Mode jumping in MCMC: Adapting proposals to the local environment

Christopher Jennison

and Richard Sharp,

Dept of Mathematical Sciences,

University of Bath, UK

“Allan day”

University of Durham,

March 2007

Plan of talk

1. Basics of MCMC
2. The challenge: difficulties in mixing
3. New algorithm, definition and illustration
4. Illustrative mode-jumping example
5. Relation to algorithm of Tjelmeland & Hegstad
6. Harder examples
7. Distributions with “thin” support
8. Distributions on manifolds
9. Conclusions

1. Markov chain Monte Carlo sampling (MCMC)

Aim: To sample from a complex distribution $\pi(x)$ by running a Markov chain with ergodic distribution π.

Typically, X is high-dimensional and π not particularly tractable.

The minimal requirement is that $\pi(x)$ can be evaluated, up to a multiplicative constant, for any specified x.

Method: Create a Markov chain on the state space Ω with transition matrix P satisfying

$$\pi P = \pi.$$

Let π_n denote the distribution of the state X_n after n transitions from an initial state x_0.

Then, if the Markov chain is irreducible and aperiodic,

$$\pi_n \to \pi \quad \text{as} \quad n \to \infty.$$

Detailed balance

It is convenient to work with Markov chains with the property of ***detailed balance***,

$$\pi(x)\,P(x,y) \;=\; \pi(y)\,P(y,x) \quad \text{for all } x,\ y \text{ in } \Omega.$$

The key property $\pi\,P = \pi$ follows since

$$\int_\Omega \pi(x)\,P(x,y)\,dx \;=\; \int_\Omega \pi(y)\,P(y,x)\,dx \;=\; \pi(y).$$

Chains with this property were constructed by Metropolis et al. (1953). Hastings (1970) enhanced their generality — hence the *Metropolis-Hastings* algorithm.

The Gibbs sampler of Geman & Geman (1984) is a special case of the Metropolis-Hastings algorithm.

A very neat implementation of the Gibbs sampler lies at the heart of the BUGS software (Spiegelhalter, Thomas, Best & Gilks, 1996).

The Metropolis-Hastings algorithm

From state $X_n = x$, generate a proposal y from the kernel $q(x, y)$ and calculate the “acceptance probability”

$$\alpha(x, y) \;=\; \min\{1, \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\}.$$

With probability $\alpha(x, y)$, accept and move to $X_{n+1} = y$, with probability $1 - \alpha(x, y)$, reject and stay at $X_{n+1} = x$.

Detailed balance: We need to show, for all $x \neq y$,

$$\pi(x)\, q(x, y)\, \alpha(x, y) \;=\; \pi(y)\, q(y, x)\, \alpha(y, x).$$

The LHS is

$$\pi(x)\, q(x, y)\; \min\{1, \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\}.$$

It is straightforward to check this is equal to

$$\pi(y)\, q(y, x)\; \min\{1, \frac{\pi(x)\, q(x, y)}{\pi(y)\, q(y, x)}\},$$

which equals the RHS.

A variety of “move types”

We may wish to use several “types” of move, indexed by a parameter $\phi \in \Phi$, with transition matrix P_ϕ for move type ϕ.

As long as each P_ϕ satisfies detailed balance, we can deduce that $\pi \, P_\phi = \pi$, i.e.,

$$\int_\Omega \pi(x) \, P_\phi(x,y) \, dx \; = \; \pi(y).$$

Transitions can be generated using a pre-fixed sequence of move types ϕ. Or, the type of each transition may be selected at random (independently of the current state x).

In either case, the chain has ergodic distribution π, as long as the chain is irreducible.

In the second case, with ϕ generated from $f(\phi)$, the overall transition matrix is

$$P \; = \; \int_\Phi f(\phi) \, P_\phi \, d\phi.$$

A variety of move types

Thus, to show that $\pi P = \pi$, write

$$\begin{aligned}
\int_\Omega \pi(x)\, P(x,y)\, dx &= \int_\Omega \pi(x) \int_\Phi f(\phi)\, P_\phi(x,y)\; d\phi\; dx \\
&= \int_\Phi f(\phi) \int_\Omega \pi(x)\, P_\phi(x,y)\, dx\; d\phi \\
&= \int_\Phi f(\phi)\, \pi(y)\; d\phi \\
&= \pi(y).
\end{aligned}$$

A simple example of separate move types is the updating of single elements of the vector x. This is usually done systematically, cycling through the elements in order.

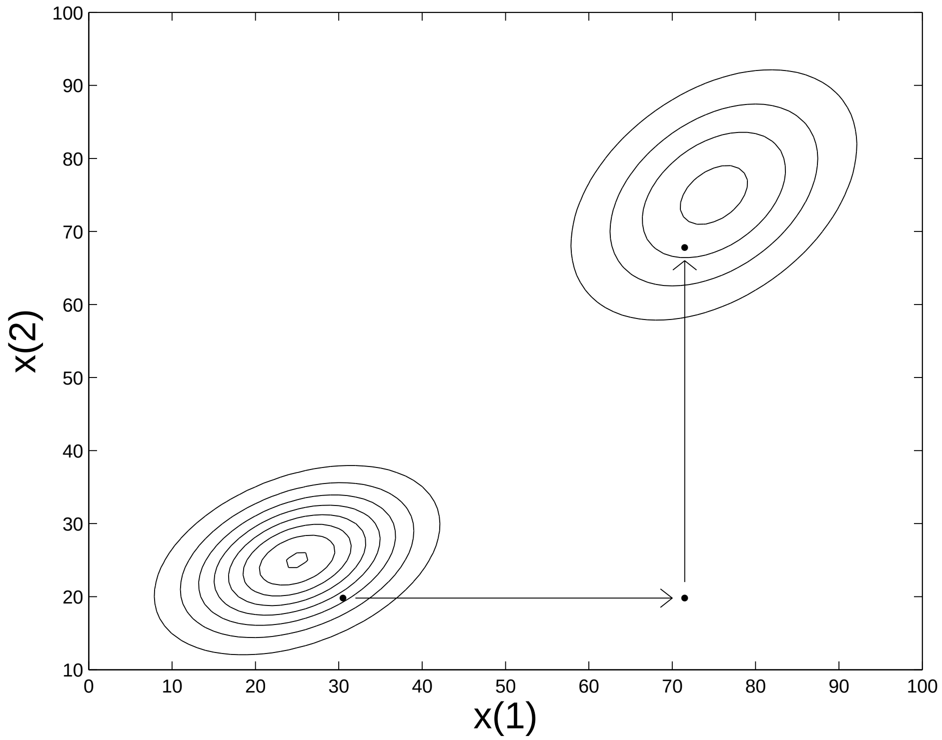
More interesting variety occurs if some moves operate locally, proposing small displacements in x, while other moves aim farther afield in the hope of a more substantial change in x.

2. Mixing problems

Efficient sampling needs π_n to converge rapidly to π.

The Markov chain must forget its initial state quickly and rapidly produce near-independent samples.

