5. Metropolis-Hastings sampling

In this chapter we discuss the Metropolis-Hastings algorithm to draw samples from a pdf, which would in the data-asimilation case typically (not not necessarily) be the posterior pdf. After the basic algorithm we discuss several variants, like the Slice sampler, and variants that explore gradient information, like Hybrid and Langevin Monte-Carlo. Finally, we briefly discuss Simulated annealing.

a. Metropolis-Hastings

The metropolis-hastings sampler proposes a new sample (sometimes called a *move*), and then decides to accept it or not given some acceptance criterion. It works as follows:

1) Draw a starting point x^0 from some initial pdf p^0.

2) Move the chain to a new value z drawn from a proposal density $q(z|x^{n-1})$

3) Evaluate the acceptance probability of the move $\alpha(x^{n-1}, z)$ with

$$\alpha(x^{n-1}, z) = min\left\{1, \frac{p(z)}{p(x^{n-1})}\frac{q(x^{n-1}|z)}{q(z|x^{n-1})}\right\} \tag{1}$$

4) Draw a random number u from $U(0, 1)$ and accept the move , i.e. $x^n = z$ when $u < \alpha$, otherwise reject it, so $x^n = x^{n-1}$.

5) Change n to $n + 1$ and return to 2) until convergence.

Several other forms of the acceptance criterion can be envisaged, but the one above by Hastings ensures that the chain is always reversible, or fulfils detailed balance, irrespective of the chosen proposal density q, and it is homogeneous. For $p(x)$ to be the invariant or equilibrium density the chain does also have to be ergodic. This is again satisfied if e.g. all transition densities are nonzero.

The *proposal density* is a very important density in Metropolis-Hastings, but not only there, also particle filters can benefit from a good proposal density. How should be choose it? First one has to realise that that is the density from which new potential samples are drawn, so it should be easy to draw from. Most used are the (multivariate) Gaussian and the Cauchy (or Lorentz) density centered around the current value of the chain, so x^{n-1} in this case. The covariance or width of these distributions determines the size of the step in state space.

The actual transition density $p(z|x^{n-1}) = q(z|x^{n-1})\alpha(x^{n-1}, z)$ if $y \neq x^{n-1}$, and there is a finite probability that the chain remains at x^{n-1}, given by

$$p(x^{n-1}|x^{n-1}) = 1 - \int q(z|x^{n-1})\alpha(x^{n-1}, z)\, dy \tag{2}$$

These two forms can be grouped in the general expression:

$$p(z|x^{n-1}) = q(z|x^{n-1})\alpha(x^{n-1}, z) + \delta(z - x^{n-1})\left[1 - \int q(z|x^{n-1})\alpha(x^{n-1}, z)\, dz\right] \tag{3}$$

so the new state of the system has a mixture distribution.

Obviously, this method works if the acceptance rate is not too low. One way to achieve this is to make only small moves, so that α is close to 1, and hence the probability of acceptance is high. This would correspond to a small covariance or width of the proposal density. Unfortunately, this means that the chain moves very slowly through state space, and convergence will be slow. One somehow has to construct moves that are large enough to probe state space efficiently, while at the same time keep acceptance rates high. There are no general cnstruction rules for the proposal q to do this. A practical solution is to monitor the acceptance rate, and adjust q such that acceptance rates are between $20\% - 50\%$.

The acceptance ratio can be written as:

$$\alpha(x^{n-1}, z) = min\left\{1, \frac{p(z)/q(z|x^{n-1})}{p(x^{n-1})/q(x^{n-1}|z)}\right\} \tag{4}$$

and is based on the ratio betwen target and proposal densities. This does remind us of the resampling schemes discussed earlier. Also note that only ratio's of densities are needed, so normalisation constants are of no importance.

The remarks in the previous chapter about the sampling scheme's, e.g. one single long or multiple shorter chains ar valid here too. Also the convergence criteria cary over directly.

b. Specific cases

In this section we discuss some specific often used Metroplois-Hastings algorithms.

