Advanced Simulation - Lecture 6

George Deligiannidis

February 4th, 2020
Markov chains - discrete space

• Let $X$ be discrete, e.g. $X = \mathbb{Z}$.

• $(X_t)_{t \geq 1}$ is a Markov chain if

$$
P(X_t = x_t | X_1 = x_1, \ldots, X_{t-1} = x_{t-1}) = P(X_t = x_t | X_{t-1} = x_{t-1}).$$

The future is conditionally independent of the past given the present.

• Homogeneous Markov chains:

$$
\forall m \in \mathbb{N}: P(X_t = y | X_{t-1} = x) = P(X_{t+m} = y | X_{t+m-1} = x).
$$

• The Markov transition kernel is a stochastic matrix

$$
K(i, j) = K_{ij} = P(X_t = j | X_{t-1} = i).
$$
Markov chains - discrete space

- Let $\mu_t(x) = \mathbb{P}(X_t = x)$, the chain rule yields

$$
\mathbb{P}(X_1 = x_1, X_2 = x_2, ..., X_t = x_t) = \mu_1(x_1) \prod_{i=2}^{t} K_{x_{i-1}x_i}.
$$

- The $m$-transition matrix $K^m$ as

$$
K^m_{ij} = \mathbb{P}(X_t + m = j \mid X_t = i).
$$

- Chapman-Kolmogorov equation:

$$
K^{m+n}_{ij} = \sum_{k \in \mathcal{X}} K^m_{ik} K^n_{kj}.
$$

- We obtain

$$
\mu_{t+1}(j) = \sum_i \mu_t(i) K_{ij}
$$

i.e. using "linear algebra notation",

$$
\mu_{t+1} = \mu_t K.
$$
Roadmap

- We will see that we can choose the transition matrix $K$ such that if $\mu_0 = \pi$ then $\mu_t = \pi$ for all $t$.
- In practice we will have $\mu_0 \neq \pi$;
- We will see that under certain conditions, not matter what $\mu_0$ is, $\mu_t \to \pi$ in total variation.
- This is enough to guarantee us a law of large numbers and a central limit theorem;
- Making this convergence precise, e.g. in terms of the dimension, is still an active research area.
Irreducibility and aperiodicity

**Definition**

A Markov chain is said to be **irreducible** if all the states communicate with each other, that is

\[
\forall x, y \in \mathbb{X} \quad \min \left\{ t : K_{xy}^t > 0 \right\} < \infty.
\]

A state \( x \) has **period** \( d(x) \) defined as

\[
d(x) = \gcd \{ s \geq 1 : K_{xx}^s > 0 \}.
\]

An irreducible chain is **aperiodic** if all states have period 1.

Example: \( K_\theta = \begin{pmatrix} \theta & 1 - \theta \\ 1 - \theta & \theta \end{pmatrix} \) is irreducible if \( \theta \in [0, 1) \) and aperiodic if \( \theta \in (0, 1) \). If \( \theta = 0 \), the gcd is 2.
Transience and recurrence

Introduce the number of visits to \( x \):

\[
\eta_x := \sum_{k=1}^{\infty} 1\{X_k = x\}.
\]

**Definition**

A state \( x \) is termed **transient** if:

\[
E_x (\eta_x) < \infty,
\]

where \( E_x \) refers to the law of the chain starting from \( x \).

A state is called **recurrent** otherwise and

\[
E_x (\eta_x) = \infty.
\]

**Proposition**

If a finite state chain is irreducible, then either all states are **recurrent** or **transient**. In addition all states have the same period.
Invariant distribution

**Definition**

A distribution $\pi$ is **invariant**, or **stationary**, for a Markov kernel $K$, if

$$\pi K = \pi.$$  

Note: if there exists $t$ such that $X_t \sim \pi$, then

$$X_{t+s} \sim \pi$$

for all $s \in \mathbb{N}$.

**Example:** for any $\theta \in [0, 1]$

$$K_\theta = \begin{pmatrix} \theta & 1 - \theta \\ 1 - \theta & \theta \end{pmatrix}$$

admits the invariant distribution

$$\pi = \begin{pmatrix} \frac{1}{2} & \frac{1}{2} \end{pmatrix}.$$
Detailed balance

Definition
A Markov kernel $K$ satisfies detailed balance for $\pi$ if
\[ \forall x, y \in X : \pi(x)K_{xy} = \pi(y)K_{yx}. \]

Lemma
If $K$ satisfies detailed balance for $\pi$ then $K$ is $\pi$-invariant.