Problem 1. Two modes

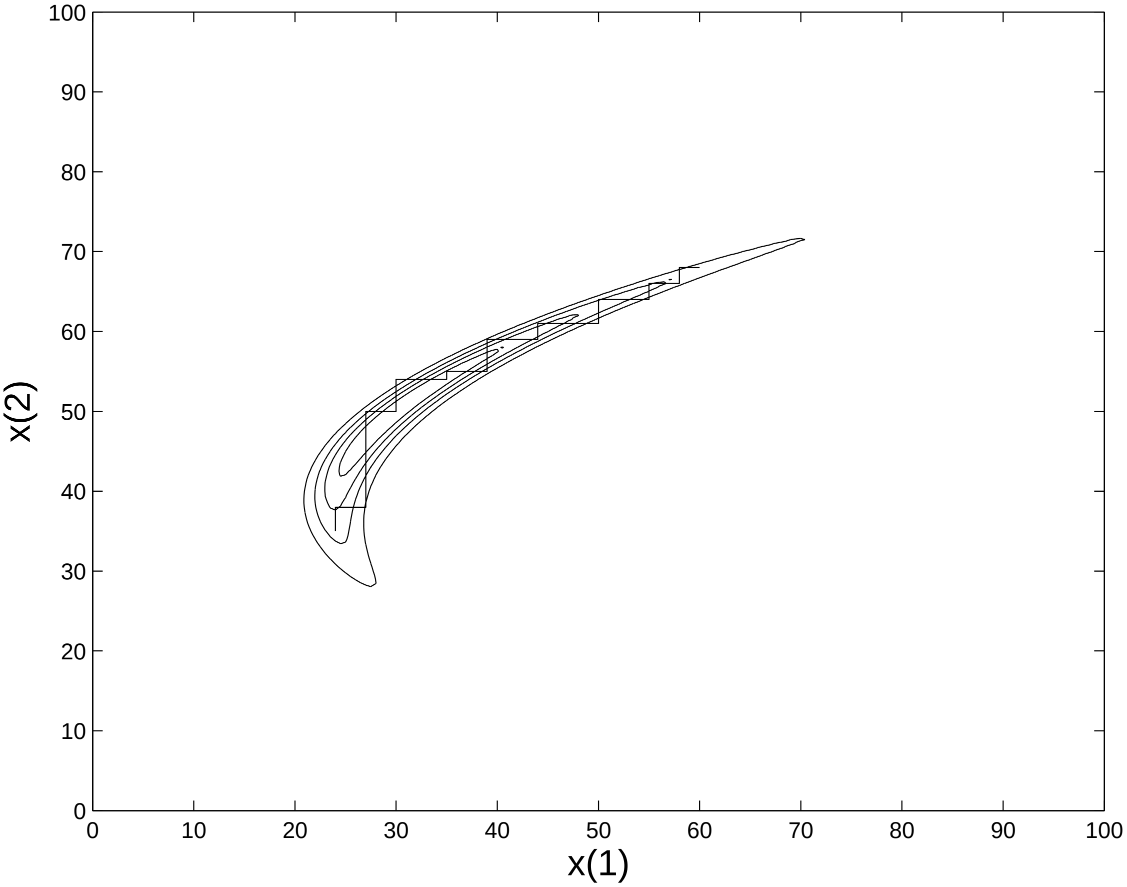


To move between modes, updating one element of x at a time, requires a visit to a state with very low $\pi(x)$ — and there is very little probability of accepting such a move.

Mixing problems

Problem 2. Very thin region of support for π

Example: Over-parameterisation, leading almost to a functional relation in parameters' posterior distribution.

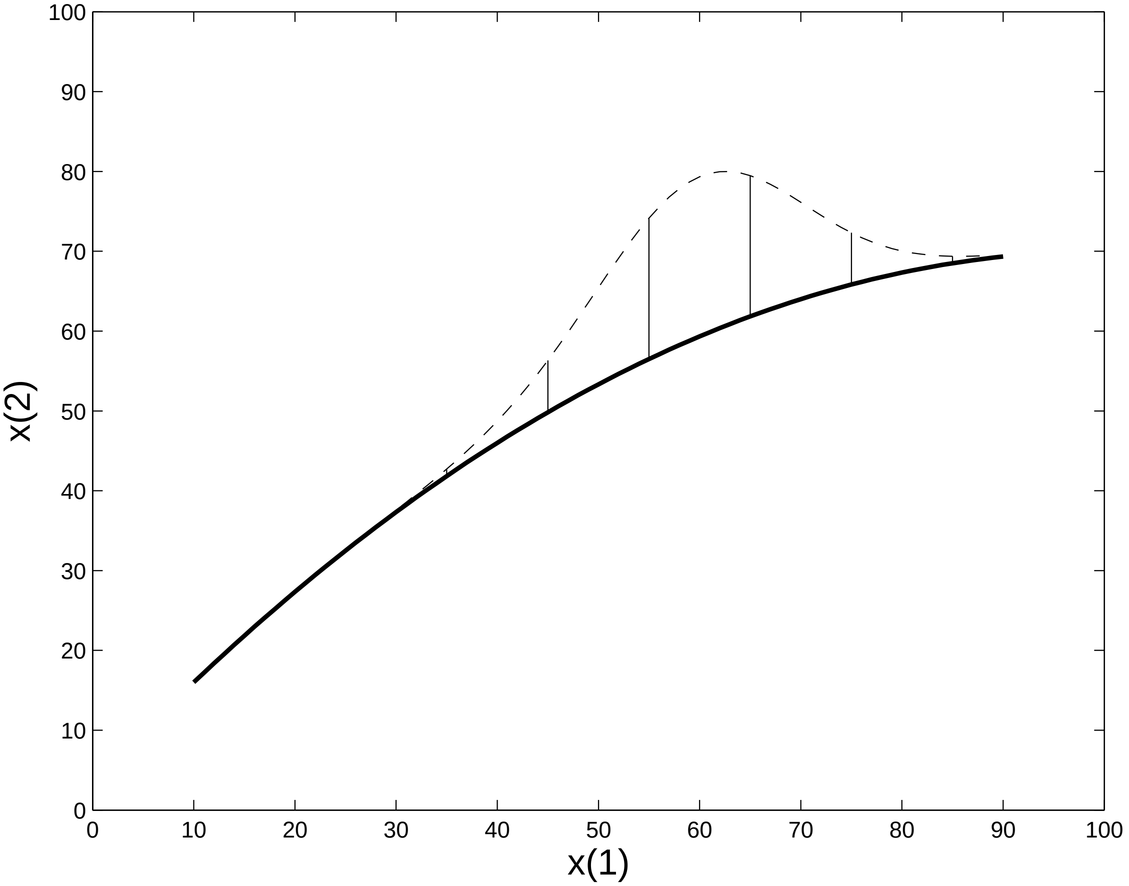


Traversing the modal region of π with updates of $x(1)$ and $x(2)$ requires a great many small steps.

Mixing problems

Problem 3. Limiting case of problem 2 — support of π confined to a sub-space of Ω

Example: An exact functional relation in the posterior distribution of a parameter vector.

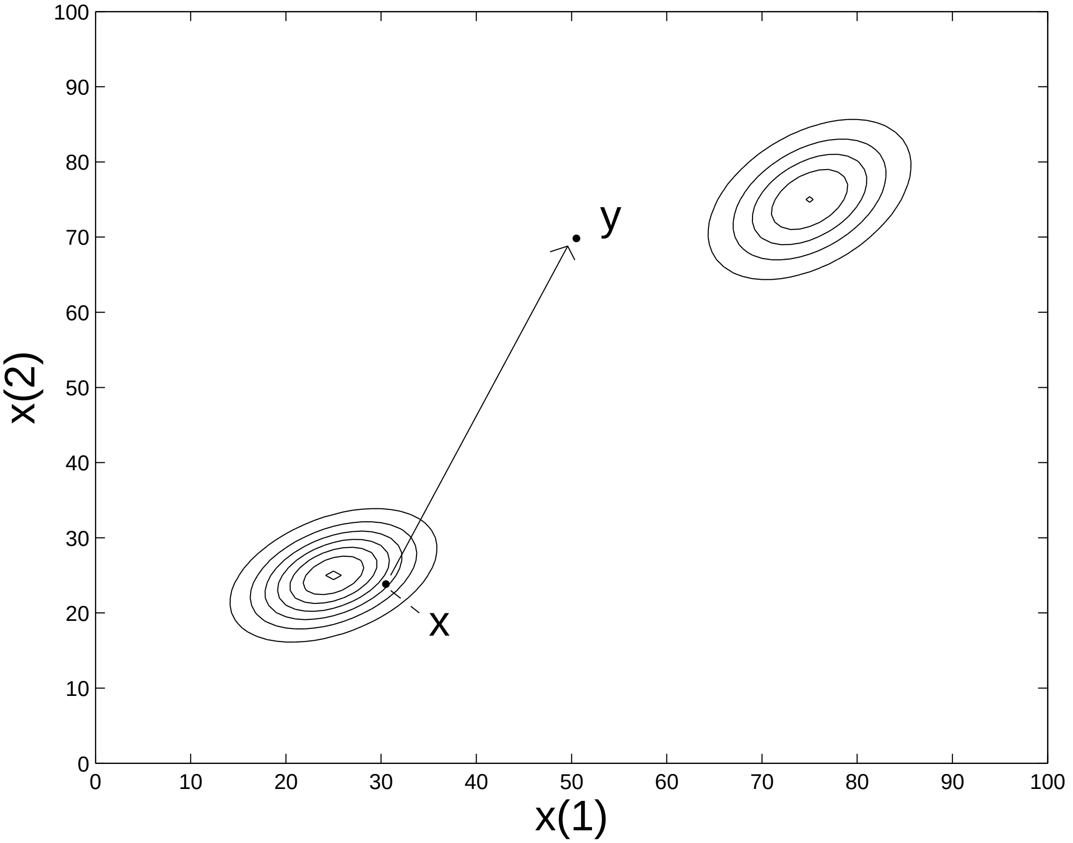


Can we apply MCMC methods to sample a distribution defined on a manifold?

3. Mode jumping

Modes of π are small in a high-dimensional space.

The kernel $q(x, y)$ generates jumps into the unknown.

