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 $A$,

$$
P(X_t \in A | X_1 = x_1, X_2 = x_2, \ldots, X_{t-1} = x_{t-1}) = \int_A K(x, y) \, dy = K(x, A),$$

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

• $K : \mathbb{X}^2 \rightarrow \mathbb{R}$ is the **kernel** of the Markov chain.

\textit{The future is conditionally independent of the past given the present.}
Markov chains - continuous space

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

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

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

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

and similarly for $m > 1$

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

where

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

• Consider the autoregressive (AR) model

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

where \( V_t \) i.i.d. \( \mathcal{N}(0, \tau^2) \). 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

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

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

**Definition**

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

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

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

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

Otherwise the chain is aperiodic.
Recurrence and Harris Recurrence

For any measurable set $A$ of $\mathcal{X}$, 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 $A \subset \mathcal{X}$: $\mu(A) > 0$, 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 $A \subset \mathcal{X}$: $\mu(A) > 0$, then

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

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

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

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

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

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

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

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

when \( |\rho| < 1 \).
Law of Large Numbers

**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 \( \varphi: X \to \mathbb{R} \):

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

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*

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

where the asymptotic variance can be written

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

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

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

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

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

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

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

\[ \vdots \]

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

\[ \vdots \]

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

A few questions one can ask about this algorithm:

• Is the joint distribution $\pi$ uniquely specified by the conditional distributions $\pi_{X_i|X_{-i}}$?
  
  • **A: Not in general!**

• Does the Gibbs sampler provide a Markov chain with the correct stationary distribution $\pi$?
  
  • **A: Not in general!**

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

• It will turn out to be the case under some mild conditions.

---

1J.P. Hobert, C.P. Robert, C. Goutis, Connectedness conditions for the convergence of the Gibbs sampler (1997)
Hammersley-Clifford Theorem

**Theorem**

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

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

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

\[
\pi(x_1, x_2, ..., x_d) \propto \prod_{j=1}^{d} \frac{\pi_{X_j \mid X_{-j}} \left( x_j \mid x_{1:j-1}, z_{j+1:d} \right)}{\pi_{X_j \mid X_{-j}} \left( z_j \mid x_{1:j-1}, z_{j+1:d} \right)}.
\]

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

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

\[
\pi(x_1, ..., x_d) > 0.
\]
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

\[
\pi(x_1:d-1, z_d) = \pi(x_1:d-2, z_{d-1}, z_d) \frac{\pi(x_1:d-1, z_d)}{\pi(x_1:d-2, z_{d-1}, z_d)}
\]

\[
= \pi(x_1:d-2, z_{d-1}, z_d) \frac{\pi(x_1:d-1, z_d)}{\pi(x_1:d-2, z_{d-1}, z_d)} \frac{\pi(x_1:d-2, z_{d-1}, z_d)}{\pi(x_1:d-2, z_{d-1}, z_d)} / \pi(x_1:d-2, z_{d-1}, z_d)
\]

\[
= \pi(x_1:d-2, z_{d-1}, z_d) \frac{\pi(x_1:d-1, z_{d-1}) | x_1:d-2, z_d}{\pi(x_1:d-1, z_{d-1}) | x_1:d-2, z_d}
\]

hence

\[
\pi(x_1:d) = \pi(x_1:d-2, z_{d-1}, z_d) \frac{\pi(x_1:d-1, z_{d-1}) | x_1:d-2, z_d}{\pi(x_1:d-1, z_{d-1}) | x_1:d-2, z_d}
\]

\[
\times \frac{\pi(x_1:d-1, z_{d-1}) | x_1:d-2, z_d}{\pi(x_1:d-1, z_{d-1}) | x_1:d-2, z_d}
\]

\[
\times \frac{\pi(x_1:d-1, z_{d-1}) | x_1:d-2, z_d}{\pi(x_1:d-1, z_{d-1}) | x_1:d-2, z_d}
\]
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

$$
\pi_{X_j|X_{-j}}(x_j | x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_d) = \frac{\pi(x_1, \ldots, x_{j-1}, z_{j+1}, \ldots, z_d)}{\pi(x_1, \ldots, x_{j-1}, z_j, z_{j+1}, \ldots, z_d)} > 0.
$$