If $K$ satisfies detailed balance for $\pi$ then the Markov chain is reversible, i.e. at stationarity,
\[ \forall x, y \in X : \mathbb{P}(X_t = x, X_{t+1} = y) = \mathbb{P}(X_t = x, X_{t-1} = y). \]
Lack of reversibility

- Let \( P = \begin{pmatrix} \frac{1}{3} & \frac{1}{3} & \frac{1}{3} \\ 1 & 0 & 0 \\ 0 & 1 & 0 \end{pmatrix} \).

- Check \( \pi P = \pi \) for \( \pi = (1/2, 1/3, 1/6) \).

- \( P \) cannot be \( \pi \) reversible as

\[ 1 \rightarrow 3 \rightarrow 2 \rightarrow 1 \]

is a possible sequence whereas

\[ 1 \rightarrow 2 \rightarrow 3 \rightarrow 1 \]

is not (as \( P_{2,3} = 0 \)).

- Detailed balance does not hold as \( \pi_2 P_{23} = 0 \neq \pi_3 P_{32} \).
Remarks

- All finite space Markov chains have at least one stationary distribution but not all stationary distributions are also limiting distributions.

\[
P = \begin{pmatrix}
0.4 & 0.6 & 0 & 0 \\
0.2 & 0.8 & 0 & 0 \\
0 & 0 & 0.4 & 0.6 \\
0 & 0 & 0.2 & 0.8
\end{pmatrix}
\]

Two left eigenvectors of eigenvalue 1:

\[
\pi_1 = \left( \frac{1}{4}, \frac{3}{4}, 0, 0 \right),
\]
\[
\pi_2 = \left( 0, 0, \frac{1}{4}, \frac{3}{4} \right)
\]

depending on the initial state, two different stationary distributions.
Equilibrium

**Proposition**

If a discrete space Markov chain is aperiodic and irreducible and admits an invariant distribution $\pi(\cdot)$, then

$$\forall x \in \mathcal{X} \quad \mathbb{P}_\mu(X_t = x) \xrightarrow{t \to \infty} \pi(x),$$

for any starting distribution $\mu$.

- In the Monte Carlo perspective, we will be primarily interested in convergence of empirical averages, such as

$$\hat{I}_n = \frac{1}{n} \sum_{t=1}^{n} \varphi(X_t) \xrightarrow{a.s. \ \ n \to \infty} I = \sum_{x \in \mathcal{X}} \varphi(x) \pi(x).$$

- Before turning to these “ergodic theorems”, let us consider continuous spaces.
Markov chains - continuous space

- The state space $\mathbb{X}$ is now continuous, e.g. $\mathbb{R}^d$.

- $(X_t)_{t \geq 1}$ is a Markov chain if for any (measurable) set $A$,

$$
P( X_t \in A | X_1 = x_1, X_2 = x_2, \ldots, X_{t-1} = x_{t-1} )
= P( X_t \in A | X_{t-1} = x_{t-1} ).$$

*The future is conditionally independent of the past given the present.*

- We have

$$
P( X_t \in A | X_{t-1} = x ) = \int_A K(x, y) \, dy = K(x, A),$$

that is conditional on $X_{t-1} = x$, $X_t$ is a random variable which admits a probability density function $K(x, \cdot)$.

- $K : \mathbb{X}^2 \to \mathbb{R}$ is the **kernel** of the Markov chain.
Markov chains - continuous space

- Denoting $\mu_1$ the pdf of $X_1$, we obtain directly

$$P(X_1 \in A_1, ..., X_t \in A_t) = \int_{A_1 \times \cdots \times A_t} \mu_1(x_1) \prod_{k=2}^{t} K(x_{k-1}, x_k) \, dx_1 \cdots dx_t.$$  

- Denoting by $\mu_t$ the distribution of $X_t$, Chapman-Kolmogorov equation reads

$$\mu_t(y) = \int_{X} \mu_{t-1}(x) K(x, y) \, dx$$

and similarly for $m > 1$

$$\mu_{t+m}(y) = \int_{X} \mu_t(x) K^m(x, y) \, dx$$

where

$$K^m(x_t, x_{t+m}) = \int_{X^{m-1}} \prod_{k=t+1}^{t+m} K(x_{k-1}, x_k) \, dx_{t+1} \cdots dx_{t+m-1}.$$
Example

