Monte-Carlo Methods and Particle Filters: Sampling from simple densities and Markov Chains

Peter Jan van Leeuwen [*]

Department of Meteorology, University of Reading, United Kingdom

January 31, 2011

[*] *Corresponding author address:* Peter jan van Leeuwen, Department of Meteorology, University of Reading, United Kingdom
E-mail: p.j.vanleeuwen@reading.ac.uk

Contents

1. Introduction

When the data-assimilation problem is strongly nonlinear and the underlying probability density functions are far from Gaussian, traditional methods like variants of the Kalman filter and variational methods like 3D and 4DVar are unlikely to describe the solution to the problem satisfactorily. It is recalled that the full solution of the data-assimilation problem is the posterior probability density function. The Kalman filter and its variants assume this pdf to be Gaussian, so that only the first two moments, mean and covariance, are sufficient to describe this pdf. Clearly, when the pdf is multi-model or highly skewed, the mean and the covariance are not effective for this description. Variational methods look for the state vector that maximises the pdf, the so-called mode, using iterative linearisations. It is clear that a multimodal pdf will pose problems, since the methods are likely to end up in a local maximum. Furthermore, again, the mode is unlikely to be an effective description of the posterior pdf.

This forces us to look at more general method that do not assume a unimodal and/or Gaussian like posterior pdf. Since the pdf can have any shape, the most efficient representation will be in terms of (independent) samples from that pdf in general. This module deals with efficient ways to generate samples from a pdf.

In the first part of the module we discuss sampling from pdf's that have known shape. Numerical procedures to sample from the uniform density will be treated, followed by several methods to draw from simple 'arbitrary but known' shaped pdf's exploiting drawing from the uniform density. Examples are the Gaussian and the exponential pdf.

The rest of the module deals with sampling from either very complicated pdf's, or pdf's with unknown shape (i.e. defined by a known generation process). We treat the Gibbs Sampler and discuss its strong points and drawbacks. This is followed by the more general Metropolis Hastings Algorithm. We discuss several variants and applications.

Finally, we confront Particle Filters, and discuss its relations with Metropolis Hastings. It is shown that the straightforward implementation leads to a degenerate filter and discuss the 'curse of dimensionality'. Several variants that might have potential for larger-dimensional problems are discussed. Of these special emphasis is given to methods that explore the so-called proposal density, as these methods seem to be most promising for large-dimsnional applications. Finally, approximations to the full particle filter are discussed, and links with e.g. the EnKF are explored.

2. Sampling from simple densities

We will discuss sampling from probability densities that are relatively simple, based on sampling from the uniform density.

a. Sampling from the uniform density

The uniform density is denoted by $U(0,1)$ and is one in the real interval $[0,1]$, and zero elsewhere. Modern computers generate samples from this density by drawing from $U(0,M)$, in which M is very large. For instance, the ISML library uses $M = 2^{31} - 1$, and the NAG library uses $M = 2^{59}$. Samples are generated according to:

$$v_i = av_{i-1} \quad mod \quad M \tag{1}$$

with a again a large number. The NAG library uses $a = 13^{13}$. These samples are then scaled by M to obtain samples from the standard uniform density as $u_i = v_i/M$.

It is clear that the generated sequence is iterative and deterministic, so not fully random. We call them psuedo-random numbers. However, the generated numbers are uniform and independent. Also, a initial value v_0 for the sequence has to be chosen. This value is called the *seed*. By choosing the seed the same we obtain exactly the same sequence of random numbers, which can by useful for testing. Sometimes the sequence should not be the same, and the seed can be based on e.g. the internal clock of the computer.

b. *Discrete sampling*

We now discuss sampling from simple discrete distributions.

1) The Bernoulli distribution.

   The discrete random variable x has a Bernoulli distribution with probability p if $Pr(x = 0) = 1 - p$ and $Pr(x = 1) = p$. To draw from this distribution we divide the unit interval in two pieces of lengths p and $1 - p$. We draw a random number u from the uniform distribution and if $u \leq p$ we set $x = 1$ and otherwise $x = 0$.

2) Generalised Bernoulli.

   Assume that x can take values from $x_1, ..., x_k$ with probabilities $p_1, ..., p_k$, with $\sum_i p_i = 1$. We split the unit interval in intervals I_1 to I_k with lengths p_1 to p_k. We draw u from the uniform density and determine the interval in which u falls, say I_i The value for x is now chosen as x_i.