1) If the proposal density q is symmetric, i.e. $q(x|z) = q(z|x)$ the acceptance rate α reduces to $\alpha = min\{1, p(z)/p(x^{n-1}\}$, simplifying the calculation. This can be achived for instance by choosing q a function of $|z - x|$.

2) An often used proposal density is that of the random walk, i.e. $x^n = x^{n-1} + w^n$, with w^n a random variable with distribution independent of the chain. Popular choices for q are Normal and Student's t. Clearly, the width of q determines the average size of the moves, and is typically chosen as a constant in the range $(0.5, 3)$ times the covariance of the chain.

   Efficient chains have been reported that use auto-regressive models like $x^n = a + bx^{n-1} + w^n$, especially when choosing $b = -1$, generating alternating chains. Chains of this sort are called *antithetic* and tend to reduce the variance in the estimates, as discussed earlier (see notes of Sarah Dance).

3) The proposal density can also be chosen as not depending on the previous estimate x^{n-1}. (Note that the actual transition density $p(z|x)$ still does depend on x^{n-1}, so the chain is Markov.) A popular choice is the prior density, i.e. our best guess of the density before the new observations come into play. In this case, the acceptance ratio becomes the ratio of the likelihoods as Bayes Theorem shows. The main advantage is that α becomes very easy to calculate. However, when the prior and the likelihood disagree the prior samples will not be well positioned to describe the posterior. This scheme does allow one to try to make the weights vary as little as possible, which can be achieved in the limit by choosing the proposal as close as possible to the posterior, which rules out the prior as a good choice.

4) Instead of updating the complete state vector at once, the so-called *global Metropoli-Hastings sampler*, one can also update individual components, or groups of components of the state vector, the *local Metropolis-Hastings sampler*.

5) Gibbs sampling can be seen as a special case of Metropolis-Hastings in which the proposal density is the density conditional of a certain component given the current values of the other components. When we update component x_i^{n-1} the other component do not change, so $x_{-i}^{n-1} = z_{-i}$, and we find for the acceptance rate:

$$\begin{aligned}
\alpha(x^{n-1}, z) &= min\left\{1, \frac{p(z)}{p(x^{n-1})}\frac{q(x^{n-1}|z)}{q(z|x^{n-1})}\right\} \\
&= min\left\{1, \frac{p(z)}{p(x^{n-1})}\frac{p(x_i^{n-1}|z_{-i})}{p(z_i|x_{-i}^{n-1})}\right\} \\
&= min\left\{1, \frac{p(z_i|z_{-i})p(z_{-i})}{p(x_i^{n-1}|x_{-i}^{n-1})p(x_{-i}^{n-1})}\frac{p(x_i^{n-1}|z_{-i})}{p(z_i|x_{-i}^{n-1})}\right\} \\
&= min\left\{1, \frac{p(z_i|x_{-i}^{n-1})p(x_{-i}^{n-1})}{p(x_i^{n-1}|x_{-i}^{n-1})p(x_{-i}^{n-1})}\frac{p(x_i^{n-1}|x_{-i}^{n-1})}{p(z_i|x_{-i}^{n-1})}\right\} \\
&= min\{1, 1\} = 1
\end{aligned} \tag{5}$$

in which we used the notation $z_{-i} = (z_i, ..., z_{i-1}, z_{i+1}, ..., z_d)^T$ for a d-dimensional state vector z. This shows that all moves are accepted, and the transition densities are the conditionals, as in the Gibbs sampler.

6) The previous result allows one to generate a new version of the Gibbs sampler in which the conditional are only approximations to the actual conditionals, so the acceptance rate is not equal to one, but close to it.

c. The Slice Sampler

One of the main problems with the Metropolis-Hastings scheme is its sensitivity to the step size. If it is too small the acceptance rate will be large, but the samples will be highly correlated, while if it is too large the samples decorrelate faster, but their acceptance rate is low. The Slice Sampler provides an adaptive step size that is automatically adjusted to the characteristics of the target density. As always with this kind of samplers, we only have to know $p(x)$ up to a normalisation constant. Let's call this unormalised density $\tilde{p}(x)$.