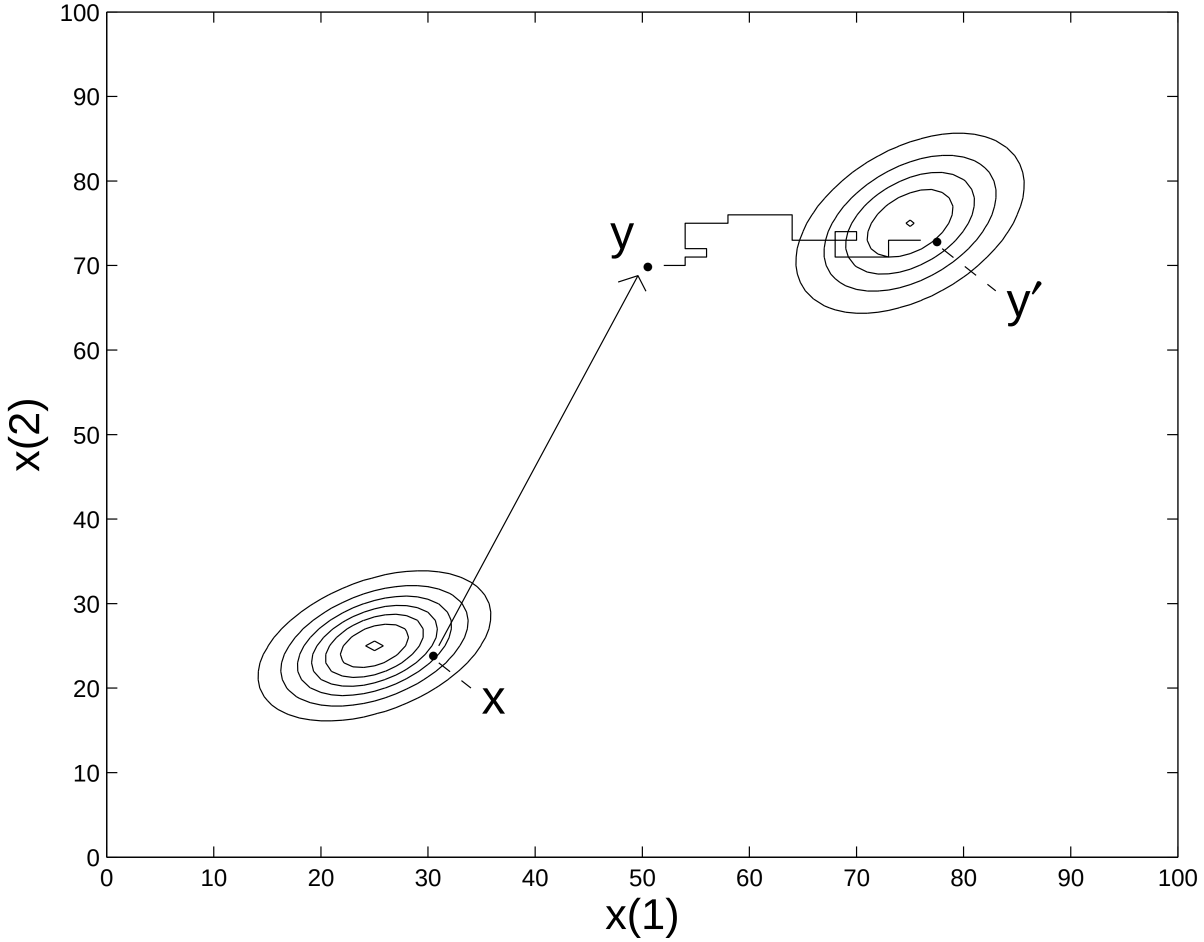


The current state x has fairly high $\pi(x)$ but proposals y are unlikely to hit the centre of another mode, so

$$\alpha(x, y) = \min\{1, \frac{\pi(y)\, q(y, x)}{\pi(x)\, q(x, y)}\} \approx 0.$$

Mode jumping

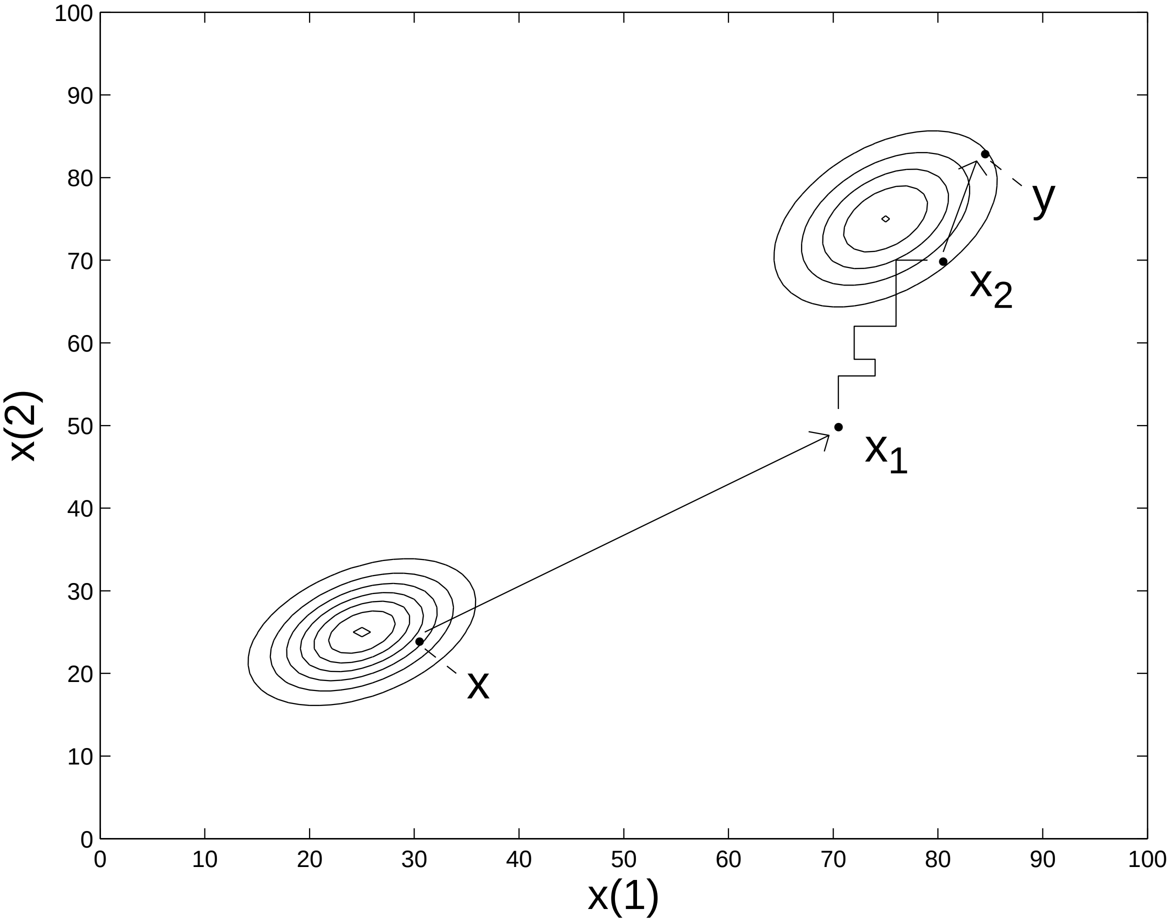
Paradox: There are simple ways to move from y to a y' with much higher $\pi(y')$, e.g., run n iterations of small-step M-H updates.



But then it is not feasible to calculate $q(x, y')$ for use in the formula for $\alpha(x, y')$.