- Consider the autoregressive (AR) model

\[ X_t = \rho X_{t-1} + V_t \]

where \( V_t \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \tau^2) \). This defines a Markov chain such that

\[ K(x, y) = \frac{1}{\sqrt{2\pi\tau^2}} \exp\left(-\frac{1}{2\tau^2} (y - \rho x)^2\right). \]

- We also have

\[ X_{t+m} = \rho^m X_t + \sum_{k=1}^{m} \rho^{m-k} V_{t+k} \]

so in the Gaussian case

\[ K^m(x, y) = \frac{1}{\sqrt{2\pi\tau^2_m}} \exp\left(-\frac{1}{2 \frac{\tau^2_m}{\tau^2_m}} (y - \rho^m x)^2\right) \]

with \( \tau^2_m = \tau^2 \sum_{k=1}^{m} (\rho^2)^{m-k} = \tau^2 \frac{1-\rho^{2m}}{1-\rho^2}. \)
Irreducibility and aperiodicity

**Definition**

Given a probability measure $\mu$ over $\mathbb{X}$, a Markov chain is $\mu$-irreducible if

$$\forall x \in \mathbb{X} \ \forall A: \mu(A) > 0 \ \exists t \in \mathbb{N} \ K^t(x, A) > 0.$$ 

A $\mu$-irreducible Markov chain of transition kernel $K$ is periodic if there exists some partition of the state space $\mathbb{X}_1, ..., \mathbb{X}_d$ for $d \geq 2$, such that

$$\forall i, j, t, s: \mathbb{P}(X_{t+s} \in \mathbb{X}_j \mid X_t \in \mathbb{X}_i) = \begin{cases} 1 & j = i + s \mod d \\ 0 & \text{otherwise.} \end{cases}.$$ 

Otherwise the chain is aperiodic.
Recurrence and Harris Recurrence

For any measurable set \( A \) of \( \mathbb{X} \), let

\[
\eta_A = \sum_{k=1}^{\infty} 1_A (X_k),
\]

the number of visits to the set \( A \).

**Definition**

A \( \mu \)-irreducible Markov chain is **recurrent** if for any measurable set \( A \subset \mathbb{X} : \mu(A) > 0 \), then

\[
\forall x \in A \quad \mathbb{E}_x (\eta_A) = \infty.
\]

A \( \mu \)-irreducible Markov chain is **Harris recurrent** if for any measurable set \( A \subset \mathbb{X} : \mu(A) > 0 \), then

\[
\forall x \in \mathbb{X} \quad \mathbb{P}_x (\eta_A = \infty) = 1.
\]

Harris recurrence is stronger than recurrence.
Invariant Distribution and Reversibility

**Definition**

A distribution of density $\pi$ is invariant or *stationary* for a Markov kernel $K$, if

$$\int_X \pi(x) K(x, y) \, dx = \pi(y).$$

A Markov kernel $K$ is $\pi$-reversible if

$$\forall f \quad \iint f(x, y) \pi(x) K(x, y) \, dx \, dy = \iint f(y, x) \pi(x) K(x, y) \, dx \, dy$$

where $f$ is a bounded measurable function.
Detailed balance

In practice it is easier to check the detailed balance condition:

\[ \forall x, y \in \mathbb{X} \quad \pi(x)K(x, y) = \pi(y)K(y, x) \]

Lemma

*If detailed balance holds, then \( \pi \) is invariant for \( K \) and \( K \) is \( \pi \)-reversible.*

Example: the Gaussian AR process is \( \pi \)-reversible, \( \pi \)-invariant for

\[ \pi(x) = \mathcal{N}\left(x; 0, \frac{\tau^2}{1 - \rho^2}\right) \]

when \( |\rho| < 1 \).
Theorem

Suppose the Markov chain \( \{X_i; i \geq 0\} \) is \( \pi \)-irreducible, with invariant distribution \( \pi \), and suppose that \( X_0 = x \). Then for any \( \pi \)-integrable function \( \phi : X \to \mathbb{R} \):

\[
\lim_{t \to \infty} \frac{1}{t} \sum_{i=1}^{t} \phi(X_i) = \int_{X} \phi(w) \pi(w) \, dw
\]

almost surely, for \( \pi \)-almost every \( x \).