   Indeed, $Pr(x = x_i) = Pr(u \in I_i) = p_i$.

3) Binomial distribution.

   The binomial distribution bin(n,p) is given by:

$$Pr(x = i) = \binom{n}{i} p^i (1 - p)^{n-i} \tag{2}$$

   Sampling each term separately using the method described above is quite expensive. A more efficient method arises when it is realised that the sum of iid samples from the Bernoulli distribution is a sample from the binomial distribution. Hence draw n samples u_i from $U(0, 1)$ and form samples x_i from the binomial distribution as the number of u_i with $u_i < p$.

c. *Continuous sampling*

Several methods for sampling from continuous densities exist that are all based on sampling first from the uniform density.

1) Probability integral transforms

If x has distribution $F(x)$ then $u = F(x) \approx U[0,1]$. This can be proved as follows. Assume $u = F(x)$. Then:

$$F_u(z) = Pr(u \leq z) = Pr(F(x) \leq z) = Pr(x \leq F^{-1}(z)) = F_x(F^{-1}(z)) = z \tag{3}$$

So, if $u = F(x)$ then $F_u(z) = z$, meaning that u is uniformly distributed. This can be used to draw from any density by drawing from a uniform density and transforming the variable. For example, suppose we want to draw from an exponential density $1/\lambda \ \ \exp(-\lambda x)$. Clearly

$$F(x) = \int_0^x \frac{1}{\lambda} e^{-\lambda z}\, dz = -\left. e^{-\lambda z}\right|_0^x = 1 - e^{-\lambda x} \tag{4}$$

If we now use the above we find $u = F(x) = 1 - exp(-\lambda x)$, so $x = -1/\lambda \log(1-u)$. Since if u is uniform distributed so is $1-u$ we can generate samples from the exponential density by drawing samples u_i from the uniform density and forming samples $x_i = -1/\lambda \ \ \log(u_i)$.

2) Ratio of uniforms methods

This is a very general method for generation of samples with arbitrary density $p(x)$ that maybe known up to a proportionality constant. It works as follows:

Let u_1 and u_2 be two uniformly distributed variables in the region $C = \{(u_1,u_2) : 0 \geq u_1 \geq \sqrt{p^*(u_2/u_1)}\}$, in which $p^*(x)$ is the un-normalised density of x. Then $x = u_2/u_1$ has density $p(x) = p^*(x)/\int p^*(x)\, dx$.

This can be proven by introducing the transformations $z = u_1$ and $x = u_2/u_1$. We know that if $(y_1,y_2) = g(x_1,x_2)$

$$p_y(y_1,y_2) = p_x(x_1,x_2)J = p_x(g^{-1}(y_1,y_2))J \tag{5}$$

where J is given by:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \frac{\partial x_2}{\partial y_1} \\ \frac{\partial x_1}{\partial y_2} & \frac{\partial x_2}{\partial y_2} \end{vmatrix} \tag{6}$$

To calculate J we need the inverse transformations $u_1 = z$ and $u_2 = xz$, to find $J = z$. Because $p(u_1,u_2) = k$ for $(u_1,u_2) \in C$ where $k = 1/area(C)$, then $p(z,x) = kz$ for $0 \geq z \geq \sqrt{p^*(x)}$. To find the marginal density $p(x)$ from the joint density $p(z,x)$ we integrate z out:

$$p(x) = \int_0^{\sqrt{p^*(x)}} kz\, dz = \frac{k}{2}\left. z^2\right|_0^{\sqrt{p^*(x)}} = \frac{k}{2} p^*(x) \tag{7}$$

proving that x has a density proportional to $p^*(x)$.