Our proposed method

Creating the proposal, y



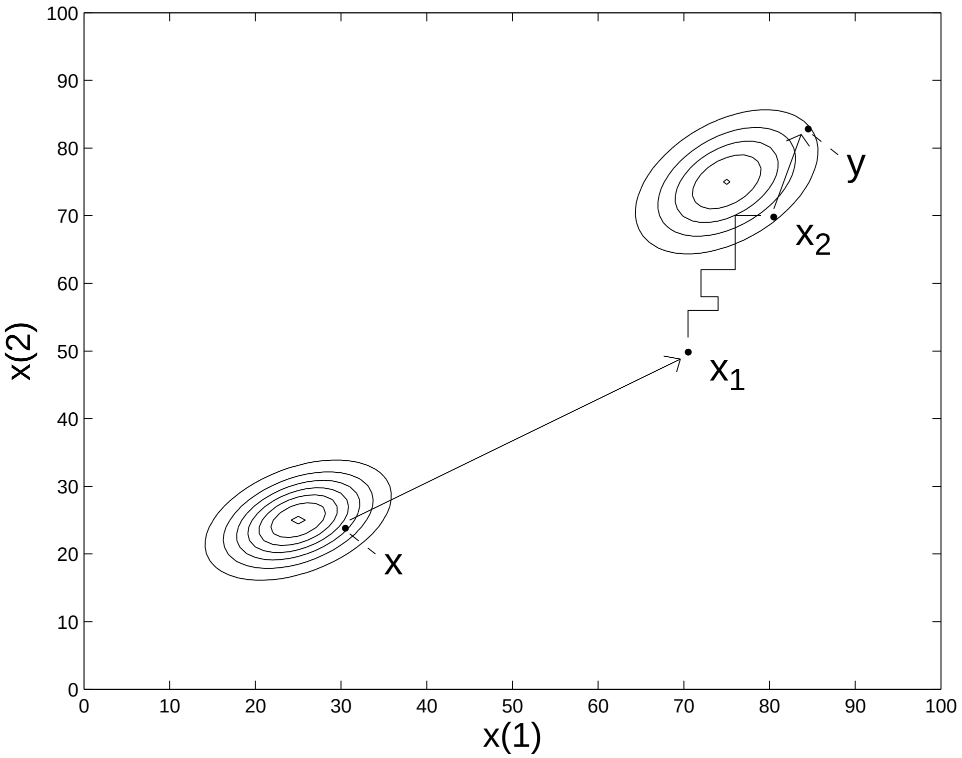
$$\phi \sim f(\phi)$$

$x_1 = x + \phi$ deterministic step,

$x_2 \sim g(x_1, x_2)$ e.g., n MCMC steps,

$y \sim h(x_2, y)$ h a local approximation to $\pi(y)$ near the point x_2.

Proposed method

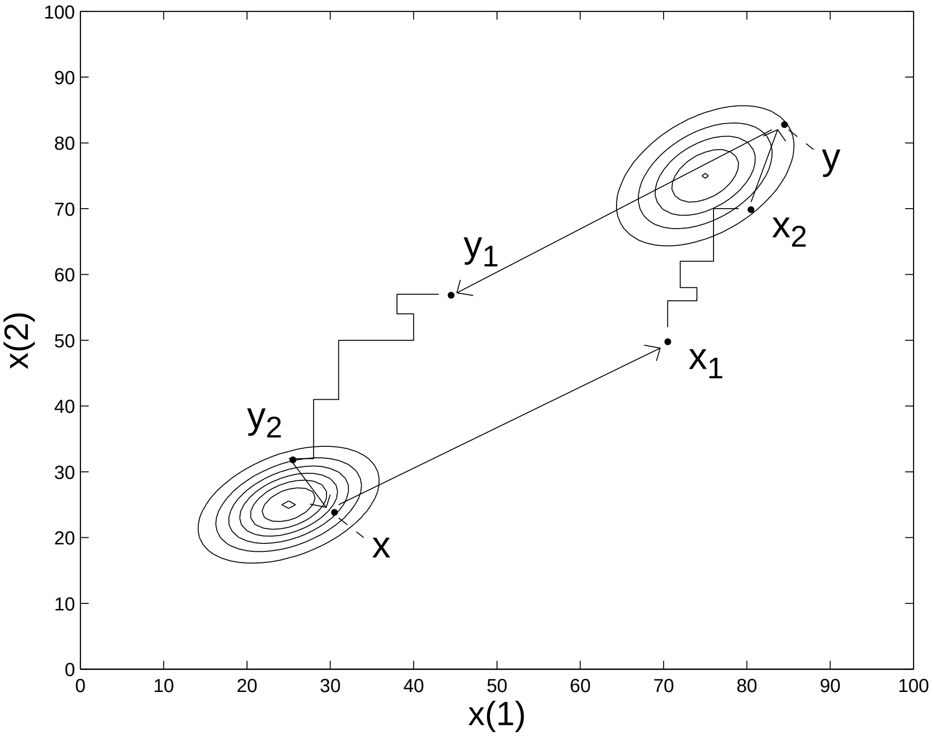


We need to find a suitable acceptance probability $\alpha(x, y)$

- NOT computing the overall distribution of y given x,
- but conditioning, in a sense, on ϕ and the random path from x_1 to x_2,
- leaving just $h(x_2, y)$.

The trick is to generate analogous parts of a return path from y to x and “condition” on these too.

Proposed method



$$\phi \sim f(\phi)$$

$$x_1 = x + \phi \qquad y_1 = y - \phi$$

$$x_2 \sim g(x_1, x_2) \qquad y_2 \sim g(y_1, y_2)$$

$$y \sim h(x_2, y) \qquad x \sim h(y_2, x).$$

Assuming $f(\phi) = f(-\phi)$, set acceptance probability for y, $\alpha(x, y)$, to be

$$\alpha_{\phi,x_1,x_2,y_1,y_2}(x, y) = \min\{1, \frac{\pi(y)\, h(y_2, x)}{\pi(x)\, h(x_2, y)}\}.$$

Proposed method

Note that for this choice of $\alpha(x,y)$,

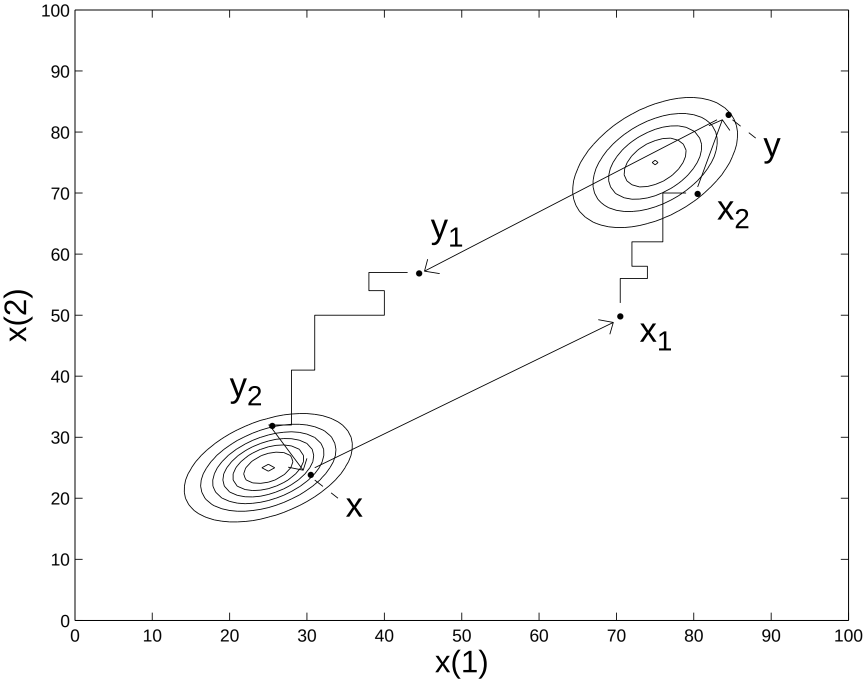
$$\begin{aligned}
&\pi(x)\; h(x_2,y)\; \alpha_{\phi,x_1,x_2,y_1,y_2}(x,y) \\
&\quad = \pi(x)\; h(x_2,y)\;\; \mathrm{min}\{1,\, \frac{\pi(y)\, h(y_2,x)}{\pi(x)\, h(x_2,y)}\} \\
&\quad = \pi(y)\; h(y_2,x)\;\; \mathrm{min}\{1,\, \frac{\pi(x)\, h(x_2,y)}{\pi(y)\, h(y_2,x)}\} \\
&\quad = \pi(y)\; h(y_2,x)\; \alpha_{-\phi,y_1,y_2,x_1,x_2}(y,x)
\end{aligned}$$

— by the same piece of algebra that justifies the standard Metropolis-Hastings algorithm.

Proof of detailed balance

We need to show that, for $x \neq y$,

$P\{$At x under π then $\rightarrow y\} \;=\; P\{$At y under π then $\rightarrow x\}$.



Here

$$\text{LHS} \;=\; \pi(x) \int d\phi\, f(\phi) \int dx_2\; g(x_1, x_2)\; h(x_2, y)$$

$$\int dy_2\; g(y_1, y_2)\; \alpha_{\phi, x_1, x_2, y_1, y_2}(x, y),$$

where $x_1 = x + \phi$ and $y_1 = y - \phi$.