To explain the method, first consider the univariate case. The variable x is augmented with an additional variable u, and samples are drawn from the joint (x, u) space. The goal is to sample uniformly from the area under the density given by:

$$\begin{aligned}
p(x, u) &= 1/X_p \quad \text{if} \quad 0 \leq u \leq \tilde{p}(x) \\
&= 0 \quad \text{otherwise}
\end{aligned} \tag{6}$$

where $X_p = \int \tilde{p}(x)\, dx$. The marginal density over x is given by:

$$\int p(x, u)\, du = \int_0^{\tilde{p}(x)} \frac{1}{X_p}\, du = \frac{\tilde{p}(x)}{X_p} = p(x) \tag{7}$$

so, we can sample from $p(x)$ by sampling from $p(x, u)$ and ignoring the u values. This can be done in the following steps:

1) Start with a value for x (the last accepted value) and evaluate $\tilde{p}(x)$.

2) Sample u uniformly from $U(0, \tilde{p}(x))$.

3) Fix this u and sample x uniformly from the slice through the density defined by $\{x : \tilde{p}(x) > u\}$. Hence the acceptance rate is related to u and the steps can be large, dependent of the shape of $p(x)$.

In practice sampling from a slice through a density can be difficult, so the following steps replaces step 3) in the scheme above:

3') Choose a region with width w around x and test if the end points of the region are inside the slice. If they are we extend the region such that both end points are outside the slice.

4) A candidate x' is chosen uniformly in the region, and if it is in the slice it forms the new candidate in the Metropolis-Hastings algorithm. If not the region is shrunk such that x' is the new end point.

5) Repeat 4) until x' is within the slice.

This scheme can be extended to a multi-variate scheme by repeatedly sampling each variable in turn, in the manner of Gibbs sampling. Hence we need to be able to sample from $p(x_i|x_{-i})$.

d. The Hybrid Monte-Carlo Method

As mentioned before, if new candidates in the Metropolis-Hastings algorithm are chosen as in a random walk, as is usually done, the distance travelled through state space grows only as the square of the number of steps taken. Larger steps do not solve this as they lead to low acceptance rates.

Using ideas from dynamical systems we can make the method more efficient. The following is an introduction to dynamical systems, followed by the combination with Metropolis-Hasting to the hybrid scheme. The main difference with the standard Metropolis-Hastings scheme is that the hybrid method explores gradient information from the pdf, or more specifically $-\log p(x)$. In this sense it resamples variational methods to find the mode of the pdf, like 3DVar or 4DVar. Recall, however, that our goal with these sampling methods is not to find the mode, but to try to represent the full pdf by a set of samples. The random ingredients in the method ensure that the methods do not just walk towards the mode.

1) Dynamical systems

We will exploit the evolution of a system under *Hamiltionian dynamics*, to be defined shortly. Consider the evolution of state variable x under continuous time τ. In classical mechanics Newton's law describes how particles are accelerated by forces, the famous $F = ma$ in which m the mass of the particle and a the acceleration, i.e. the second time derivative of

the coordinates of a particle. This law being a second derivative in time, the evolution of a system of particles is fully determined if the forces are given, together with the position and velocities of the particles. So each particle is fully determined by its position and velocity coordinates, and these are related through $v_i = dx_i/dt$. The space spanned by the position and velocity variables is called *phase space*.

