Definition: If a measure π satisfies

$$\pi(x)\, p(x,y) = \pi(y)\, p(y,x), \quad \forall x, y \in \mathcal{J} \qquad (*)$$

then π is called a reversible measure.

Equation $(*)$ is called the detailed balance condition and it is a fundamental building block of the MCMC methods we will study.

Here is why:

Theorem: If a measure π satisfies the detailed balanced condition $(*)$ then it is stationary.

Proof:

$$\sum_{x \in \mathcal{J}} \pi(x) p(x,y) = \sum_{x \in \mathcal{J}} \pi(y)\, p(y,x)$$

(↑ detailed balance)

$$= \pi(y) \underbrace{\sum_{x \in \mathcal{J}} p(y,x)}_{=1 \text{ because } p \text{ is transition matrix}} = \pi(y).$$

∎

Let X_n be a Markov chain which admits a stationary distribution π and suppose $X_0 \sim \pi$.

For fixed n, consider $Y_m = X_{n-m}$. ⇒ For every $n \in \mathbb{N}$, Y_m is a time-homogeneous Markov chain with $Y_0 \sim \pi$. The transition probabilities $q(x,y)$ of Y_m are given by:

$$\begin{aligned} q(x,y) &= \mathbb{P}(Y_1 = y \mid Y_0 = x) \\ &= \mathbb{P}(X_{n-1} = y \mid X_n = x) \\ &= \mathbb{P}(X_n = x \mid X_{n-1} = y)\, \pi(y)/\pi(x) \\ &= \frac{p(x,y)\, \pi(y)}{\pi(x)} \end{aligned}$$

⇒ if the detailed balance condition holds, $q(x,y) = p(x,y)$ for all x and y, and the chain X_n is said to be time-reversible. $\pi(x)\, q(x,y) = \pi(y)\, p(x,y)$!

This is why we call π a reversible measure.

The final and most important theoretical concept we need to define is that of ergodicity.

Definiton: A Markov chain is said to be ergodic if it admits a unique stationary distribution π. In this case, π, the invariant distribution, is said to be the ergodic measure for the chain.

There are a few things that a process needs to do in order to be ergodic. In particular, it has to

(a) eventually explore the entire state space: $\forall x \in \mathcal{S}$, $\exists n$ such that X_n is in some sense close to x.

(b) explore the space in a "homogeneous way": the measure π controls how frequently the process will explore a given region of space. If $A \subset \mathcal{S}$, if $\pi(A)$ is small then X_n will visit A rarely, whereas if $\pi(A)$ is large then X_n will visit A often.
This connects to a common interpretation of ergodicity: "space average equals time average".

(c) The limiting behaviour forgets the initial condition it started from.

For discrete state spaces, there are three conditions which are sufficient in order for X_n to be ergodic. We will state these now. For continuous state space, there are stronger conditions needed. For more details, see

Meyn and Tweedie. Markov chains and stochastic stability. Communication and Control Engineering Series. Springer-Verlag London, London 1993.

The sufficient conditions are:

1) Irreducibility: The chain X_n must be irreducible: any set A can be reached from any other set B with nonzero probability.

2) Positive recurrent: For any set A, the expected number of steps required for the chain to return to A is finite.

3) Aperiodic: For any set A, the number of steps required to return to A must not always be a multiple of some value k.

These three conditions are sufficient because, roughly speaking, positive recurrence ensures existence of an invariant measure, irreducibility ensures this invariant measure is unique, and aperiodicity ensures convergence: for all x

$$\sum_{y \in J} |p^n(x,y) - \pi(y)| \longrightarrow 0.$$

example: The Ornstein-Uhlenbeck process, solution of the SDE

$$dX_t = -\alpha X_t \, dt + \sigma \, dW_t$$

is ergodic with respect to the distribution

$$\pi = \mathcal{N}(0, \sigma^2/2\alpha).$$

(Recall that solutions of Itô SDEs are Markov processes, their numerical approximations Markov chains. We will be able to use this fact in order to estimate integrals for SDEs as well!)

We are now ready to state the central limit theorem for stationary Markov chains.