As an example we discuss the application of this method to generating samples from the Cauchy density given by:

$$p(x) = \frac{1}{\pi}\frac{1}{1+x^2} \tag{8}$$

The recipe above tells us to first draw two uniform random numbers form the interval

$$C = \left\{ (u_1,u_2) : 0 \geq u_1 \geq \sqrt{\frac{1}{1+(u_2/u_1)^2}} \right\} \tag{9}$$

By rearanging and taking $u_1 \geq 0$ we find the interval:

$$C = \left\{ (u_1, u_2) : 0 \geq 1 \geq \sqrt{\frac{1}{u_1^2 + u_2^2}} \right\} \tag{10}$$

which is the same interval as

$$C = \left\{ (u_1, u_2) : 0 \geq 1 \geq \frac{1}{u_1^2 + u_2^2} \right\} \tag{11}$$

The procedure is now to draw two uniform random numbers u_1 and u_2 and accept them where squared sum is smaller than 1, as the interval above points out. If accepted, a sample from the Cauchy density is obtained as $x = u_2/u_1$.

3) Methods based on mixtures

Suppose one needs samples from a bivariate density $p(x_1, x_2)$ that is hard to draw from. If however $p(x_1|x_2)$ and $p(x_2)$ are easy to draw from one explores:

$$p(x_1, x_2) = p(x_1|x_2)p(x_2) \tag{12}$$

So, to generate a sample from $p(x_1, x_2)$ we first sample x_2 from $p(x_2)$, followed by sampling from $p(x_1|x_2)$ given the sampled value x_2.

This can also be used if the problem is in drawing from the marginal $p(x_1)$. In that case one can explore

$$p(x_1) = \int p(x_1, x_2)\, dx_2 = \int p(x_1|x_2)p(x_2)\, dx_2 \tag{13}$$

An example is generating samples from the Student's t density:

$$p(x) = \frac{\Gamma((n+1)/2)}{\Gamma(n/2)\Gamma(1/2)} \frac{1}{\sqrt{n}} \left[1 + \frac{1}{n} x^2 \right]^{-(n+1)/2} \tag{14}$$

One first samples x_2 from χ_n^2, followed by sampling x_1 from $N(0, n/x_2)$.

Another use of mixtures is replacing the difficult generation from $p(x)$ by sampling from another density $p_1(x)$ from which it is easy to sample with high probability, and a density from which it is difficult to sample with low probability.

Assume that $p(x) - a_1 p_1(x) > 0$, so that

$$\int (p(x) - a_1 p_1(x))\, dx = 1 - a_1 \tag{15}$$

We can now define a density $g_1(x) = (p(x) - a_1 p_1(x))/(1 - a_1)$, and the original density can be written as:

$$p(x) = a_1 p_1(x) + (1 - a_1) g_1(x) \tag{16}$$

Since $p_1(x)$ is easy to sample from, one wants to choose a_1 as large as possible, with the conditions $0 < a_1 \leq 1$ and $p(x) - a_1 p_1(x) > 0$, which leads to a_1 being the minimum of $p(x)/p_1(x)$. What is left is the problem of sampling from $g_1(x)$. Interestingly, we can use the same idea on $g_1(x)$ as on $p(x)$, so we can choose a density $p_2(x)$ that is easy to draw from and specify a constant b_2 such that $g_1(x) - b_2 p_2(x) > 0$, etc. leading to an original density given by:

$$p(x) = a_1 p_1(x) + a_2 p_2(x) + (1 - a_1 - a_2) g_2(x) \tag{17}$$

where $a_2 = (1 - a_1)b_2$. Obviously, we can repeat this process as many times as we like, making the coefficient in front of the density $g_i(x)$, from which it is hard to draw, as small as we like.

We illustrate the resulting sampling procedure when the method is applied twice, see equation (17). A two-step sampling procedure arises in which we first draw from the Generalised Bernoulli distribution with probabilities $[a_1, a_2, 1 - a_1 - a_2]$, by drawing one sample u from the uniform density $U(0, 1)$ and determine in which interval u lies (see section Discrete Sampling). That interval then determines from which of the three densities $p_1(x), p_2(x)$, or $g_2(x)$ we sample next. When $a_1 + a_2$ is close to 1, the change of having to draw from $g_2(x)$ is very small.

d. Sampling vectors and matrices

One often has to sample a vector or a matrix from a multivariate density. The trick is to first sample independent univariate quantities on which specific transformations are applied to introduce the correlation structures between the elements of the vector or the matrix. We simply use that if $(y_1, ..., y_d) = g(x_1, ..., x_d)$ and g is differentiable one to one, the densities of vectors y and x are related by:

$$p_y(y_1, ..., y_d) = p_x(x_1, ..., x_d)J = p_x(g^{-1}(y_1, ..., y_d))J \tag{18}$$

in which the Jacobian J is given by:

$$J = \begin{vmatrix} \frac{\partial x_1}{\partial y_1} & \cdots & \frac{\partial x_d}{\partial y_1} \\ & \vdots & \\ \frac{\partial x_1}{\partial y_d} & \cdots & \frac{\partial x_d}{\partial y_d} \end{vmatrix} \tag{19}$$