The probability density of the system can be written as:

$$p(x) = \frac{1}{Z_p} e^{-E(x)} \tag{8}$$

in which $E(x)$ is called the *potential energy* of the system when in state x. In a Hamiltinian system forces are conservative, which means that they can be written as the gradient of the potential energy, and Newton's law becomes:

$$\frac{dv_i}{d\tau} = -\frac{\partial E(x)}{\partial x_i} \tag{9}$$

The kinetic energy is defined as $K(v) = 1/2 \sum_i v_i^2$, the sum of the squares of the velocities of the particles. The total energy of the system is the sum of the kinetic and potential energies:

$$H(x, v) = E(x) + K(v) \tag{10}$$

where H is called the *Hamiltonian* of the system. Because the velocities are the time derivatives of the positions and only the kinetic part of the Hamiltinian depends on the velocities, we find the Hamiltonian equations:

$$\begin{aligned} \frac{dx_i}{d\tau} &= \frac{\partial H}{\partial v_i} = v_i \\ \frac{dv_i}{d\tau} &= -\frac{\partial H}{\partial x_i} \end{aligned} \tag{11}$$

During the evolution of the system the Hamiltonian is constant:

$$\begin{aligned} \frac{dH}{d\tau} &= \sum_i \left\{ \frac{\partial H}{\partial x_i}\frac{dx_i}{d\tau} + \frac{\partial H}{\partial v_i}\frac{dv_i}{d\tau} \right\} \\ &= \sum_i \left\{ \frac{\partial H}{\partial x_i}\frac{\partial H}{\partial v_i} - \frac{\partial H}{\partial v_i}\frac{\partial H}{\partial x_i} \right\} = 0 \end{aligned} \tag{12}$$

which is just energy conservation in this case.

Another important feature of Hamiltonian systems is that they preserve volume in phase space, the so-called *Liouville Theorem*. This can be derived by calculating the divergence of the flow field in phase space. This flow field is given by $V = (dx/dt, dv/dt)$ and its divergence is:

$$\begin{aligned} divV &= \sum_i \left\{ \frac{\partial}{\partial x_i}\frac{dx_i}{d\tau} + \frac{\partial}{\partial v_i}\frac{dv_i}{d\tau} \right\} \\ &= \sum_i \left\{ -\frac{\partial}{\partial x_i}\frac{\partial H}{\partial v_i} + \frac{\partial}{\partial v_i}\frac{\partial H}{\partial x_i} \right\} = 0 \end{aligned} \tag{13}$$

Having established these relations we can generate the laws that describe the evolution of the density of the system in a new way. Define the joint density of positions and velocities as:

$$p(x,v) = \frac{1}{Z_H} e^{-H(x,v)} \tag{14}$$

Because both H and the volume in phase space are conserved, the Hamiltonian dynamics will leave $p(x,v)$ invariant. The connection with sampling becomes apparent when we realise that although H is constant, x and v may change, and large changes in x are possible when v is large, avoiding random walk behaviour.

To exploit this we have to integrate the Hamiltonian equations numerically. Obviously, this will lead to numerical errors, that we would like to minimise. Especially, scheme's that preserve Liouville's Theorem are of importance, as will become clear in the next section. One of those is the *leap-frog* scheme:

$$\begin{aligned} v_i(\tau + \epsilon/2) &= v_i(\tau) - \frac{\epsilon}{2}\frac{\partial E}{\partial x_i}(x(\tau)) \\ x_i(\tau + \epsilon) &= x_i(\tau) + \epsilon v_i(\tau + \epsilon/2) \\ v_i(\tau + \epsilon) &= v_i(\tau + \epsilon/2) - \frac{\epsilon}{2}\frac{\partial E}{\partial x_i}(x(\tau + \epsilon)) \end{aligned} \tag{15}$$

Although each finite ϵ will lead to numerical errors, we will show in the next section that the hybrid scheme is able to compensate for them. But even with these numerical errors we see that numerical scheme follows Liouville's theorem because the first step changes all v_i by an amount that only depends on x_i. So along each line x_i is constant all v_i change the same amount, and this first step conserves volume. The same is true for the other two steps in the leap-frog scheme: the updated variable is changed by an amount that only depends on the other variable.

Finally, note that the main difference between the Metropolis-Hastings and the hybrid scheme's is that the latter uses gradient information of the density (or rather the log density) while the former does not.