If the chain in addition is Harris recurrent then this holds for every starting value \( x \).
Convergence

**Theorem**

Suppose the kernel $K$ is $\pi$-irreducible, $\pi$-invariant, aperiodic. Then, we have

$$\lim_{t \to \infty} \int_X |K^t(x, y) - \pi(y)| \, dy = 0$$

for $\pi$–almost all starting values $x$.

Under some additional conditions, one can prove that there exists a $\rho < 1$ and a function $M : X \to \mathbb{R}^+$ such that for all measurable sets $A$ and all $n$

$$|K^n(x, A) - \pi(A)| \leq M(x) \rho^n.$$

The chain is then said to be **geometrically ergodic**.
Central Limit Theorem

**Theorem**

*Under regularity conditions, for a Harris recurrent, \( \pi \)-invariant Markov chain, we can prove*

\[
\sqrt{t} \left[ \frac{1}{t} \sum_{i=1}^{t} \varphi(X_i) - \int_{\mathbb{X}} \varphi(x) \pi(x) \, dx \right] \xrightarrow{d} \mathcal{N} \left( 0, \sigma^2(\varphi) \right),
\]

*where the asymptotic variance can be written*

\[
\sigma^2(\varphi) = \mathbb{V}_{\pi} \left[ \varphi(X_1) \right] + 2 \sum_{k=2}^{\infty} \text{Cov}_{\pi} \left[ \varphi(X_1), \varphi(X_k) \right].
\]

This formula shows that (positive) correlations increase the asymptotic variance, compared to i.i.d. samples for which the variance would be \( \mathbb{V}_{\pi}(\varphi(X)) \).
Central Limit Theorem

Example: for the AR Gaussian model,
\[ \pi(x) = \mathcal{N}\left(x; 0, \tau^2 / (1 - \rho^2)\right) \] for \(|\rho| < 1\) and

\[ \text{Cov}(X_1, X_k) = \rho^{k-1} \forall [X_1] = \rho^{k-1} \frac{\tau^2}{1 - \rho^2}. \]

Therefore with \(\varphi(x) = x\),

\[ \sigma^2(\varphi) = \frac{\tau^2}{1 - \rho^2} \left(1 + 2 \sum_{k=1}^{\infty} \rho^k\right) = \frac{\tau^2}{1 - \rho^2} \frac{1 + \rho}{1 - \rho} = \frac{\tau^2}{(1 - \rho)^2}, \]

which increases when \(\rho \to 1\).
Markov chain Monte Carlo

- We are interested in sampling from a distribution $\pi$, for instance a posterior distribution in a Bayesian framework.

- Markov chains with $\pi$ as invariant distribution can be constructed to approximate expectations with respect to $\pi$.

- For example, the Gibbs sampler generates a Markov chain targeting $\pi$ defined on $\mathbb{R}^d$ using the full conditionals

  $$\pi(x_i \mid x_1, \ldots, x_{i-1}, x_{i+1}, \ldots, x_d).$$
Gibbs Sampling

• Assume you are interested in sampling from

\[ \pi(x) = \pi(x_1, x_2, ..., x_d), \quad x \in \mathbb{R}^d. \]

• Notation: \( x_{-i} := (x_1, ..., x_{i-1}, x_{i+1}, ..., x_d) \).

**Systematic scan Gibbs sampler.** Let \( (X_1^{(1)}, ..., X_d^{(1)}) \) be the initial state then iterate for \( t = 2, 3, ... \)

1. Sample \( X_1^{(t)} \sim \pi_{X_1|X_{-1}} (\cdot | X_2^{(t-1)}, ..., X_d^{(t-1)}) \).

\[ \vdots \]

j. Sample \( X_j^{(t)} \sim \pi_{X_j|X_{-j}} (\cdot | X_1^{(t)}, ..., X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)}) \).

\[ \vdots \]

d. Sample \( X_d^{(t)} \sim \pi_{X_d|X_{-d}} (\cdot | X_1^{(t)}, ..., X_{d-1}^{(t)}) \).
Gibbs Sampling

A few questions one can ask about this algorithm:

• Is the joint distribution $\pi$ uniquely specified by the conditional distributions $\pi_{X_i|X_{-i}}$?