As a simple illustration we use the multi-variate Gaussian distribution, given by

$$p(y; \mu_y, \Sigma_y) = \frac{1}{(2\pi)^{d/2}|\Sigma_y|^{1/2}} \exp\left[-\frac{1}{2}(y - \mu_y)^T \Sigma_y^{-1} (y - \mu_y)\right] \tag{20}$$

with mean μ_y an covariance Σ_y. Samples from this density can be obtained by first drawing d independent samples x_i from the univariate Gaussian $N(0, 1)$, and form the vector x with entries these x_i. This vector is distributed as a multi-variate Gaussian with mean $\mu_x = 0$ and covariance $\Sigma_x = I_d$. A linear transformation is applied to x to form y, so:

$$y = Ax + b \quad \text{so} \quad x = A^{-1}(y - b) = g^{-1}(y) \tag{21}$$

leading to

$$p_y(y_1, ..., y_d) = p_x(g^{-1}(y_1, ..., y_d))J = \frac{|A^{-1}|}{(2\pi)^{d/2}} \exp\left[-\frac{1}{2}(y - b)^T (A^{-1})^T A^{-1} (y - b)\right] \tag{22}$$

Equating this to the desired density for y given above results in $b = \mu_y$, and $A = \Sigma_y^{1/2}$. So, after a Cholesky decomposition of Σ samples from $p_y(y)$ can be generated by generating d independent Gaussian samples from $N(0,1)$ and transforming them according to $y = \Sigma_y^{1/2}x + \mu_y$.

It will not come as a surprise that one can generate samples from $N(\mu_y, \Sigma_y)$ using samples from $N(\mu_x, \Sigma_x)$ when $\mu_x \neq 0$ and $\Sigma_x \neq I_d$ using the transformation method explained above. In fact, as long as the inverse of A exists the dimensions of x and y do not have to be the same.

e. Resampling methods

We now discuss methods in which samples are first drawn from a proposal density and then either accepted as samples form the target with a certain probability, so-called Rejection Sampling, or weighted and resampled from the weighted samples, so-called Importance Resampling. The latter is at the heart of Particle Filters, to be discussed later.

1) Rejection Sampling

This method is also known as Acceptance-Rejection sampling. Suppose one wants to sample from a density $p(x)$ that is difficult to draw from, but which can be evaluated relatively easy. The method also works when the density can only be evaluated up to a normalisation constant, so when $p(x) = \tilde{p}(x)/A$ with A unknown. To sample from this density we take the followng steps:

1) Generate a *proposal density* $q(x)$ such that $kq(x) \geq p(x)$ for all x. Not surprisingly, $kq(x)$ is also called the envelope density. The factor k plays an important role, as we will see later on.

2) Generate a start value x_0 from $q(x)$.

3) Generate a u_0 from $U(0, kq(x_0))$.

4) If $u_0 > p(x_0)$ reject x_0, otherwise accept it. It is easy to show that this does lead to draws from $p(x)$.

The original samples from $q(x)$ are accepted with a probability $p(x)/kq(x)$. So the acceptance rate is given by:

$$p(accept) = \int \frac{p(x)}{kq(x)} q(x)\, dx = \frac{1}{k} \int p(x)\, dx = \frac{1}{k} \tag{23}$$

So the smaller k, the higher the acceptance rate. In the limit of accepting all draws from $q(x)$ we find $k = 1$, which means that $q(x) = p(x)$, and we draw $p(x)$ directly. Obviously, this is not what we want, but it does show that high acceptance rates can be expected when the proposal density $q(x)$ is close to the target density $p(x)$.

Several methods exist to make this method more efficient. One of them is so-called Adaptive Rejection Sampling, in which the proposal density $q(x)$ is adapted every time a sample is rejected. The method works when $p(x)$ is concave, i.e. when $\log p(x)$ has derivatives that are non increasing functions of x.