Recall that before, we used $\{X_n\}_{n\geq 0}$, where X_n are i.i.d. r.v. distributed according to a density π and we used the strong law of large numbers to guarantee the existence of the limit

$$S_n = \frac{1}{n}\sum_{i=1}^{n} X_i \quad \text{as } n \to \infty.$$

Now we consider that $\{X_n\}_{n\geq 0}$ is a Markov chain. This means that the X_i will not necessarily have the same distribution and will certainly be correlated! However, if the Markov chain is ergodic, we can obtain a similar result.

Theorem (Ergodic Theorem): Let X_n be an ergodic Markov chain with unique invariant distribution π. Then, for any integrable function f, the limit

$$\frac{1}{n}\sum_{i=0}^{n-1} f(X_i) \longrightarrow \int f(x)\,\pi(dx) \quad \text{as } n \to \infty$$

is valid, for $\mathbb{P}$-almost surely every x.

Note that:

- if $f = \mathbb{1}_A$ for some set $A \in \mathcal{S}$, then the above limit says exactly that asymptotically, space averages equal time averages.
- We still need to ~~state~~ a central limit theorem to characterise the fluctuations of $\frac{1}{n}\sum f(X_i)$ around $\mathbb{E}(f(X))$, $X \sim \pi$.

Theorem (Central Limit Theorem for Stationary, Reversible Markov Chains). If X_n is an ergodic, reversible Markov chain with invariant distribution π, and supposing that $X_0 \sim \pi$ (so that X_n is stationary), then the central limit theorem applies, i.e.,

$$\sqrt{n}\left(\frac{1}{n}\sum_{i=0}^{n-1}\left(f(X_i) - \mathbb{E}_\pi f(X_1)\right)\right) \xrightarrow{d} \mathcal{N}(0, \sigma^2(f))$$

provided that

$$0 \leq \sigma^2(f) = \mathrm{Var}(f(X_0)) + 2\sum_{i=1}^{\infty}\mathrm{Cov}(f(X_0), f(X_k)) < \infty$$

If it exists, $\sigma^2(f)$ is known as the asymptotic variance.

Most of the results stated were focused on countable or discrete state spaces. However, most of them can be easily extended to $\mathbb{R}^d$-valued Markov chains.

We are now ready to look at Markov-Chain Monte Carlo (MCMC) methods!

Suppose we have a Markov chain X_n which is ergodic with unique stationary distribution π. this implies three things:

1) If $X_n \sim \pi$, then $X_{n+1} \sim \pi$.
2) If the distribution of X_n is $\mathcal{L}(X_n)$, then $\mathcal{L}(X_n) \longrightarrow \pi$ as $n \to \infty$, regardless of the distribution of X_0.
3) The ergodic theorem implies that

$$\frac{1}{n}\sum_{i=0}^{n} f(X_i) \longrightarrow \mathbb{E}_\pi f \quad \text{as } n \to \infty$$

for all integrable functions f.

Suppose we want to compute

$$I = \mathbb{E}_\pi(f(X)) = \int_{\mathbb{R}^d} f(x)\,\pi(x)\,dx.$$

then, if we could somehow construct an ergodic process with unique invariant distribution π, the most natural estimator for I would be to simulate X_i up to some time n, and use the time average

$$\hat{I}_n = \frac{1}{n}\sum f(X_i)$$

as an estimator.

Note that the iid samples we considered in section 2 are a particular case of this.

This idea, of specifically constructing a Markov chain which is ergodic with respect to π is the underpinning of MCMC methods.
These methods were developed almost at the same time as standard MC methods; they both originate from Los Alamos in the 1940s.

From the ergodic theorem, we know that $\hat{I}_n$ is a consistent estimator. However, one fundamental difference is that

$$\mathbb{E}(\hat{I}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(f(X_i)) \neq \frac{1}{n}\sum \mathbb{E}_{X\sim\pi}(f(X)) = I$$

$\Rightarrow$ $\hat{I}_n$ is a biased estimator of I. If we set $X_1 \sim \pi$ then $\hat{I}_n$ is unbiased, but this assumes that we can somehow generate samples of π (in which case we might as well use standard MC methods).

transient phase

When n is small, the distribution of X_n can be very different from π. However, from the third property, we know that if we wait long enough X_n will be distributed according to π.