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

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

\[
\pi_{X_1|X_2}(x_1|x_2) = x_2 \exp(-x_2 x_1)
\]

\[
\pi_{X_2|X_1}(x_2|x_1) = x_1 \exp(-x_1 x_2).
\]

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

- Hammersley-Clifford would give

\[
\pi(x_1, x_2, ..., x_d) \propto \frac{\pi_{X_1|X_2}(x_1|z_2) \pi_{X_2|X_1}(x_2|x_1)}{\pi_{X_1|X_2}(z_1|z_2) \pi_{X_2|X_1}(z_2|x_1)} = \frac{z_2 \exp(-z_2 x_1) x_1 \exp(-x_1 x_2)}{z_2 \exp(-z_2 z_1) x_1 \exp(-x_1 z_2)} \propto \exp(-x_1 x_2).
\]

- However, $\int \int \exp(-x_1 x_2) \, dx_1 \, dx_2 = \infty$ so

$\pi_{X_1|X_2}(x_1|x_2) = x_2 \exp(-x_2 x_1)$ and

$\pi_{X_2|X_1}(x_1|x_2) = x_1 \exp(-x_1 x_2)$ 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 violated: any density of the form 
\[ f(x) = \alpha 1_{[-1,0]^2} + (1 - \alpha) 1_{[0,1]^2}, \]
has same conditionals.
Invariance of the Gibbs sampler I

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 $d > 2$:

$$K(x^{(t-1)}, x^{(t)}) = \prod_{j=1}^{d} \pi_{X_j|X_{-j}}(x_j^{(t)} | x_1^{(t)}, x_2^{(t-1)}, x_{j+1:d})$$

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 $y = (y_1, y_2)$. Then we have

$$\int K(x, y)\pi(x)dx = \int \pi(y_2 | y_1)\pi(y_1 | x_2)\pi(x_1, x_2)dx_1 dx_2$$

$$= \pi(y_2 | y_1) \int \pi(y_1 | x_2)\pi(x_2)dx_2$$

$$= \pi(y_2 | y_1)\pi(y_1) = \pi(y_1, y_2) = \pi(y).$$
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, \(^a\)

(One step)Irreducibility. Let $\mathcal{X} \subset \mathbb{R}^d$, such that $\pi(\mathcal{X}) = 1$. Write $K$ for the kernel and let $A \subset \mathcal{X}$ such that $\pi(A) > 0$. Then for any $x \in \mathcal{X}$

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

$$= \int_A \pi_{X_1|X_{-1}}(y_1 \mid x_2, \ldots, x_d) \times \cdots \times \pi_{X_d|X_{-d}}(y_d \mid y_1, \ldots, y_{d-1})dy.$$  

\(^a\)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

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

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

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

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

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

Note: Positivity not necessary for irreducibility; e.g. \( f \propto 1_{|x| \leq 1} \).
Theorem

If the positivity condition is satisfied then for any \( \pi \)-integrable function \( \varphi : \mathbb{X} \rightarrow \mathbb{R} \):

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

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

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

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

- 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.
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.
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.
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 $\overline{\pi}(x_1, ..., x_d, z_1, ..., z_p)$ such that

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

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

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

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

- Associate to each $Y_i$ an auxiliary variable $Z_i \in \{1, \ldots, K\}$ such that
  \[ P(Z_i = k|\theta) = p_k \quad \text{and} \quad Y_i|Z_i = k, \theta \sim \mathcal{N}(\mu_k, \sigma_k^2) \]

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

- The extended posterior is given by
  \[ p(\theta, z_1, \ldots, z_n|y_1, \ldots, y_n) \propto p(\theta) \prod_{i=1}^{n} P(z_i|\theta) p(y_i|z_i, \theta) . \]

- Gibbs samples alternately
  \[ P(z_{1:n}|y_{1:n}, \mu_{1:K}) \]
  \[ p(\mu_{1:K}|y_{1:n}, z_{1:n}) . \]
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 \( n_k = \sum_{i=1}^{n} \mathbf{1}_{\{k\}}(z_i) \), \( n_k \bar{y}_k = \sum_{i=1}^{n} y_i \mathbf{1}_{\{k\}}(z_i) \) 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.
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.
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.
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.
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

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

in which case \( X^{(t)} = X \), otherwise \( X^{(t)} = X^{(t-1)} \).
- This algorithm clearly defines a Markov chain \( (X^{(t)})_{t \geq 1} \).
Transition Kernel and Reversibility

**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

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

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

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

\[ = q(y \mid x)\alpha(y \mid x) + \left\{ 1 - a(x) \right\} \delta_x(y). \]
Reversibility

Proposition

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

Proof.