In that case the envelope density of $\log p(x)$ is found using tangent lines at certain points, and the log of the envelope density is a set of linear lines. This means that the envelope density itself is a set of piecewise exponential functions:

$$k_i q(x) = k_i \lambda_i e^{[-\lambda(x - x_{i-1})]} \qquad x_{i-1} < x \leq x_i \tag{24}$$

The method uses as follows. If a sample is rejected it is used to draw a new tangent line and $q(x)$ is refined using that new line. This will result in a $q(x)$ that is closer and closer to $p(x)$ and the number of rejected sampled will decrease rapidly.

Unfortunately, rejection sampling is not very efficient in higher dimensions. The reason is that the acceptance rate is the ratio of the volume of $p(x)$ and $kq(x)$ which tends to be small in high dimensions. To illustrate this consider the following example. Suppose $p(x) = N(0, \sigma_p^2 I_D)$ and $q(x) = N(0, \sigma_q^2 I_D)$, in which I_D is the identity matrix in D dimensions. Since $kq(x) \geq p(x)$ we must have $\sigma_q > \sigma_p$. Furthermore, the minimal value for k can be obtained by ensuring $kq(0) = p(0)$. We know:

$$p(0) = \frac{1}{(2\pi)^{D/2} |\sigma_p^2 I_D|} \tag{25}$$

and

$$kq(0) = \frac{k}{(2\pi)^{D/2} |\sigma_q^2 I_D|} \tag{26}$$

Equating these two gives

$$k = \frac{|\sigma_q^2 I_D|}{|\sigma_p^2 I_D|} = \left(\frac{\sigma_q}{\sigma_p}\right)^D \tag{27}$$

Hence the acceptance rate is $p(accept) = 1/k = (\sigma_p/\sigma_q)^D$ which will be very small when D is large. For instance, when $D = 1000$ and $\sigma_q = 1.01\sigma_p$ we find $p(accept) \approx 1/20,000$.

2) Importance Resampling

Importance resampling has its roots in Importance Sampling. Importance sampling is not a sampling scheme, but a way to evaluate expectations. Assume $p(x)$ is difficult to sample from, but relatively easy to evaluate. Now suppose we want to evaluate the expectation of a function $f(x)$ under $p(x)$. By definition

$$E[f(x)] = \int f(x)p(x)\, dx \approx \sum_i p(x_i) f(x_i) \tag{28}$$

in which one could use a uniform grid to choose the x_i. As is well known, the number of evaluations needed grows exponentially with the dimension of the problem, so that is not very efficient.

Another way would be to draw samples x_i from $p(x)$ and approximate

$$E[f(x)] = \int f(x)p(x)\, dx \approx \sum_i f(x_i) \tag{29}$$

but by assumption, drawing from $p(x)$ is not easy. A way forward is to draw from a proposal density $q(x)$:

$$E[f(x)] = \int f(x)p(x)\, dx = \int f(x)\frac{p(x)}{q(x)}q(x)\, dx \approx \sum_i f(x_i)\frac{p(x_i)}{q(x_i)} = \sum_i w_i f(x_i) \tag{30}$$

in which $\sum_i w_i = 1$. Note that we did draw from the proposal density $q(x)$ to approximate the integral. The condition on the sum of the weights ensures that we can also use this method when normalisation constants of $p(x)$ and $q(x)$ are unknown or difficult to evaluate. The weights w_i should not vary too much: if some weights are very low the value of $f(x_i)$ of those weights contribute little to the expectation, and these samples are a waist of time and energy. For weights that are all of the same order of magnitude we find again $q(x) \approx p(x)$, with equality when all weights are $1/N$ in which N is the number of samples.

As mentioned above, Importance sampling is the basis for Particle filtering, and we will discuss these issues further in that chapter.

This method becomes a sampling method when samples are drawn from the distribution set by the weights, leading to Importance Resampling. We will discuss this further when discussing the Particle Filters.

3. Markov Chains

a. Introduction

A Markov chain is a stochastic process where given the present, past and future states are independent. This can be stated mathematically as:

$$Pr(X^n = x^n | X^{n-1} = x^{n-1}, ..., X^0 = x^0) = Pr(X^n = x^n | X^{n-1} = x^{n-1}) \tag{31}$$

If the above probabilities do not depend on n the Markov chain is said to be *homogeneous*. In that case one can define the *transition probability*, or *transition function*, or *kernel function* as $p(x^n|x^{n-1})$ for each n.