So this property suggests that we can reduce the bias by introducing a "burn-in phase": we can discard the first n_0 samples, for some $n_0 > 0$ and instead use the estimator

$$\hat{I}_{n_0,n} = \frac{1}{n} \sum_{i=n_0}^{n_0+n} f(X_i).$$

We will see that it is possible to construct a Markov chain which is ergodic with respect to a given distribution π for a wide range of distributions π.
However, the biggest challenge is to decide when to stop (how big n needs to be).

We cannot really use the CLT to construct confidence intervals because $\hat{I}_{n_0,n}$ is based on a single Markov chain $\Rightarrow$ can't really compute variances.

We will see how to deal with this, but first we will focus on how to construct the Markov chains.

5.2. The Metropolis-Hastings Algorithm

The Metropolis-Hastings (MH) algorithm is a generalisation of the rejection sampling method we studied in section 2. It is based in an accept-reject step, which ensures that the resulting Markov chain has the correct stationary distribution

Suppose we are given a target density $\pi(x)$, known up to a normalisation constant and we have an associated conditional density $q(\cdot|x)$, which is easy to sample from — q is known as proposal density. The Metropolis-Hastings algorithm is as follows:

Algorithm: Metropolis-Hastings

Suppose that the chain has state X_n at time n.

1) Generate $Y \sim q(\cdot, X_n)$
2) Set $X_{n+1} = Y$ with probability $\alpha(X_n, Y)$ where
$$\alpha(x,y) = \min\left\{1, \frac{\pi(y)\, q(x|y)}{\pi(x)\, q(y|x)}\right\}$$
3) Otherwise reject the proposal Y and set $X_{n+1} = X_n$.

$\alpha(x,y)$ is known as the Metropolis-Hastings acceptance probability.

Similarly to the rejection sampler, we accept Y with probability p by generating $U \sim U(0,1)$ and accept Y if $U < p$.

Note that: π appears in the numerator and denominator
$\Rightarrow$ Normalisation constants cancel, which is why we only need to know π up to a constant.

We will now see why the MH algorithm works (i.e., that it generates reversible and ergodic Markov chains) and after this we will discuss the importance of the proposal density and how to choose it.

We will prove things in a discrete state space, but the arguments for continuous state spaces are similar.

The chain generated by the MH algorithm is clearly a time-homogeneous Markov process. Let us compute its transition matrix.
If the current state is $X_n = x$, then for $y \neq x$,

$$p(x,y) = q(y|x)\,\alpha(x,y)$$

($X_{n+1} = y$ if MH proposes y and it is accepted).
If $x = y$, either MH proposes x and it is accepted or proposes some other z and it is rejected.
So in this case

$$p(x,x) = q(x|x)\,\alpha(x,x) + \sum_{z \in S} (1 - \alpha(x,z))\, q(z|x)$$

So, in the end, we have

$$\boxed{p(x,y) = q(y|x)\,\alpha(x,y) + \delta_x(y) \sum_{z \in S} (1 - \alpha(x,z))\, q(z|x).}$$

Proposition: The density π is reversible with respect to the above transition density.

Proof: If $x \neq y$, then

$$\begin{aligned}
\pi(x)\,p(x,y) &= \pi(x)\,q(y|x)\,\alpha(x,y) \\
&= \pi(x)\,q(y|x)\,\min\left\{1, \frac{\pi(y)\,q(x|y)}{\pi(x)\,q(y|x)}\right\} \\
&= \min\left(q(y|x)\,\pi(x),\ q(x|y)\,\pi(y)\right) \\
&= \min\left(\frac{q(y|x)\,\pi(x)}{\pi(y)\,q(x|y)},\ 1\right)\pi(y)\,q(x|y)\,\pi(y) \\
&= \alpha(y,x)\,\pi(y)\,q(x|y) = \pi(y)\,p(y,x).
\end{aligned}$$

The analogous condition for $x = y$ holds immediately and the chain is reversible

Since π is reversible, then it is stationary. and therefore it is an invariant distribution for X_n.