For any $x, y \in \mathcal{X}$, 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
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 $\varphi : \mathbb{X} \to \mathbb{R}$:

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

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

$$X^* = X + \epsilon, \quad \epsilon \sim g;$$

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

$$\pi(x^*) / \pi(x) \leq 1.$$
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 \( \rho \left( X^{(t-1)}, X^{(t)} \right) \) between subsequent values (why?).

- **Two sources of correlation**:
  - between the current state \( X^{(t-1)} \) and proposed value \( X \sim q \left( \cdot | X^{(t-1)} \right) \),
  - 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

\[ g(\epsilon) = \mathcal{N}(\epsilon; 0, \sigma^2 \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}) \].

- 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 $\approx 94\%$. 
Metropolis–Hastings algorithm

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.

⚠️ Extreme care is needed when designing adaptive algorithms: it’s easy to make an algorithm with the wrong invariant distribution.
Sophisticated Proposals

- "Langevin" proposal relies on

\[ X^* = X^{(t-1)} + \frac{\sigma}{2} \nabla \log \pi \left( X^{(t-1)} \right) + \sigma W \]

where \( W \sim \mathcal{N}(0, I_d) \), so the Metropolis-Hastings acceptance ratio is

\[
\frac{\pi(X^*) q(X^{(t-1)} | X^*)}{\pi(X^{(t-1)}) q(X^* | X^{(t-1)})} = \frac{\pi(X^*)}{\pi(X^{(t-1)})} \frac{\mathcal{N}(X^*; X^{(t-1)} + \frac{\sigma}{2} \nabla \log \pi \left( X^{(t-1)} \right); \sigma^2)}{\mathcal{N}(X^{(t-1)}; X^* + \frac{\sigma}{2} \nabla \log \pi \left( X^* \right); \sigma^2)}.
\]

- Possibility to use higher order derivatives:

\[ X^* = X^{(t-1)} + \frac{\sigma}{2} \left[ \nabla^2 \log \pi \left( X^{(t-1)} \right) \right]^{-1} \nabla \log \pi \left( X^{(t-1)} \right) + \sigma W. \]
Sophisticated Proposals

- We can use

\[ q(X^*|X^{(t-1)}) = g(X^*; \varphi(X^{(t-1)})) \]

where \( g \) is a distribution on \( \mathbb{X} \) of parameters \( \varphi(X^{(t-1)}) \) and \( \varphi \) is a deterministic mapping

\[
\frac{\pi(X^*)q(X^{(t-1)}|X^*)}{\pi(X^{(t-1)})q(X^*|X^{(t-1)})} = \frac{\pi(X^*)g(X^{(t-1)}; \varphi(X^*))}{\pi(X^{(t-1)})g(X^*; \varphi(X^{(t-1)}))}.
\]

- For instance, use heuristics borrowed from optimization techniques.
Sophisticated Proposals

The following link shows a comparison of

- adaptive Metropolis-Hastings,
- Gibbs sampling,
- No U-Turn Sampler (e.g. Hamiltonian MCMC)

on a simple linear model.

twiecki.github.io/blog/2014/01/02/visualizing-mcmc/
Sophisticated Proposals

- Assume you want to sample from a target $\pi$ with $\text{supp}(\pi) \subset \mathbb{R}^+$, e.g. the posterior distribution of a variance/scale parameter.

- Any proposed move, e.g. using a normal random walk, to $\mathbb{R}^-$ is a waste of time.

- Given $X^{(t-1)}$, propose $X^* = \exp(\log X^{(t-1)} + \varepsilon)$ with $\varepsilon \sim \mathcal{N}(0, \sigma^2)$. What is the acceptance probability then?