Proof of detailed balance

Rearranging terms,

$$\text{LHS} = \pi(x) \int d\phi\, f(\phi) \int dx_2\, g(x_1, x_2)\, h(x_2, y)$$

$$\int dy_2\, g(y_1, y_2)\, \alpha_{\phi, x_1, x_2, y_1, y_2}(x, y)$$

$$= \int d\phi\, f(\phi) \int dx_2\, g(x_1, x_2) \int dy_2\, g(y_1, y_2)$$

$$\pi(x)\, h(x_2, y)\, \alpha_{\phi, x_1, x_2, y_1, y_2}(x, y)$$

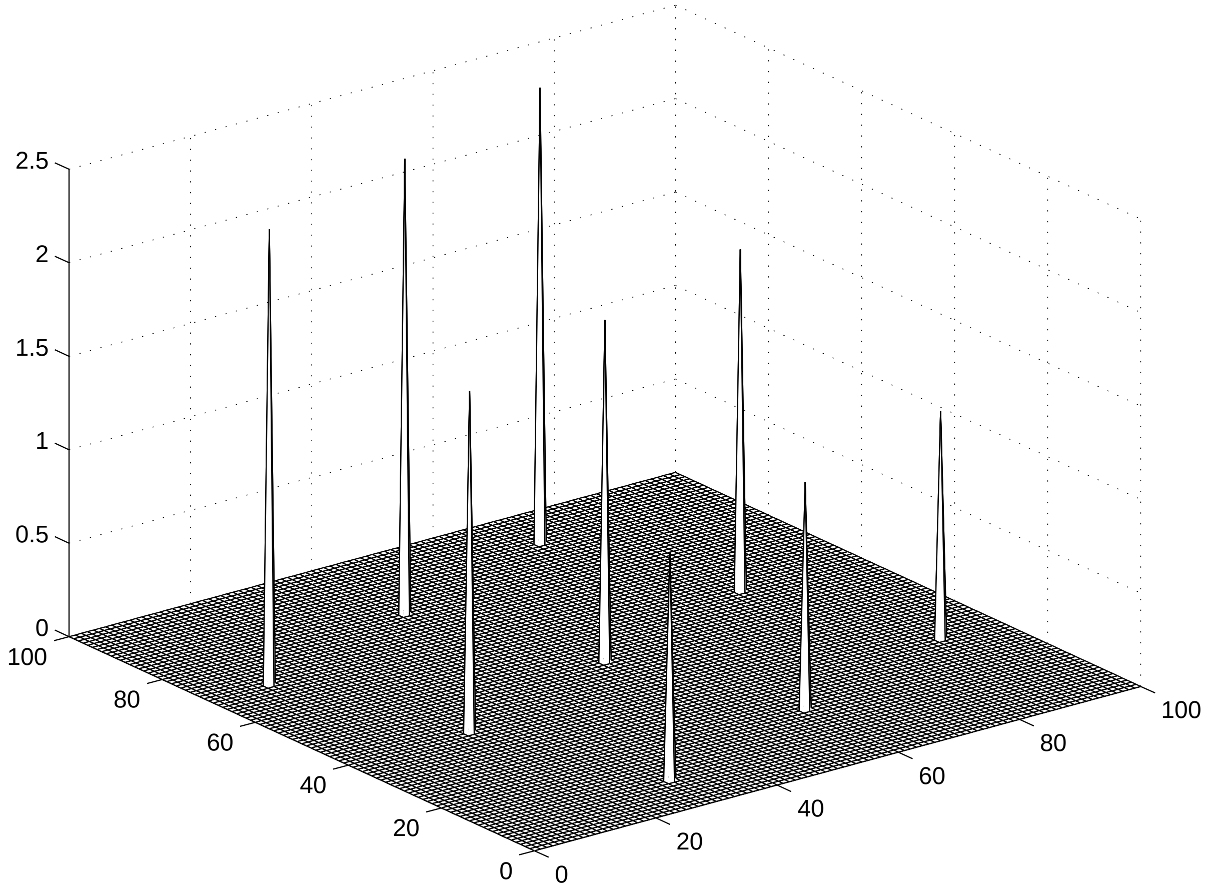
$$= \int d\phi\, f(\phi) \int dx_2\, g(x_1, x_2) \int dy_2\, g(y_1, y_2)$$

$$\pi(y)\, h(y_2, x)\, \alpha_{-\phi, y_1, y_2, x_1, x_2}(y, x)$$

$$= \pi(y) \int d(-\phi)\, f(-\phi) \int dy_2\, g(y_1, y_2)\, h(y_2, x)$$

$$\int dx_2\, g(x_1, x_2)\, \alpha_{-\phi, y_1, y_2, x_1, x_2}(y, x) = \text{RHS.}$$

4. Example 1



Mixture of 9 bivariate normal distributions in window $(0, 100) \times (0, 100)$. Each bivariate normal distribution has $Var(X(1)) = Var(X(2)) = 0.1^2$ and correlation $\rho = 0.25$.

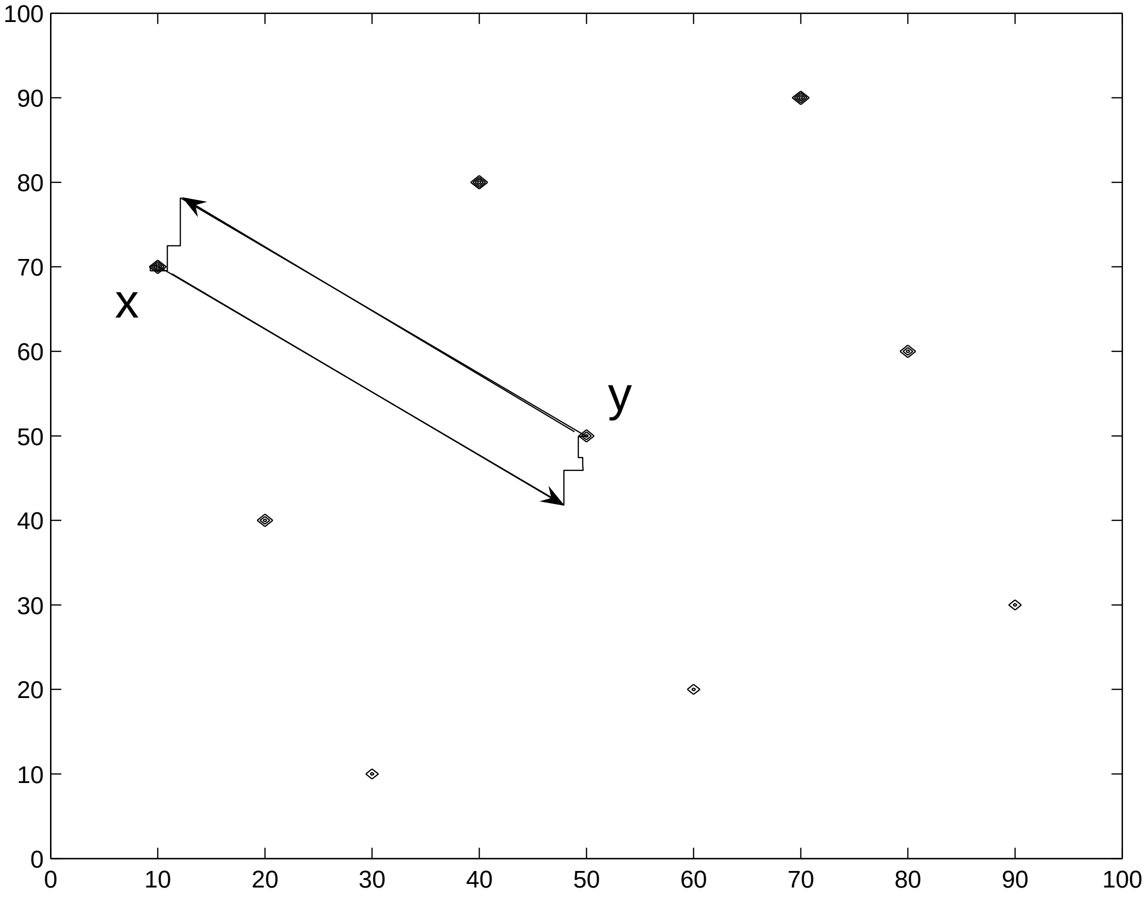
Plan to use two move types: (1) small changes within a mode and (2) jumps between modes.

Example 1

For large steps, move to a point chosen uniformly within a circle of radius 50, centred on the current state.

1. It is not surprising that large steps with no adaptation are nearly always rejected.

2. With adaptation, large steps are frequently accepted:



Example 1: Details of adaptation

1) The random walk stage “x_1 to x_2” comprises 10 M-H updates of both elements of x with $N(0, 4^2)$ proposals followed by 10 updates with $N(0, 1)$ proposals.