Now we just need to establish conditions under which X_n generated by the MH algorithm is ~~ste~~ ergodic.

Theorem: Assume that π is bounded and positive on every compact set. If there exist $\varepsilon, \delta > 0$ such that

$$q(x|y) > \varepsilon \quad \text{if } |x-y| < \delta,$$

then the MH Markov chain is ergodic with respect to π. In particular, for all integrable functions f,

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} f(X_i) = \mathbb{E}_\pi(f(x)), \text{ for a.e. } X_1.$$

and we have convergence of $\mathcal{L}(X_n)$ to π in total variation, i.e.,

$$\lim_{n\to\infty} \sum_{y\in S} \left| \sum_{x\in S} p^n(x,y)\mu(x) - \pi(y) \right| = 0,$$

~~where $X_1 \sim \mu$~~, for every initial distribution $X_1 \sim \mu$ and where $p^n(x,y)$ is the transition matrix for n steps of the Markov chain.

Note that: the above 2 conditions are not very stringent and are very easy to verify in general.

- Having an ergodic chain is sometimes not enough for it to be useful in practice for sampling; we need fast convergence to the stationary distribution.

To quantify convergence to equilibrium, we can use a "qualitative" convergence rate property: uniform ergodicity. This states that

$$\| P^n(x, \cdot) - \pi(\cdot) \|_{TV} \leq M \rho^n, \quad n = 1, 2, 3, \ldots$$

for some $\rho < 1$, $M < \infty$, where $P^n(x, \cdot)$ is the law of X_n at time n, given that $X_1 = x$. Furthermore $\| \cdot \|_{TV}$ denotes total variation distance, i.e.

$$\| \mu - \nu \|_{TV} = \sup_{A \in \mathcal{F}} | \mu(A) - \nu(A) |.$$

Another (weaker) property is geometric ergodicity, which holds if

$$\| P^n(x, \cdot) - \pi(x) \|_{TV} \leq M(x) \rho^n, \quad n = 1, 2, 3$$

for some $\rho < 1$, $M < \infty$, for π-a.e. $x \in \mathcal{F}$. Note that here M depends on x.

Note that: if $\mathcal{F}$ is finite, then all irreducible and aperiodic Markov chains are geometric (and in fact uniformly) ergodic. However, if $\mathcal{F}$ is infinite, this is not the case.

X_n generated by the MH algorithm may or may not be geometrically/uniformly ergodic depending on the proposal density and the tails of π.

To understand the importance of the proposal density and of the tails of π, let us consider an example.

example. Suppose we wish to sample from a bimodal distribution known up to a normalisation constant $\pi(x) = \exp\left(-(x^2-1)^2\right)$.
As proposal density, we will use a uniform distribution $q(\cdot|x) \sim \mathcal{U}(x-r, x+r)$, i.e., $q(y|x) = \frac{1}{2r} \mathbb{1}_{(x-r, x+r)}(y)$.
Noting that $q(y|x) = q(x|y)$, we have that

$$\alpha(x,y) = \min\left(1, \pi(y)/\pi(x)\right).$$

So the MH algorithm reduces to:

Given the state X_n

- Sample $Y \sim \mathcal{U}(X_n - r, X_n + r)$, $u \sim \mathcal{U}(0,1)$
- If $u < \min\left(1, \pi(Y)/\pi(X_n)\right)$ set $X_{n+1} = Y$
- Otherwise $X_{n+1} = X_n$.

⇒ the acceptance rule is straightforward: we accept Y if $\pi(Y) > \pi(X_n)$ and reject it with probability $\pi(Y)/\pi(X_n)$ otherwise.