2) Hybrid Monte-Carlo

The hybrid scheme works as follows:

1) Add a random quantity to the velocity variables v by sampling from $p(v|x)$. Since that density is Gaussian it is relatively easy.

2) Update the position variable (which is the actual variable in the target density) with the leap-frog scheme, choosing a positive or negative value for ϵ each with probability $1/2$. Take L of those leap-frog steps.

3) Accept the new state (x^*, v^*) with probability

$$\alpha = min\left\{1, \exp\left[H(x,v) - H(x^*,v^*)\right]\right\} \tag{16}$$

4) Return to 1).

Note that one leap-frog step would make the scheme close to a random walk again, which is what we wanted to avoid. This is the reason to take L leap frog steps, with L not too small. Also note that the stochastic first step is needed to change the total energy of the system. If H would remain constant the chain would not be ergodic.

If the numerical scheme used for solving the Hamiltonian equations was without error the value of H would not change, and each step will be accepted with probability one. Due to numerical errors, the value of H may sometimes decrease, leading to a bias in the Metropolis-Hastings scheme. To avoid this we need detailed balance. One way to achieve this is to choose the step in the leap-frog scheme random with equal probability for a positive or a negative value for ϵ, as shown below.

Consider a region R that is transformed after L iterations of the leap-frog scheme with random step sign to a region R'. The volume of the region δV remains the same under the leap-frog scheme as proven above. The probability of starting in R and ending up in R' is given by:

$$p(R') = p(R'|R)p(R)\delta V = \frac{1}{Z_H} e^{-H(R)} \delta V \frac{1}{2} min\left\{1, e^{[H(R)-H(R')]}\right\} \tag{17}$$

where the factor $1/2$ comes from the probability of choosing a positive step instead of a negative step. Similarly, we can start form R' and move backwards to R with probability:

$$p(R) = p(R|R')p(R')\delta V = \frac{1}{Z_H} e^{-H(R')} \delta V \frac{1}{2} min\left\{1, e^{[H(R')-H(R)]}\right\} \tag{18}$$

Using

$$a\ min\{1, b\} = min\{a, ab\} \tag{19}$$

we see that these two probabilities are the same, so detailed balance holds.

In some applications a slight modification of the scheme is needed to avoid that the scheme returns to its initial position and ergodicity is lost. This can be avoided by choosing the step size in the leap-frog scheme random too. Finally, one could use a slightly different Hamiltonian in the leap-frog steps as long as the acceptance rate is determined using the correct Hamiltonian. This allows one to simplify the actual number of calculations needed.

3) Generalised Hybrid Monte-Carlo

In the generalised method we replace the Hamiltonian equations by:

$$\begin{aligned} \frac{dx_i}{d\tau} &= Av_i \\ \frac{dv_i}{d\tau} &= -A^T \frac{\partial H}{\partial x_i} \end{aligned} \tag{20}$$

and try to choose matrix A such that the correlation between the subsequent samples is as small as possible. This set of equations can be discretised as:

$$\begin{aligned} x(\tau+\epsilon) &= x(\tau) + \epsilon Av - \frac{\epsilon^2}{2} AA^T \frac{\partial E}{\partial x}(\tau) \\ v(\tau+\epsilon) &= v(\tau) - \frac{\epsilon}{2} A^T \left(\frac{\partial E}{\partial x_i}(\tau) + \frac{\partial E}{\partial x}(\tau+\epsilon) \right) \end{aligned} \tag{21}$$