• Does the Gibbs sampler provide a Markov chain with the correct stationary distribution $\pi$?

• If yes, does the Markov chain converge towards this invariant distribution?

• It will turn out to be the case under some mild conditions.
Hammersley-Clifford Theorem

**Theorem**

Consider a distribution with continuous density $\pi(x_1, x_2, \ldots, x_d)$ such that

$$\text{supp}(\pi) = \text{supp}\left( \bigotimes_{i=1}^d \pi_{X_i} \right).$$

Then for any $(z_1, \ldots, z_d) \in \text{supp}(\pi)$, we have

$$\pi(x_1, x_2, \ldots, x_d) \propto \prod_{j=1}^d \frac{\pi_{X_j|X_{-j}}(x_j|x_{1:j-1}, z_{j+1:d})}{\pi_{X_j|X_{-j}}(z_j|x_{1:j-1}, z_{j+1:d})}.$$

The condition above is known as the **positivity condition**.

Equivalently, if $\pi_{X_i}(x_i) > 0$ for $i = 1, \ldots, d$, then

$$\pi(x_1, \ldots, x_d) > 0.$$
Proof of Hammersley-Clifford Theorem

Proof.

We have

\[ \pi(x_{1:d-1}, x_d) = \pi_{X_d|X_{-d}}(x_d | x_{1:d-1}) \pi(x_{1:d-1}), \]
\[ \pi(x_{1:d-1}, z_d) = \pi_{X_d|X_{-d}}(z_d | x_{1:d-1}) \pi(x_{1:d-1}). \]

Therefore

\[ \pi(x_{1:d}) = \pi(x_{1:d-1}, z_d) \frac{\pi(x_{1:d-1}, x_d)}{\pi(x_{1:d-1}, z_d)} \]
\[ = \pi(x_{1:d-1}, z_d) \frac{\pi(x_{1:d-1}, x_d) / \pi(x_{1:d-1})}{\pi(x_{1:d-1}, z_d) / \pi(x_{1:d-1})} \]
\[ = \pi(x_{1:d-1}, z_d) \frac{\pi_{X_d|X_{1:d-1}}(x_d | x_{1:d-1})}{\pi_{X_d|X_{1:d-1}}(z_d | x_{1:d-1})}. \]
Proof.

Similarly, we have

\[ \pi(x_{1:d-1}, z_d) = \pi(x_{1:d-2}, z_{d-1}, z_d) = \frac{\pi(x_{1:d-2}, z_{d-1}, z_d)}{\pi(x_{1:d-2}, z_{d-1}, z_d)} \]

\[ = \pi(x_{1:d-2}, z_{d-1}, z_d) = \frac{\pi(x_{1:d-1}, z_d)}{\pi(x_{1:d-1}, z_d) / \pi(x_{1:d-2}, z_d)} \]

\[ = \pi(x_{1:d-2}, z_{d-1}, z_d) = \frac{\pi X_{d-1}X_{-(d-1)}(x_{d-1} | x_{1:d-2}, z_d)}{\pi X_{d-1}X_{-(d-1)}(z_{d-1} | x_{1:d-2}, z_d)} \]

hence

\[ \pi(x_{1:d}) = \pi(x_{1:d-2}, z_{d-1}, z_d) = \pi X_{d-1}X_{-(d-1)}(x_{d-1} | x_{1:d-2}, z_d) \]

\[ \times \frac{\pi X_{d}X_{d-1}(x_d | x_{1:d-1})}{\pi X_{d}X_{d-1}(z_d | x_{1:d-1})} \]
Proof.

By \( z \in \text{supp}(\pi) \) we have that \( \pi_{X_i}(z_i) > 0 \) for all \( i \). Also, we are allowed to suppose that \( \pi_{X_i}(x_i) > 0 \) for all \( i \). Thus all the conditional probabilities we introduce are positive since

\[
\pi_{X_j|X^{-j}}(x_j \mid x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_d) \\
= \frac{\pi(x_1, \ldots, x_{j-1}, x_j, z_{j+1}, \ldots, z_d)}{\pi(x_1, \ldots, x_{j-1}, z_j, z_{j+1}, \ldots, z_d)} > 0.
\]

By iterating we have the theorem.
Example: Non-Integrable Target

- Consider the following conditionals on $\mathbb{R}^+$

\[
\begin{align*}
\pi_{X_1|X_2}(x_1|x_2) &= x_2 \exp(-x_2x_1) \\
\pi_{X_2|X_1}(x_2|x_1) &= x_1 \exp(-x_1x_2).
\end{align*}
\]

We might expect that these full conditionals define a joint probability density $\pi(x_1, x_2)$.