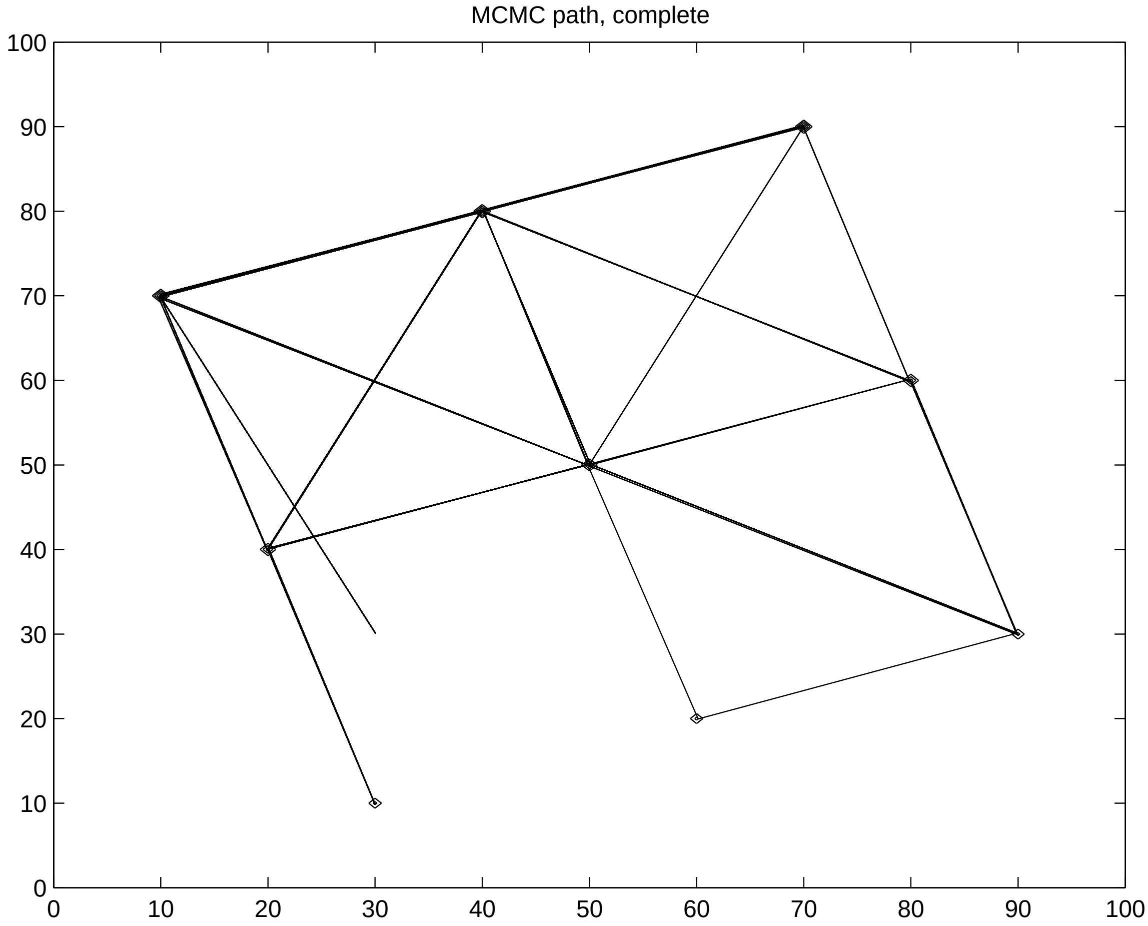
In both cases, the target distribution is an “annealed” version of π proportional to $\pi(x)^5$.

2) In calculating the local approximation to π at x_2, we fit a univariate distribution in each component direction, matching the density π at three points.

The two univariate distributions are combined assuming the two components of X to be independent — to demonstrate that this local fit will often be only an approximation.

Example 1: Results

After 1000 iterations, shared between small steps, large steps (unadapted), and adaptive large steps:



A total of 42 out of 333 adaptive large steps were successful in giving an accepted jump between modes.

Example 1: Results

From a longer simulation, we find:

	Success rate	*Average no. of function evaluations*	*Evaluations per mode jump*
Unadapted	0.00009	1	11,111
Adapted	0.20	59	295

Hence, the adapted method improves on the unadapted method's efficiency by a factor of 38.

Higher dimensional versions of Example 1

Suppose $\pi(x)$ is a distribution on $x \in \Re^n$ and the form of π within each dimension does not change with n.

Unadapted method

The volume of the mode decreases exponentially in n.

The expected number of attempts before a random shot hits the mode increases exponentially in n.

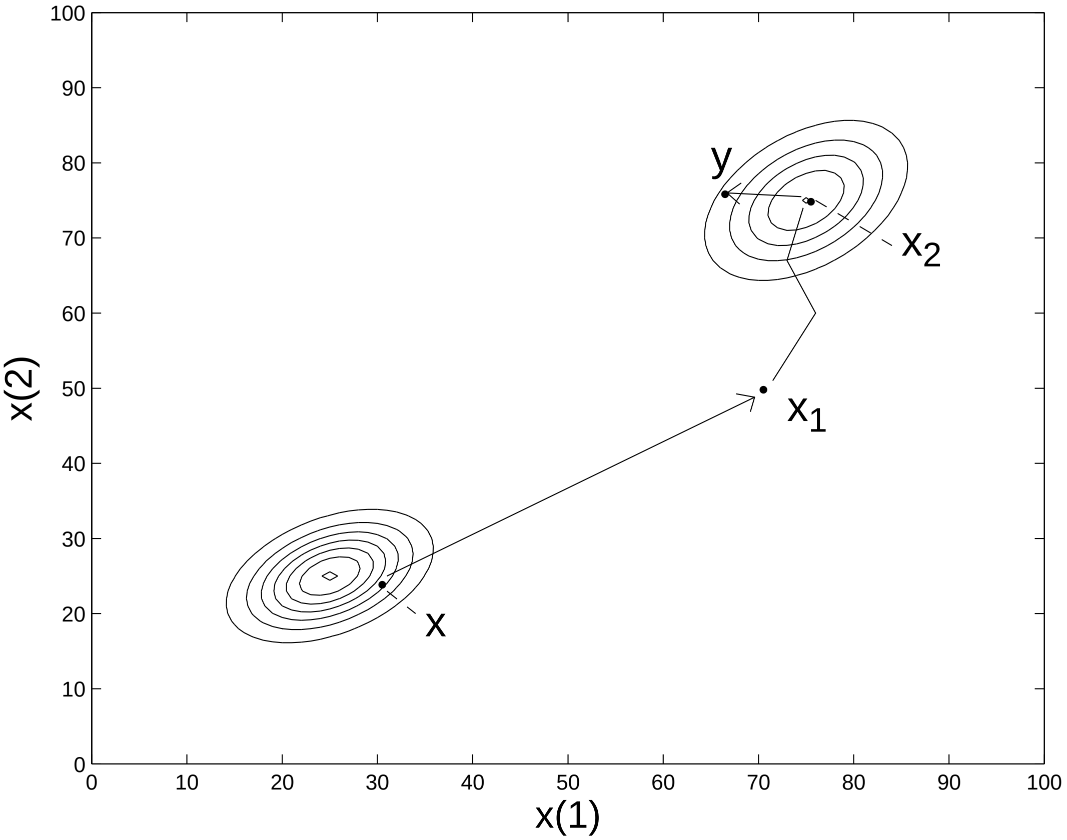
Adapted method

The number of function evaluations for a useful rate of mode-jumping is liable to scale linearly in n.

Relative efficiency of the adapted method increases substantially with n, e.g., to values of 1,000 or 50,000.

For more awkward distributions π, the adapted method may need longer MCMC sequences to allow time to travel to a mode and find an x with high $\pi(x)$ within the mode.