Show numerical example, MH.m
Run it first with $r = 1$, $X_0 = 1$, $N = 10^4$.
→ we see that generated MC is really close to the target distribution: this is because the proposal can make jumps of size 1 and this makes it easier to jump from $x = -1$ to $x = 1$.
Then repeat with $r = 0.1$, $X_0 = 1$, $N = 10^4$
→ in this case the chain can only make jumps of length 0.1 ⇒ the chain can get stuck in one of the modes and it can be hard to escape.
If we use $r = 0.01$, the chain does not even leave the $x = 1$ mode!

this example shows how important it is to choose the proposal density correctly in order to obtain a Markov chain with the correct behaviour and estimate $\hat{I}_n$ correctly.

Like with rejection sampling, there are many possible choices for the proposal density. We will explore three classes of proposals.

5.2.1. The independence sampler.

Although we defined the MH algorithm using a proposal density which depends on the current state, this is not necessary: we can choose a proposal to be independent on the current state, $q(y|x) = g(y)$. The resulting algorithm becomes

Algorithm: Independence Sampler:

Given the state X_n

1) Sample $Y \sim g(y)$
2) If $u < \min\left(1, \frac{\pi(Y)g(X_n)}{\pi(X_n)g(Y)}\right)$ accept $X_{n+1} = Y$
3) Otherwise $X_{n+1} = X_n$.

Note that: even though the proposal generates independent samples, the Markov chain X_n is not because of the acceptance probability, which depends on the previous state.

We can ask ourselves: why don't we use rejection sampling straight away in this case? The independence sampler is advantageous in cases where we don't know what A (the upper bound needed for RS) is, or when it is too big so that the algorithm performs poorly!

Another advantage is that if $M = \sup_x \frac{\pi(x)}{g(x)} = \infty$,

we cannot use rejection sampling, but the independence sampler will still work in theory. But its performance can be very poor.

On the other hand, if $M = \sup_x \pi(x)/g(x) < \infty$, we have the following result.

Theorem (Mengersen and Tweedie (1996)): The independence sampler produces a uniformly ergodic chain if there exists a constant such that

$$\pi(x) \leq M g(x), \quad x \in \text{supp } \pi. \qquad (\star)$$

In this case,

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

On the other hand, if there exists a set of x with positive measure such that $(\star)$ does not hold, then X_n is not geometrically ergodic.

Let us return to the example of sampling from a Cauchy distribution using a Gaussian proposal. Run IS.m. We see that the chain fails to capture the behaviour of π at the tails! This is because the tails of a Cauchy distribution are fatter than those of a Gaussian.

For example, for $X_n = 10$, the probability of accepting a state is given by

$$\int_{-\infty}^{+\infty} \frac{e^{-x^2/2}}{1+y^2}\frac{(1+x^2)}{e^{-y^2/2}}\,dy \approx 6.12 \times 10^{-20} \quad (x=10).$$

$\Rightarrow$ the sampler is unlikely to accept a proposal far from the origin.

5.2.2. Random Walk Metropolis Hastings

While the independence sampler is a good approach, it does not take advantage of the fact that the proposals need not be independent. A more natural approach is to consider a local exploration of the neighbourhood of the current state of the Markov chain. The idea is to generate Y according to

$$Y = X_n + \xi$$

where ξ is a random perturbation, with some given distribution g which is assumed to be symmetric around 0 ($g(-y) = g(y)\ \forall y$).

Possible choices are/include

1) $g(y)$ uniform ($Y \sim U(X_n - \delta, X_n + \delta)$, as we saw before)
2) $g(y)$ is Gaussian ($Y \sim N(X_n, \delta^2)$).

the parameter δ controls the size of the jump from the current state.

Note that: the proposal density can be expressed as $q(y|X_n) = g(y - X_n)$ and the symmetry assumption implies that $q(y|X_n) = q(X_n|y)$.

The algorithm is as follows:

Algorithm: Random Walk Metropolis Hastings.

Given the state X_n

1) Sample $Y = X_n + \xi$, $\xi \sim g$
2) If $u < \min\left\{1, \frac{\pi(Y)}{\pi(X_n)}\right\}$ accept $X_{n+1} = Y$
3) Otherwise set $X_{n+1} = X_n$.