- Hammersley-Clifford would give

\[
\begin{align*}
\pi(x_1, x_2, \ldots, x_d) &\propto \frac{\pi_{X_1|X_2}(x_1|z_2) \pi_{X_2|X_1}(x_2|x_1)}{\pi_{X_1|X_2}(z_1|z_2) \pi_{X_2|X_1}(z_2|x_1)} \\
&= \frac{z_2 \exp(-z_2x_1) x_1 \exp(-x_1x_2)}{z_2 \exp(-z_2z_1) x_1 \exp(-x_1z_2)} \propto \exp(-x_1x_2).
\end{align*}
\]

- However $\int \int \exp(-x_1x_2) \, dx_1 \, dx_2 = \infty$ so

\[
\begin{align*}
\pi_{X_1|X_2}(x_1|x_2) &= x_2 \exp(-x_2x_1) \text{ and} \\
\pi_{X_2|X_1}(x_1|x_2) &= x_1 \exp(-x_1x_2)
\end{align*}
\]

are not compatible.
Example: Positivity condition violated

Figure: Gibbs sampling targeting \(\pi(x, y) \propto 1_{[-1,0] \times [-1,0] \cup [0,1] \times [0,1]}(x, y)\).

Positivity condition is sufficient for the Gibbs sampler to be irreducible.
Invariance of the Gibbs sampler

The kernel of the Gibbs sampler (case $d=2$) is

$$K(x^{(t-1)}, x^{(t)}) = \pi_{X_1|X_2}(x_1^{(t)} | x_2^{(t-1)})\pi_{X_2|X_1}(x_2^{(t)} | x_1^{(t)})$$

Case $d > 2$:

$$K(x^{(t-1)}, x^{(t)}) = \prod_{j=1}^{d} \pi_{X_j|X_{-j}}(x_j^{(t)} | x_{1:j-1}^{(t)}, x_{j+1:d}^{(t-1)})$$

**Proposition**

The systematic scan Gibbs sampler kernel admits $\pi$ as invariant distribution.
Invariance of the Gibbs sampler II

Proof for $d = 2$.

Let $x = (x_1, x_2)$ and $y = (y_1, 2)$. Then we have

\[
\int K(x, y)\pi(x)dx = \int \pi(y_2 \mid y_1)\pi(y_1 \mid x_2)\pi(x_1, x_2)dx_1 dx_2 \\
= \pi(y_2 \mid y_1) \int \pi(y_1 \mid x_2)\pi(x_2)dx_2 \\
= \pi(y_2 \mid y_1)\pi(y_1) = \pi(y_1, y_2) = \pi(y).
\]
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, \(^a\)

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 $x \in X$

$$K(x, A) = \int_A K(x, y)dy$$

$$= \int_A \pi_{X_1|X_{-1}}(y_1 \mid x_2, \ldots, x_d) \times \cdots$$

$$\times \pi_{X_d|X_{-d}}(y_d \mid y_1, \ldots, y_{d-1}) dy.$$ 

\(^a\)Meyn and Tweedie, Markov chains and stochastic stability, Prop’n 10.1.1.
Proof.

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

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

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

Therefore, by the Hammersley-Clifford theorem, we must also have that

\[
\pi(y_1, y_2, \ldots, y_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j | X^{-j}}(y_j | y_1:j-1, x_{j+1:d})}{\pi_{X_j | X^{-j}}(x_j | y_1:j-1, x_{j+1:d})} = 0,
\]

for almost all \( y = (y_1, \ldots, y_d) \in A \) and thus \( \pi(A) = 0 \) obtaining a contradiction.
If the positivity condition is satisfied then for any $\pi$-integrable function $\varphi : X \rightarrow \mathbb{R}$:

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

for $\pi$–almost all starting values $X^{(1)}$. 