An example of a Markov chain is the random walk process for which the transition equation reads:

$$x^n = x^{n-1} + w^n \tag{32}$$

in which the w^i are independent random variables with probability density $p_w(w^i)$. In the special case that w^i are discrete with $p_w(1) = p$, $p_w(-1) = q$, and $p_w(0) = 1 - p - q$ the transition density is given by

$$p(x^n|x^{n-1}) = \begin{cases} p & \text{if } x^n = x^{n-1} + 1 \\ q & \text{if } x^n = x^{n-1} - 1 \\ 1 - p - q & \text{if } x^n = x^{n-1} \\ 0 & \text{if otherwise} \end{cases} \tag{33}$$

One can define an *absorbing state* as a state x for which $p(x^n|x^{n-1}) = 1$.

For the multi-variate case we can define a transition matrix as:

$$P = \begin{pmatrix} p(x_1|x_1) & ... & p(x_r|x_1) \\ ... & ... & ... \\ p(x_1|x_r) & ... & p(x_r|x_r) \end{pmatrix} \tag{34}$$

Because $\sum_x p(x|y) = 1$ each row has to sum to 1. Such a matrix is called a stochastic matrix and has the property that at least one eigenvalue is equal to one, and the product of two stochastic matrices is also a stochastic matrix. This latter feature allows us to determine the transition matrix of moving from time n to time $n+2$ by forming:

$$\begin{aligned} P(x^{n+2}|x^n) &= Pr(X^{n+2} = x^{n+2}|X^n = x^n) \\ &= \sum_{x^{n+1}} Pr(X^{n+2} = x^{n+2}, X^{n+1} = x^{n+1}|X^n = x^n) \\ &= \sum_{x^{n+1}} Pr(X^{n+2} = x^{n+2}|X^{n+1} = x^{n+1}) Pr(X^{n+1} = x^{n+1}|X^n = x^n) \\ &= \sum_{x^{n+1}} P(x^{n+2}|x^{n+1}) P(x^{n+1}|x^n) \end{aligned} \tag{35}$$

This process can be repeated to obtain the transition matrix over m steps as:

$$P(x^{n+m}|x^n) = \sum_{x^{n+m-1}} \dots \sum_{x^{n+1}} P(x^{n+m}|x^{n+m-1}) \dots P(x^{n+1}|x^n) \tag{36}$$

and even more general:

$$P(x^{n+m_1+m_2}|x^n) = \sum_{x^{n+m_1}} P(x^{n+m_2+m_1}|x^{n+m_1}) P(x^{n+m_1}|x^n) \tag{37}$$

This equation is called the *Chapman-Kolmogorov* equation, of which the continuous form is perhaps more familiar:

$$P(x^{n+m_1+m_2}|x^n) = \int P(x^{n+m_2+m_1}|x^{n+m_1}) P(x^{n+m_1}|x^n)\, dx^{n+m_1} \tag{38}$$

Finally, if the initial state is known only approximately, so described by a marginal distribution $p^0(x^0)$, the marginal distribution at time n can be obtained from:

$$p^n(x^n) = \sum_{x^0} P(x^n|x^0) p^0(x^0) \tag{39}$$

b. Decomposition of the state space

The initial condition of a chain and the transition probabilities determine what part of state space will be visited in the future. A chain is said to be *recurrent* if starting at y it will return to y with probability one. The *return time* T_y is a random quantity whose mean μ_y can be evaluated. If μ_y is finite the chain is *positive recurrent*, otherwise it is *null recurrent*.

A chain is *transient* if it has a positive probability of not returning to y in a finite number of steps. It is possible to decompose the state space in recurrent an transient parts, which has important implications for the behaviour of the dynamical system at hand. .

c. Stationary distributions

Starting from an initial distribution $p^0(x^0)$ the transition distribution will keep on transforming this distribution. Some chains have a stationary distribution, and these chains have special interest for us as we will see in later chapters. The *stationary distribution* of a chain satisfies:

$$\pi(x) = \sum_y P(x|y)\pi(y) \tag{40}$$

so the marginal does not change when the transition distribution works on it. Note that in matrix notation we would write $\pi = P^T\pi$ (and not $\pi = P\pi$). Other names are the *invariant distribution*, *equilibrium distribution*, or *limiting distribution* when this condition is fulfilled by repeatedly performing P.