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

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

Why?

$$
\frac{q(y \mid x)}{q(x \mid y)} = \frac{1}{y\sigma\sqrt{2\pi}} \exp \left[ -\frac{(\log y - \log x)^2}{2\sigma^2} \right] = \frac{x}{y}.
$$
Random Proposals

• Assume you want to use $q_{\sigma^2}(X^*|X^{(t-1)}) = \mathcal{N}(X; X^{(t-1)}, \sigma^2)$ but you don’t know how to pick $\sigma^2$. You decide to pick a random $\sigma_{2,*}^2$ from a distribution $f(\sigma^2)$:

$$\sigma_{2,*}^2 \sim f(\sigma_{2,*}^2), \quad X^*|\sigma_{2,*}^2 \sim q_{\sigma_{2,*}^2}(\cdot|X^{(t-1)})$$

so that

$$q(X^*|X^{(t-1)}) = \int q_{\sigma_{2,*}^2}(X^*|X^{(t-1)}) f(\sigma_{2,*}^2) d\sigma_{2,*}^2.$$ 

• Perhaps $q(X^*|X^{(t-1)})$ cannot be evaluated, e.g. the above integral is intractable. Hence the acceptance probability

$$\min\{1, \frac{\pi(X^*) q(X^{(t-1)}|X^*)}{\pi(X^{(t-1)}) q(X^*|X^{(t-1)})}\}$$

cannot be computed.
Random Proposals

• Instead you decide to accept your proposal with probability

\[
\alpha_t = \min \left\{ 1, \frac{\pi(X^*) q_{\sigma^2,(t-1)} \left( \frac{X^{(t-1)}}{X^*} \right)}{\pi(X^{(t-1)}) q_{\sigma^2,*} \left( \frac{X^*}{X^{(t-1)}} \right)} \right\}
\]

where \( \sigma^2,(t-1) \) corresponds to parameter of the last accepted proposal.

• With probability \( \alpha_t \), set \( \sigma^2,(t) = \sigma^2,* \), \( X^{(t)} = X^* \), otherwise \( \sigma^2,(t) = \sigma^2,(t-1) \), \( X^{(t)} = X^{(t-1)} \).

• **Question**: Is it valid? If so, why?
Random Proposals

- Consider the extended target

\[ \tilde{\pi}(x, \sigma^2) := \pi(x) f(\sigma^2). \]

- Previous algorithm is a Metropolis-Hastings of target \( \tilde{\pi}(x, \sigma^2) \) and proposal

\[ q(y, \tau^2|x, \sigma^2) = f(\tau^2) q_{\tau^2}(y|x) \]

- Indeed, we have

\[
\frac{\tilde{\pi}(y, \tau^2) q(x, \sigma^2|y, \tau^2)}{\tilde{\pi}(x, \sigma^2) q(y, \tau^2|x, \sigma^2)} = \frac{\pi(y) f(\tau^2) f(\sigma^2) q_{\sigma^2}(x|y)}{\pi(x) f(\sigma^2) f(\tau^2) q_{\tau^2}(y|x)} = \frac{\pi(y) q_{\sigma^2}(x|y)}{\pi(x) q_{\tau^2}(y|x)}
\]

- Remark: we just need to be able to sample from \( f(\cdot) \), not to evaluate it.
Using multiple proposals

- Consider a target of density $\pi(x)$ where $x \in X$.
- To sample from $\pi$, you might want to use various proposals for Metropolis-Hastings $q_1(x'|x), q_2(x'|x), ..., q_p(x'|x)$.
- One way to achieve this is to build a proposal

$$q(x'|x) = \sum_{j=1}^{p} \beta_j q_j(x'|x), \quad \beta_j > 0, \quad \sum_{j=1}^{p} \beta_j = 1,$$

and Metropolis-Hastings requires evaluating

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

and thus evaluating $q_j(X^*|X^{(t-1)})$ for $j = 1, ..., p$. 
Motivating Example

• Let

\[ q(x' | x) = \beta_1 \mathcal{N}(x'; x, \Sigma) + (1 - \beta_1) \mathcal{N}(x'; \mu(x), \Sigma) \]

where \( \mu : \mathbb{X} \rightarrow \mathbb{X} \) is a clever but computationally expensive deterministic optimisation algorithm.

• Using \( \beta_1 \approx 1 \) will make most proposed points come from the cheaper proposal distribution \( \mathcal{N}(x'; x, \Sigma) \).

• ... but you won’t save time as \( \mu(X^{(t-1)}) \) needs to be evaluated at every step.
Composing kernels

• How to use different proposals to sample from $\pi$ without evaluating all the densities at each step?