One example for matrix A that has been shown to work well is the cyclic matrix

$$A = \begin{pmatrix} 1 & e^{-\theta} & e^{-2\theta} & ... \\ e^{-\theta} & 1 & e^{-\theta} & ... \\ e^{-2\theta} & e^{-\theta} & 1 & ... \\ ... & ... & ... & ... \\ e^{-d\theta} & e^{-(d-1)\theta} & e^{-(d-2)\theta} & ... \end{pmatrix} \tag{22}$$

where the constant θ can depend on the dimension d of the system. It can be shown that the subsequent sampler are indeed less dependent on each other, dependent on the value for θ. That has to be found by trial and error. As an example, one might use $\theta = 10/d$. A disadvantage is that the samples are more expensive to generate. However, a useful property of the above matrix A is that matrix vector calculations can be performed efficiently in $N \log_2 N$ operations, leading to sometimes serious efficiency savings. Also this method has not been applied to any serious geophysical problem.

e. The Langevin Monte-Carlo method

If one uses the Hybrid Monte-Carlo algorithm with only one leap-frog step the Langevin Monte-Carlo method results. The behaviour of the system is close to that of the random walk.

In a variant one omits the acceptance step by accepting all moves directly. In this case, there is no need to calculate the values of the new momentum variables v at the end of the leap-frog step since they will immediately be replaced by new values from the conditional density at the start of the next iteration. So, there is no reason to represent them at all. The scheme consists of the following steps:

1) Draw d random values β_i from d Gaussian distributions $N(0, 1)$, where d is the dimension of the system.

2) Calculate a new value for the state vector via

$$x_i^n = x_i^{n-1} - \frac{\epsilon^2}{2} \frac{\partial E(x)}{\partial x_i} + \epsilon \beta_i \tag{23}$$

which follows from contracting the momentum and position update in the leap-frog scheme.

One can show that if ϵ is small the acceptance rate in the Metropolis-Hastings version converges to one, so ignoring the acceptance step is justified.

f. Simulated annealing

As mentioned above, Metropolis-Hastings has difficulty with fast exploration of the state space because large steps tend to be rejected. A way to exploit the state space fast is simulated annealing. The word comes from metallurgy in which slow cooling (annealing) leads to tougher metal than rapid cooling (quenching). It is stressed that the purpose of the method is not to sample the pdf, but to find the global mode, or the global minimum of $-\log p(x)$.

The idea is to introduce an artificial temperature changing slowly from a large value to close to zero into the system. This is done by dividing the energy (or costfunction) $E(x) = -\log p(x)$ by this temperature in the acceptance rate, and choosing the proposal density symmetric, leading to:

$$\begin{aligned} \alpha(x^{n-1}, z) &= min\left\{1, \frac{p(z)}{p(x^{n-1})}\frac{q(x^{n-1}|z)}{q(z|x^{n-1})}\right\} = min\left\{1, \frac{p(z)}{p(x^{n-1})}\right\} \\ &= min\left\{1, \exp\left[\frac{E(x^{n-1}) - E(z)}{T}\right]\right\} \end{aligned} \tag{24}$$

For T large, say 10, the acceptance rate is much larger than when $T = 1$ in the standard Metropolis Hastings algorithm. So the idea is to start sampling with a large T allowing for large steps to explore the pdf rapidly. The new state is accepted directly when it has lower energy, i.e. higher probability, but there is a finite change to accept new states that have higher energy. This allows one to move 'up hill' in energy terms. When T is large, this probability is quite high. Then T is slowly made lower and lower, allowing the states to move towards the most probable part of the pdf, so positioning them close to the global mode.

When T approaches zero the density becomes very peaked around the the mode, making it more and more difficult to leave this area around the global mode, and moving fast towards it. The samples in this last stage could be used as samples of the pdf around the mode, allowing for an estimate of the shape of the pdf there.

Several methods exist to reduce the temperature, such as $T^i = \log(T^{i-1})$ and also $T^i = \alpha T^{i-1}$, with $0 < \alpha < 1$. Some authors use a large number of steps for each value of T to allow good exploration of the pdf before moving to a lower temperature, while others argue that T can be reduced at every step since the good exploration is not necessary.

Again, it is stressed that this is actually not so much a sampling method as well as an optimisation method. It has the advantage over 3D and 4DVar that it allows one to leave a local minimum since there is a finite probability to accept a state with higher energy, or lower probability density.