A quick note on the notation. $P(x|y)$ denotes the transition distribution of the Markov chain. As soon as that is defined, it is also the probability of event $X = x$ given that before $X = y$, so a conditional probability, which explains the notation. The main difference is that the x and y here are events for the same random variable X, so if $x \neq y$ the conditional distribution is zero unless x and y are events at different times. So the transition distribution is a special conditional distribution in which the variables refer to different times.

The following example shows how to obtain this distribution in simple cases. Imagine a 2-dimensional system that can take on values 0 and 1, with transition matrix:

$$P = \begin{pmatrix} 1-p & p \\ q & 1-q \end{pmatrix} \tag{41}$$

in which the rows sum to one by construction. The stationary distribution can be found from $\pi = P^T\pi$, so by writing:

$$\begin{aligned} \pi(0) &= (1-p)\pi(0) + q\pi(1) \\ \pi(1) &= p\pi(0) + (1-q)\pi(1) \end{aligned} \tag{42}$$

which can be solved as $p\pi(0) = q\pi(1)$. Together with the condition $\pi(0) + \pi(1) = 1$ we find $\pi(0) = q/(p+q)$ and $\pi(1) = p/(p+q)$, which is the invariant distribution for this chain.

A stationary distribution doesn't have to be a limiting distribution. One exception is a periodic chain. A chain is *periodic* with period d if all states are periodic with period d, and *aperiodic* if not so. It can be shown that if the chain is aperiodic it has at least one limiting distribution.

A sufficient but not necessary condition for ensuring that the required distribution $\pi(x)$ is invariant under the Markov Chain is that the chain satisfies *detailed balance*. A Markov chain that respects detailed balance is *reversible*, and fulfils:

$$\pi(x)P(y|x) = \pi(y)P(x|y) \tag{43}$$

It states that the rate at which the state moves from x to y when in equilibrium, $\pi(x)P(y|x)$, is the same as the rate at which it moves from y to x, given by $\pi(y)P(x|y)$. This condition is called *detailed balance*, balance because it equates rates of moves through states, and detailed because it does this for every pair of states. (Again note that the transition distribution is a conditional distribution for two (sets of) variables at different times.)

It is easy to show that detailed balance implies that the target distribution is indeed invariant:

$$\sum_y \pi(y)P(x|y) = \sum_y \pi(x)P(y|x) = \pi(x)\sum_y P(y|x) = \pi(x) \tag{44}$$

The reverse is not true. For instance the uniform distribution over the space $0, 1, 2$ is invariant with respect to the homogeneous Markov chain with transition probabilities $P(1|0) = P(2|1) = P(0|2) = 1$ and all others zero, but detailed balance does not hold.

We have just found ways that allow us to generate a Markov chain to draw samples from the invariant distribution. A remaining question is how we get to that distribution given some prior distribution $p(x)$. For that we need the chain to be ergodic, which insures that whatever distribution we start from, we will always end up with the stationary distribution. The *ergodic theorem* states that when a chain is ergodic the mean of a sequence of functions of a state from a chain is equal to the expectation of that function of the state under the stationary distribution:

$$\bar{f}_n = \frac{1}{n} \sum_1^n f(x^n) \rightarrow E_\pi[f(x)] \quad \text{with} \quad \text{probability} \;\; 1 \tag{45}$$

Note that this result holds while the subsequent function evaluations are dependent. Also note that this is the law of large numbers for Markov chains.

Ergodicity is hard to prove in general. However, it has been shown that under mild conditions on the invariant distribution a homogeneous chain is also an ergodic chain. For instance, a chain is *ergodic* if it is aperiodic and positive recurrent (which means that the mean recurrent time is finite). The mathematical expression is that $\min_{x,y} P(y|x) > 0$, i.e. all states can be reached from all other states, none of the transition probabilities is zero. Note that periodic chains are eliminated this way because $P(x|x) > 0$ too.