We can see that "uphill proposals" (those which take the chain closer to a local mode) are always accepted while "downhill proposals" are accepted with probability equal to the relative "heights" of the posterior at the proposed and current values.

Note that: Although the shape of the distribution g has some effect in the chain's performance, it is much more important to calibrate the step size δ (we saw that in the bimodal example!)

→ There is a tradeoff to be made when choosing δ: choosing δ too large might result in many proposals being generated in regions where $\pi(x)$ is small ⇒ they get rejected. On the other hand, for δ too small, proposals will be accepted, but the chain does not explore the state space properly ⇒ it will take much longer for $\mathcal{L}(x_n)$ to converge to equilibrium.

Although a random walk proposal is a natural one, the RWMH algorithm does not generate uniformly ergodic chains! (If $\pi > 0$, the RWMH is never uniformly ergodic). However, we can establish conditions under which the chain is geometrically ergodic. For example, the log-concavity of π in the tails, i.e., if there exists $k > 0$ and x_1 such that

$$\log \pi(x) - \log \pi(y) \geq k|y - x|$$

for $y < x < -x_1$ or $x_1 < x < y$, then the chain is geometrically ergodic. (if π is positive and symmetric).

5.2.3. Langevin proposals (MALA)

Even though the RWMH is a natural choice for proposal densities, it does not explore the local properties of the target densities.
Using Langevin proposals we can take advantage of that → this is what the MALA (Metropolis-adjusted Langevin algorithm) algorithm does.

Since the gradient of the target distribution $\nabla \pi(x_n)$ points towards the local mode of the distribution, it is natural to bias the proposals to prefer this direction, while still allowing some randomness to promote exploration.

This motivates proposals based on the overdamped Langevin SDE:

$$dX_t = \nabla \log \pi(X_t)\, dt + \sqrt{2}\, dW_t.$$

Then we can consider the Markov chain generated by considering an Euler-Maruyama discretisation of this SDE:

$$X_{n+1} = X_n + \nabla \log \pi(X_n)\,\delta + \sqrt{2\delta}\,\xi_n$$

where $\xi_n \sim N(0,1)$, iid.

Note that the magnitude of the random jump is controlled by the time step δ.

In this case, the proposal conditional density is given by

$$q(y|x) \propto \exp\left(-\frac{|y - x - \nabla \log \pi(x)\,\delta|^2}{4\delta}\right).$$

(i.e., proposing $X_{n+1} = Y$ using the EM approximation is equivalent to sampling from the above distribution).

Note that: the proposal density is not symmetric
⇒ we do not obtain the nice cancellations we did with the other cases when computing the acceptance probabilities.

The main advantage of the MALA scheme is that it proposes moves into regions of high target probability ⇒ they are more likely to be accepted.

However, this comes at the cost of computing the gradient of the logarithm of the density π.
In many applications this is known exactly, but if it isn't, we can replace it with numerical approximations.

The qualitative rate of convergence to equilibrium depends strongly on the tails of the distribution π.
If it has light tails (e.g. $\pi(x) \propto \exp(-\gamma|x|^\beta)$ for $\beta > 2$, then the corresponding Markov chain is not geometrically ergodic.

5.3. Performance and tuning of Metropolis Hastings.

In the previous sections we checked which conditions we need to ensure so that the Markov chain X_n converges to stationarity (i.e., to a target distribution π), and under certain additional conditions it is possible to prove that the convergence is exponential. We also saw that even though $\hat{I}_n$ is a biased estimator, we can decrease the bias by discarding a sufficiently long burn-in simulation (initial samples).
However, this does not tell us when to stop the simulations with any confidence. We would ideally like to have a test which tells us when the bias is sufficiently small and we achieved convergence, based on a single run.

This has motivated the development of convergence diagnostics → empirical tests which can give a measure in confidence that the chain has reached stationarity. We will not study these methods in detail, but examples include

- Geweke's statistic
- Gelman & Rubin's method
- Raftery and Lewis and Heidelberg and Welch Diagnostic

We will, instead, be concerned with the fluctuations of $\hat{I}_n$ around its mean. Even if we assume that the chain is (sufficiently close to) stationary, so its bias can be neglected, these will still exist.