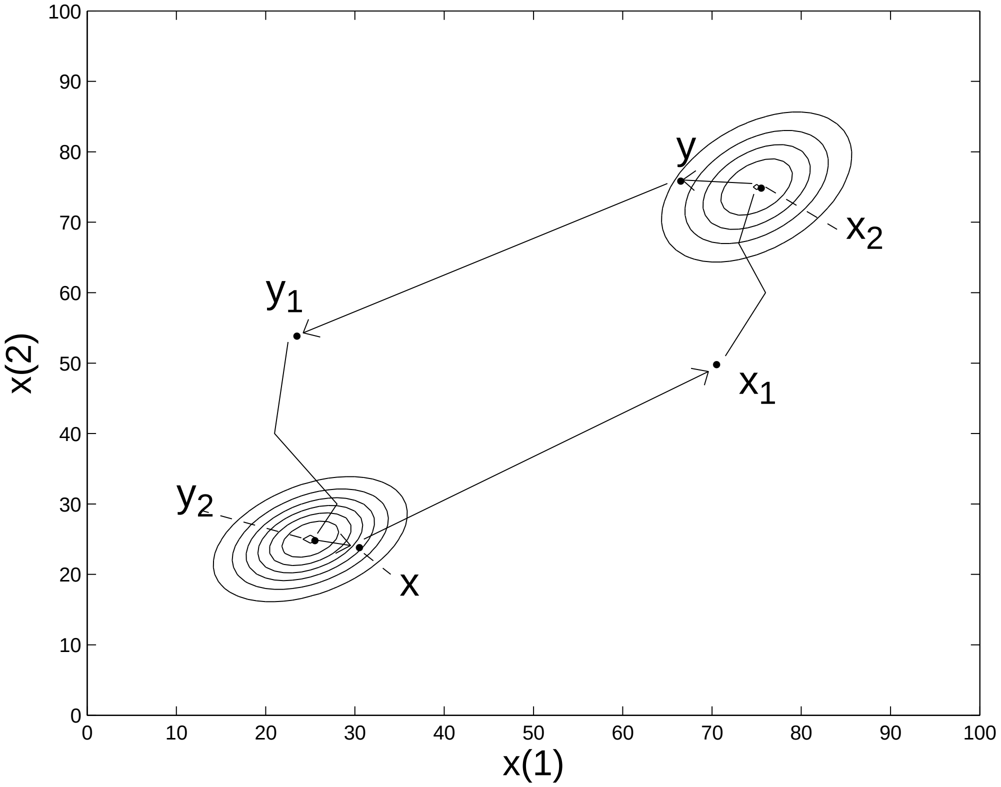
5. Relation to Tjelmeland & Hegstad (2001)



T & H's algorithm has steps:

1. Large step from x to $x_1 = x + \phi$.
2. Hill climbing from x_1 to x_2.
3. Sample y from $h(x_2, y)$, an approximation to $\pi(y)$ at x_2.

Relation to Tjelmeland & Hegstad



4. Construct a reverse step to $y_1 = y - \phi$.

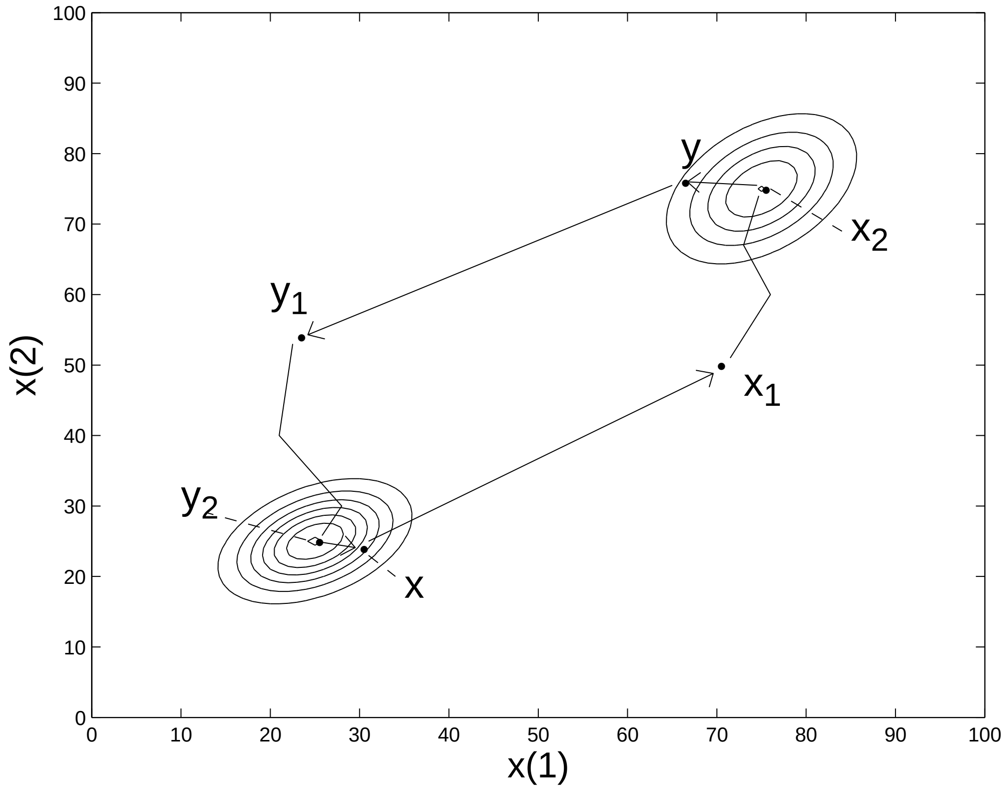
5. Hill climbing from y_1 to y_2.

6. Fit a local approximation $h(y_2, x)$ to $\pi(x)$ at y_2.

7. Accept the move from x to y with probability

$$\alpha(x, y) = \min\{1, \frac{\pi(y)\, h(y_2, x)}{\pi(x)\, h(x_2, y)}\}.$$

Relation to Tjelmeland & Hegstad

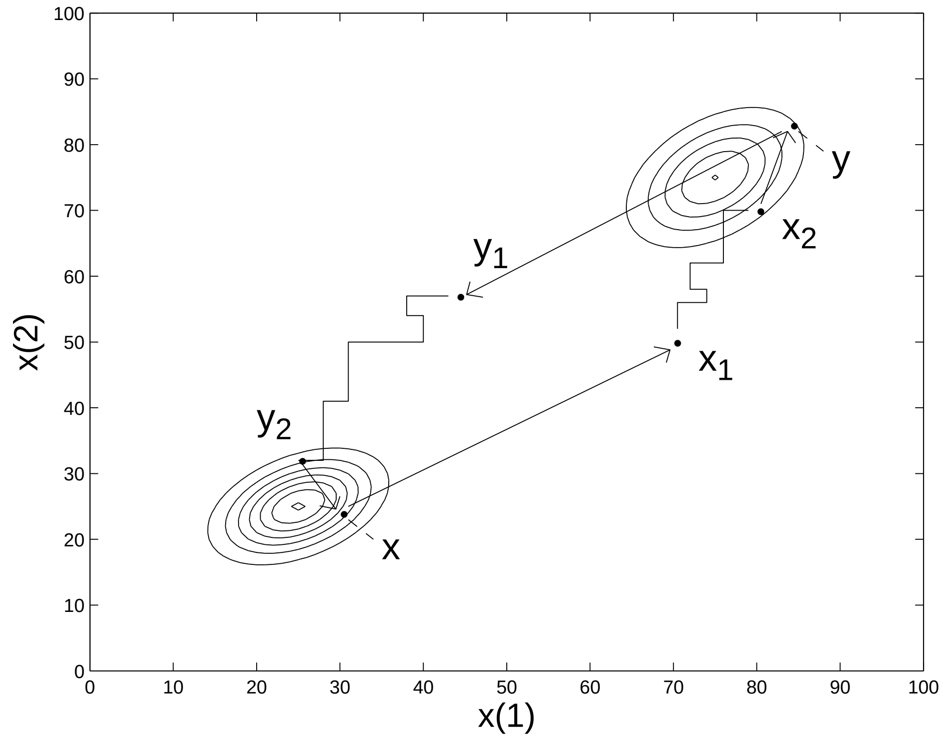


T & H prove detailed balance within move type ϕ (defined to include a random choice of step $+\phi$ and $-\phi$).

The path from x_1 to x_2 is deterministic, so randomness in y is only in sampling from $h(x_2, y)$.

The return path has to be constructed in order to discover where the hill climbing algorithm from y_1 reaches, and to compute the local approximation to π there, $h(y_2, x)$.

Relation to Tjelmeland & Hegstad



Richard argued that, in our method, the random paths $x_1 \rightarrow x_2$, for all possible values x_1, can be included in the definition of the “move type”.

Hence, T & H’s proof is enough to prove detailed balance when the $x_1 \rightarrow x_2$ transition is stochastic!

The formal proof, with multiple integrals, establishes this to be true but our motivation was from Richard’s less standard argument.

6. Harder examples

Key to applying our method is the ability to construct a good approximation $h(x_2, y)$ to $\pi(y)$ near x_2.

Suppose π is a mixture distribution with k modes,

$$\pi(y) = \sum_k p_k \, f_k(y),$$

where the f_k are probability densities and $\Sigma p_k = 1$.

Suppose also we are able to obtain $h(x, y) \approx f_k(y)$ when y is near mode k — note, we do not find the factor p_k.

For x and y_2 near mode k and x_2 and y near mode k',

$$\begin{aligned} \alpha(x,y) &= \min\{1, \frac{\pi(y)\ h(y_2,x)}{\pi(x)\ h(x_2,y)}\} \\ &\approx \min\{1, \frac{p_{k'}\ f_{k'}(y)\ \ f_k(x)}{p_k\ f_k(x)\ \ f_{k'}(y)}\} \\ &= \min\{1, \frac{p_{k'}}{p_k}\}, \end{aligned}$$

which is just right for moves between these two modes.

