. 
Figure: Metropolis–Hastings on a bivariate Gaussian target. With $\sigma^2 = 10$, the acceptance rate is $\approx 1.5\%$. 
Choice of proposal

- Aim at some intermediate acceptance ratio: 20%? 40%? Some hints come from the literature on “optimal scaling”.
- Literature suggest tuning to get .234...
- Maximize the expected square jumping distance:
  \[ \mathbb{E} [||X_{t+1} - X_t||^2] \]
- In multivariate cases, try to mimic the covariance structure of the target distribution.

Cooking recipe: run the algorithm for \( T \) iterations, check some criterion, tune the proposal distribution accordingly, run the algorithm for \( T \) iterations again…

“Constructing a chain that mixes well is somewhat of an art.”

*All of Statistics*, L. Wasserman.
The adaptive MCMC approach

- One can make the transition kernel $K$ adaptive, i.e. use $K_t$ at iteration $t$ and choose $K_t$ using the past sample $(X_1, \ldots, X_{t-1})$.

- The Markov chain is not homogeneous anymore: the mathematical study of the algorithm is much more complicated.

- Adaptation can be counterproductive in some cases (see Atchadé & Rosenthal, 2005)!

- Adaptive Gibbs samplers also exist.