Recall that, for a stationary chain, we have

$$n \operatorname{Var}(\hat{I}_n) = \operatorname{Var}_\pi(f) + 2 \sum_{k=1}^{n-1} \operatorname{Cov}(f(X_0), f(X_k))$$

$$= \operatorname{Var}_\pi(f) \left(1 + 2 \sum_{k=1}^{n-1} \rho_k\right)$$

where $\rho_k = \operatorname{Cov}(f(X_0), f(X_k))/\sigma_f^2$ is the autocorrelation of $f(X_n)$. If the autocorrelations are zero, then $\operatorname{Var}(\hat{I}_n) = \sigma_f^2/n$, which corresponds to an iid chain. However, in general, ρ_k is not zero in recrec, but we clearly want them to be as small as possible, so that the chain performs effectively (= does not fluctuate too much).

The autocorrelations depend strongly on the proposal distribution $\Rightarrow$ we need to make a good choice of q! For the particular cases of RWMH and MALA, they depend very strongly on δ.

We saw this in our examples:

*) If δ is too small, the proposals are really close to each other. This means they are very likely to be accepted but the chain is not really exploring the state space (at least not very quickly). Furthermore, we expect that subsequent samples are strongly correlated.

*) If δ is too large, proposals are very likely to be rejected $\Rightarrow$ the chain will spend a lot of time in the same state until a new proposal is accepted.

Clearly, we need to choose δ somewhere in between, and/or make an "optimal" choice of proposal density. Notice that if it was possible to use $q(\cdot|x) = \pi(\cdot)$, this would be ideal since the autocorrelation would be zero, but this is not doable in practice.

So we will adopt a practical criterion which allows to compare proposal distributions in situations where we don't know a lot about the target distribution π.

The first criterion is to use an estimate of the autocorrelation, which can be easily estimated from one realisation of the chain.

Given n samples $X_1, \dots, X_n$, the sample autocovariance function is given by

$$\hat{\gamma}_h = \frac{1}{n} \sum_{i=1}^{n-|h|} (X_{i+|h|} - \bar{X})(X_i - \bar{X}), \quad -n < h < n$$

and the sample autocorrelation function is given by $\hat{\rho}_h = \frac{\hat{\gamma}_h}{\hat{\gamma}_0}$

Note that $\bar{X}$ is the mean. h is called the lag.

Plotting the sample autocorrelation for different lags h provides a very convenient way of "eyeballing" the performance of a chain → most statistical computing libraries come equipped with functions to compute it.

Because of the autocorrelations, the variance of the MCMC estimator $\hat{I}_n$ will always be larger than if the estimator was generated from iid samples of π. This gives rise to a useful criterion: we will find out for what value of N the MC estimator $\hat{I}_N^{MC} = \frac{1}{N}\sum f(Y_i)$, $Y_i \sim \pi$ iid would have the same variance of $\hat{I}_n$, i.e, when do we have

$$\mathrm{Var}(\hat{I}_n) = \frac{1}{n}\mathrm{Var}_\pi(f(X))\left(1 + 2\sum_{k=1}^{n-1}\rho_k\right) = \frac{\mathrm{Var}_\pi(f)}{N} ?$$

from this, we can conclude that n samples of the MCMC estimator correspond to $N = n/(1 + 2\sum_{k=1}^{n-1}\rho_k)$ samples of an iid sampler.

Usually, $1 + 2\sum_{k=1}^{n-1}\rho_k > 1$, so $n > N$ (note that this is not always the case). This motivates the definition of effective sample size.

$$\mathrm{ESS}(\hat{I}_n) = \frac{n}{1 + 2\sum_{k=1}^{n-1}\rho_k}$$

A final useful criterion to look at is the acceptance rate, i.e., the average rate at which states are accepted by the MH chain. This can be computed by measuring the frequency of acceptance (during the algorithm).

Even though we can optimise the independence sampler by maximising its acceptance rate, this will not be true for RWMH or MALA. However, in certain cases it is possible to find an "optimal" acceptance rate.