• What about combining different Metropolis-Hastings updates $K_j$ using proposal $q_j$ instead? i.e.

$$K_j (x, x') = \alpha_j (x'|x) q_j (x'|x) + (1 - a_j (x)) \delta_x (x')$$

where

$$\alpha_j (x'|x) = \min \left( 1, \frac{\pi(x') q_j (x|x')}{\pi(x) q_j (x'|x)} \right)$$

$$a_j (x) = \int \alpha_j (x'|x) q_j (x'|x) dx'.$$
Composing kernels

Generally speaking, assume

- $p$ possible updates characterised by kernels $K_j(\cdot, \cdot)$,
- each kernel $K_j$ is $\pi$-invariant.

Two possibilities of combining the $p$ MCMC updates:

- **Cycle**: perform the MCMC updates in a deterministic order.
- **Mixture**: Pick an MCMC update at random.
Cycle of MCMC updates

• Starting with $X^{(1)}$ iterate for $t = 2, 3, ...$

(a) Set $Z^{(t,0)} := X^{(t-1)}$.

(b) For $j = 1, ..., p$, sample $Z^{(t,j)} \sim K_j \left(Z^{(t,j-1)}, . \right)$.

(c) Set $X^{(t)} := Z^{(t,p)}$.

• Full cycle transition kernel is

$$K \left(x^{(t-1)}, x^{(t)} \right) = \int \ldots \int K_1 \left(x^{(t-1)}, z^{(t,1)} \right) K_2 \left(z^{(t,1)}, z^{(t,2)} \right) \ldots K_p \left(z^{(t,p-1)}, x^{(t)} \right) d z^{(t,1)} \ldots d z^{(t,p-1)}.$$ 

• $K$ is $\pi$-invariant.
Mixture of MCMC updates

• Starting with \( X^{(1)} \) iterate for \( t = 2, 3, \ldots \)

  (a) Sample \( J \) from \( \{1, \ldots, p\} \) with \( \mathbb{P}(J = k) = \beta_k \).

  (b) Sample \( X^{(t)} \sim K_J\left( X^{(t-1)}, . \right) \).

• Corresponding transition kernel is

\[
K\left( x^{(t-1)}, x^{(t)} \right) = \sum_{j=1}^{p} \beta_j K_j \left( x^{(t-1)}, x^{(t)} \right).
\]

• \( K \) is \( \pi \)-invariant.

• The algorithm is different from using a mixture proposal

\[
q\left( x' \mid x \right) = \sum_{j=1}^{p} \beta_j q_j \left( x' \mid x \right).
\]
Metropolis-Hastings Design for Multivariate Targets

• If $\dim(X)$ is large, it might be very difficult to design a “good” proposal $q(x'|x)$.

• As in Gibbs sampling, we might want to partition $x$ into $x = (x_1, \ldots, x_d)$ and denote $x_{-j} := x \setminus \{x_j\}$.

• We propose “local” proposals where only $x_j$ is updated

$$
q_j(x'|x) = \underbrace{q_j(x'_j|x)}_{\text{propose new component } j} \underbrace{\delta_{x_{-j}}(x'_{-j})}_{\text{keep other components fixed}}.
$$
Metropolis-Hastings Design for Multivariate Targets

- This yields

\[
\alpha_j(x, x') = \min \left( 1, \frac{\pi(x'_j, x'_j) q_j(x_j | x_{-j}, x'_j) \delta_{x'_j}(x_{-j})}{\pi(x_{-j}, x_j) q_j(x'_j | x_{-j}, x_j) \delta_{x_{-j}}(x'_j)} \right) = 1
\]

\[
= \min \left( 1, \frac{\pi(x_{-j}, x'_j) q_j(x_j | x_{-j}, x'_j)}{\pi(x_{-j}, x_j) q_j(x'_j | x_{-j}, x_j)} \right)
\]

\[
= \min \left( 1, \frac{\pi_{X_j|X_{-j}}(x'_j | x_{-j}) q_j(x_j | x_{-j}, x'_j)}{\pi_{X_j|X_{-j}}(x'_j | x_{-j}) q_j(x'_j | x_{-j}, x_j)} \right).
\]
One-at-a-time MH (cycle/systematic scan)

Starting with $X^{(1)}$ iterate for $t = 2, 3, ...$
For $j = 1,...,d$,

- Sample $X^* \sim q_j(\cdot \mid X_1^{(t)}, ..., X_{j-1}^{(t)}, X_j^{(t-1)}, ..., X_d^{(t-1)})$.