A full prove will not be given here, but the idea is as follows. One can write any probability distribution as a linear combination of the invariant distribution and something else. Applying the transition distribution will leave the invariant distribution part invariant by definition, but part of the 'something else' will end up as the invariant distribution (because all states can be reached from all other states), and part of it doesn't. So the 'something else' will shrink over time, and in the limit the invariant distribution is reached.

To conclude, a Markov chain starts sampling from the desired distribution after some time if

1) the transition distribution is such that the target distribution is the stationary distribution of the chain. This can for example be realised by ensuring detailed balance

2 the chain has to be ergodic. This can for example be achieved by choosing a homogeneous Markov Chain, i.e. one which does not explicitly depend on the iteration index, and all transition distributions are non-zero.

The 'some time' is not well defined, and we will discuss this in further chapters.

d. Constructing Markov Chains

We want to be able to construct Markov Chains that converge to the desired invariant distribution as fast as possible. Clearly the chain has to be ergodic for this to be true. However, it is often convenient to construct the probabilities for such a chain from a set of base transition probabilities $B_1, ..., B_m$, each of them leaving the desired distribution invariant, but which do not have to be ergodic individually. An example are B_i's that change only a subset of the variables in the state. The full transition probability can be build from a mixture of these base probabilities as:

$$P(y|x) = \sum_k \alpha_k B_k(y|x) \tag{46}$$

with $\sum_k \alpha_k = 1$. Clearly, when each B_i leaves the desired distribution invariant, so does P. Also, when each B_i satisfies detailed balance, so does P, and when one of the B_i is ergodic, so is P.

Another possibility is to generate the full transition probability by applying the base transition probabilities is sequence, i.e.

$$P(y|x) = \Pi_k B_k(y|x) \tag{47}$$

Clearly, when each B_k has the desired probability as invariant, so has P. In this case, detailed balance of all B_k does not guarantee detailed balance for P. If each B_k has non-zero probability of leaving the state unchanged, than P is ergodic. The Gibbs sampler to be discussed in the next chapter, is an example of this algorithm.

e. Example: the random walk

Consider the random walk model with transition probabilities:

$$p(x^n|x^{n-1}) = \begin{cases} 1/4 & \text{if } x^n = x^{n-1}+1 \\ 1/4 & \text{if } x^n = x^{n-1}-1 \\ 1/2 & \text{if } x^n = x^{n-1} \\ 0 & \text{if otherwise} \end{cases} \tag{48}$$

and starting point 0. At each step, the state remains at its current position with probability 1/2, and moves either left of right with probability 1/4. What can we say about the state after n steps? Since the transition probabilities are symmetric around the initial point $X = 0$ we will have $E[X_n] = 0$. But how far are we likely to be from this initial state? A indication of this is given by calculating the variance $E[X_n^2]$, as follows:

$$\begin{aligned} E[X_n^2] = \sum x^2 p_n(x) &= \sum x^2 \left(\frac{1}{2}p_{n-1}(x) + \frac{1}{4}p_{n-1}(x-1) + \frac{1}{4}p_{n-1}(x+1)\right) \\ &= \frac{1}{2}E[X_{n-1}^2] + \frac{1}{4}E[(X_{n-1}+1)^2] + \frac{1}{4}E[(X_{n-1}-1)^2] \\ &= E[X_{n-1}^2] + \frac{1}{2} \end{aligned} \tag{49}$$

since $E[X_i] = 0$. Because $E[X_0^2] = 0$ we find $E[X_n^2] = n/2$. So, after n steps we are expected to have moved a total distance of about $n/2$, but we are expected to have moved only a distance $\sqrt{n/2}$ from the starting point. This is obviously because many of the steps cancel each other.

This random walk process is not invariant because the distance from the starting point grows with n. However, chains that resample random walks can have invariant distributions when they are biased appropriately, or are confined to a finite state space.

In the following example we restrict the random walk to the interval $[-5, 5]$ by assuming probability 1 to stay at its current position when the move is outside this interval. It is easy to show that this chain has the uniform distribution over this interval as its stationary distribution. How many samples does one need to approximate this uniform distribution? A fist guess would perhaps be 10, as the distance travelled is $n/2 = 5$ in this case. However, because a more appropriate measure might be the distance from the starting point, leading to the square of this number, so 100. Indeed, after about 100 steps the resulting distribution starts looking like the uniform distribution.