Theorem
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

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

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

Figure: Case where $\rho = 0.99$, first 100 steps.
Figure: Histogram of the first component of the chain after 1000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Bivariate Normal Distribution

Figure: Histogram of the first component of the chain after 10000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Bivariate Normal Distribution

Figure: Histogram of the first component of the chain after 100000 iterations. Small $\rho$ on the left, large $\rho$ on the right.
Gibbs Sampling and Auxiliary Variables

- 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.
Mixtures of Normals

- Independent data \( y_1, \ldots, y_n \)

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

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

- **Likelihood function**

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

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** \( p(\mu_1 | \mu_2, y_1, ..., y_n) \).
Auxiliary Variables for Mixture Models

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

$$\mathbb{P}(Z_i = k|\theta) = p_k \quad \text{and} \quad Y_i|Z_i = k, \theta \sim \mathcal{N}(\mu_k, \sigma_k^2)$$

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

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

- Gibbs samples alternately

$$\mathbb{P}(z_1:n|y_1:n, \mu_{1:K})$$

$$p(\mu_{1:K}|y_1:n, z_{1:n}).$$
Gibbs Sampling for Mixture Model

• 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.
Gibbs sampling in practice

- 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).
Gibbs Recap

- 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 $t = 2, 3, ...$
  (a) Sample $X^* \sim q(\cdot|X^{(t-1)})$.
  (b) Compute

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

  (c) Sample $U \sim \mathcal{U}[0,1]$. If $U \leq \alpha(X^*|X^{(t-1)})$, set $X^{(t)} = X^*$, otherwise set $X^{(t)} = X^{(t-1)}$. 
Metropolis–Hastings algorithm

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.
Metropolis–Hastings algorithm

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.
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.
Metropolis–Hastings algorithm

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.
Metropolis–Hastings algorithm

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.
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 **average acceptance probability** from the current state is

\[
a\left(x^{(t-1)}\right) := \int_{\mathcal{X}} \alpha\left(x \mid x^{(t-1)}\right) q\left(x \mid x^{(t-1)}\right) dx
\]

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

- This algorithm clearly defines a Markov chain \( (X^{(t)})_{t \geq 1} \).
Lemma

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

\[
\begin{align*}
K(x, y) &= \int q(x^* \mid x) \{ \alpha(x^* \mid x) \delta_{x^*}(y) + (1 - \alpha(x^* \mid x)) \delta_x(y) \} \, dx^* \\
&= q(y \mid x) \alpha(y \mid x) + \left\{ \int q(x^* \mid x) (1 - \alpha(x^* \mid x)) \, dx^* \right\} \delta_x(y) \\
&= q(y \mid x) \alpha(y \mid x) + \left\{ 1 - \int q(x^* \mid x) \alpha(x^* \mid x) \, dx^* \right\} \delta_x(y) \\
&= q(y \mid x) \alpha(y \mid x) + \{ 1 - a(x) \} \delta_x(y). \end{align*}
\]
Reversibility

Proposition

The Metropolis–Hastings kernel $K$ is $\pi$–reversible and thus admit $\pi$ as invariant distribution.

Proof.

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

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

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

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

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

If $x = y$, then obviously $\pi(x)K(x, y) = \pi(y)K(y, x)$. ⊢
Reducibility and periodicity of Metropolis–Hastings

- 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 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 Metropolis–Hastings

- **Independent proposal**: a proposal distribution $q(x^* | x)$ which does not depend on $x$.
  - Acceptance probability becomes
    $$
    \alpha(x^* | x) = \min \left(1, \frac{\pi(x^*)q(x)}{\pi(x)q(x^*)} \right).
    $$
  - 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 $\rho(X^{(t-1)}, X^{(t)})$ between subsequent values (why?).

- **Two sources of correlation:**
  - between the current state $X^{(t-1)}$ and proposed value $X \sim q(\cdot | X^{(t-1)})$,
  - 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: \( \sigma^2 = 0.1^2, 1, 10^2 \).
Metropolis–Hastings algorithm

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 $\approx 52\%$. 
Metropolis–Hastings algorithm

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 $\approx 1.5\%$. 
Metropolis–Hastings algorithm

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:
  \[ \mathbb{E} \left[ \| X_{t+1} - X_t \|^2 \right] \]
- In multivariate cases, try to mimick 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.