- Compute

\[
\alpha_j = \min \left(1, \frac{\pi_{X_j \mid X_{-j}} \left( X^*_j \mid X_1^{(t)}, ..., X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)} \right)}{\pi_{X_j \mid X_{-j}} \left( X_j^{(t-1)} \mid X_1^{(t)}, ..., X_{j-1}^{(t)}, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)} \right)} \times \frac{q_j \left( X_j^{(t-1)} \mid X_1^{(t)}, ..., X_{j-1}^{(t)}, X_j^*, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)} \right)}{q_j \left( X_j^* \mid X_1^{(t)}, ..., X_{j-1}^{(t)}, X_j^{(t-1)}, X_{j+1}^{(t-1)}, ..., X_d^{(t-1)} \right)} \right). 
\]

- With probability $\alpha_j$, set $X^{(t)} = X^*$, otherwise set $X^{(t)} = X^{(t-1)}$. 

One-at-a-time MH (mixture/random scan)

Starting with $X^{(1)}$ iterate for $t = 2, 3, ...$

- Sample $J$ from $\{1, ..., d\}$ with $\mathbb{P}(J = k) = \beta_k$.
- Sample $X^* \sim q_J\left(\cdot | X_1^{(t)}, ..., X_d^{(t-1)}\right)$.
- Compute

$$
\alpha_J = \min \left(1, \frac{\pi_{X_J|X_{-J}}(X^*_J | X_1^{(t-1)} ... X_{J-1}^{(t-1)}, X_{J+1}^{(t-1)} ...)}{\pi_{X_J|X_{-J}}(X_J^{(t-1)} | X_1^{(t-1)} ... X_{J-1}^{(t-1)}, X_{J+1}^{(t-1)} ...)} \times \frac{q_J(X_J^{(t-1)} | X_1^{(t-1)} ... X_{J-1}^{(t-1)}, X_*^J, X_{J+1}^{(t-1)} ... X_d^{(t-1)})}{q_J(X_J^* | X_1^{(t-1)} ... X_{J-1}^{(t-1)}, X_J^{(t-1)}, X_{J+1}^{(t-1)} ... X_d^{(t-1)})} \right).
$$

- With probability $\alpha_J$ set $X^{(t)} = X^*$, otherwise $X^{(t)} = X^{(t-1)}$. 
Gibbs Sampler as a Metropolis-Hastings algorithm

**Proposition**

The systematic Gibbs sampler is a cycle of one-at-a time MH whereas the random scan Gibbs sampler is a mixture of one-at-a time MH where

\[
q_j \left( x'_j \middle| x \right) = \pi_{X_j \mid X_{-j}} \left( x'_j \middle| x_{-j} \right).
\]

**Proof.**

It follows from

\[
\frac{\pi \left( x_{-j}, x'_j \right) q_j \left( x_j \middle| x_{-j}, x'_j \right)}{\pi \left( x_{-j}, x_j \right) q_j \left( x'_j \middle| x_{-j}, x_j \right)} = \frac{\pi \left( x_{-j} \right) \pi_{X_j \mid X_{-j}} \left( x'_j \middle| x_{-j} \right) \pi_{X_j \mid X_{-j}} \left( x_j \middle| x_{-j} \right)}{\pi \left( x_{-j} \right) \pi_{X_j \mid X_{-j}} \left( x_j \middle| x_{-j} \right) \pi_{X_j \mid X_{-j}} \left( x'_j \middle| x_{-j} \right)} = 1.
\]
This is not a Gibbs sampler

Consider a case where $d = 2$. From $X_1^{(t-1)}, X_2^{(t-1)}$ at time $t-1$:

- Sample $X_1^* \sim \pi(X_1 | X_2^{(t-1)})$, then $X_2^* \sim \pi(X_2 | X_1^*)$. The proposal is then $X^* = (X_1^*, X_2^*)$.

- Compute

$$
\alpha_t = \min \left( 1, \frac{\pi(X_1^*, X_2^*)}{\pi(X_1^{(t-1)}, X_2^{(t-1)})} \frac{q(X^{(t-1)} | X^*)}{q(X^* | X^{(t-1)})} \right)
$$

- Accept $X^*$ or not based on $\alpha_t$, where here

$$
\alpha_t \neq 1
$$

!!