Harder examples

Implicit construction of $\boldsymbol{h(x_2, y)}$

For a discrete distribution π

Apply a cycle of the Gibbs sampler starting in state x_2 and updating each element of x in turn.

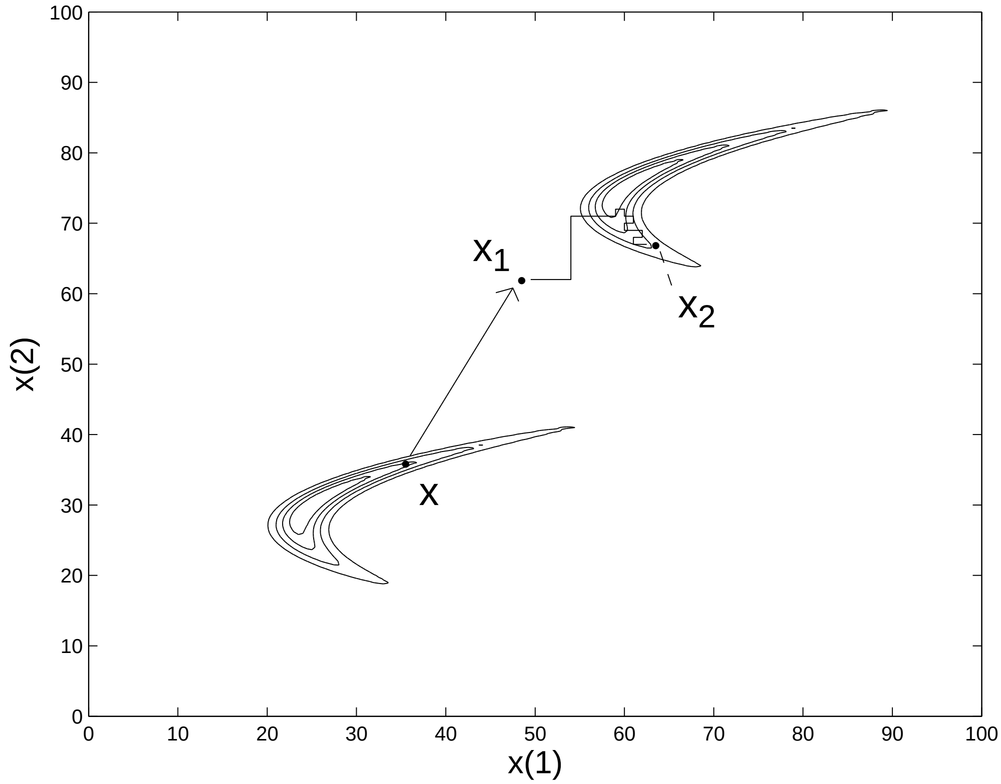
Multiplying together the conditional distributions gives the distribution $h(x_2, y)$ which has, implicitly, been sampled.

For π a continuous distribution

An analogous procedure can be implemented, sampling from a local approximation to the conditional distribution of each element at each stage.

Difficulties arise when the product of conditional distributions is not a good local approximation to π.

Example 2

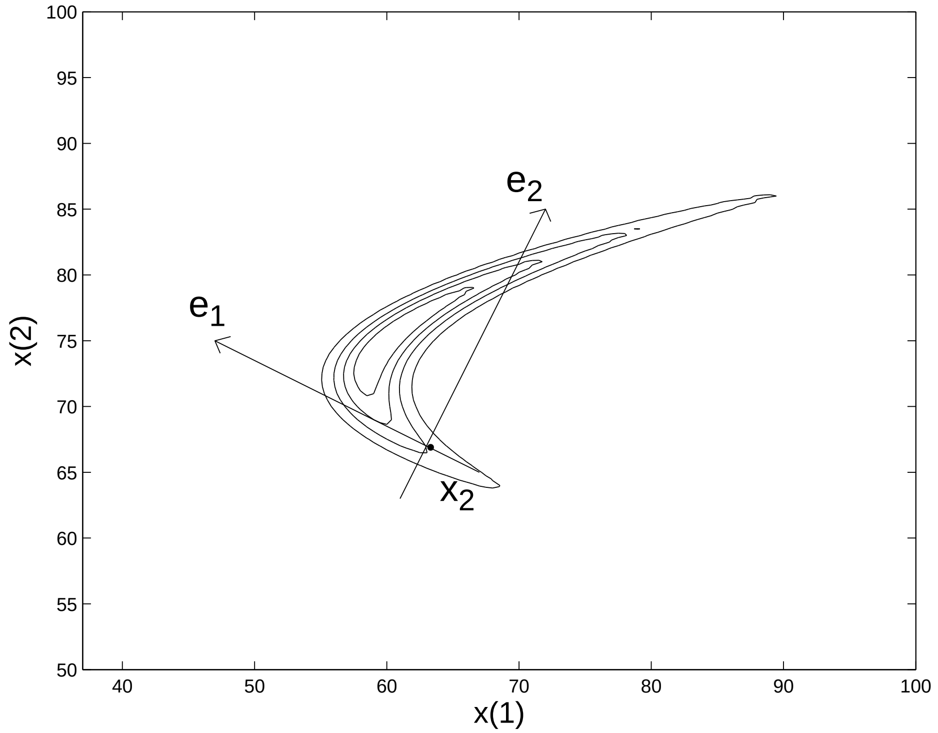


Having reached x_2, fitting $h(x_2, y)$ as a multivariate normal distribution, using values of π near x_2, will *not* give a good local approximation to $\pi(y)$.

1) Local information *can* give a good choice of directions for a sequence of updates.

2) It will be wise to wait for each update before fitting an approximation to π in a new direction.

Example 2



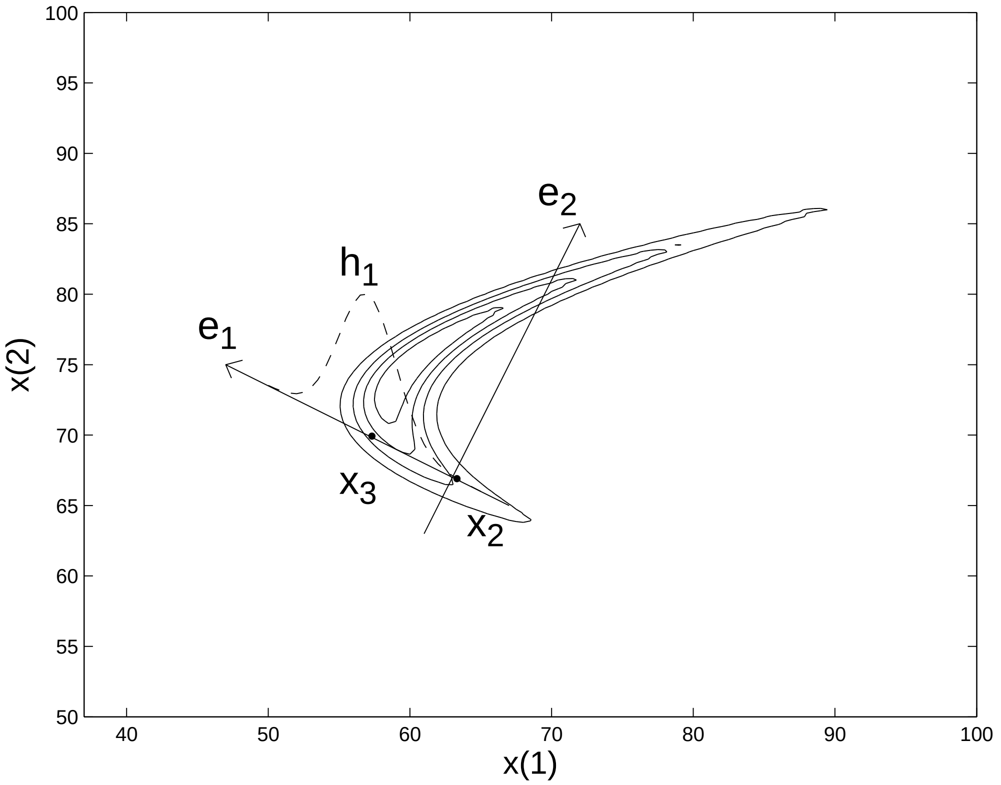
A sampling scheme to generate $\boldsymbol{h(x_2, y)}$

1. Fit a multivariate normal distribution to π, using values of π at points near x_2. In fact, we only need the fitted variance matrix — which is easy to obtain.

2. Find principal components of the variance matrix and use these to define directions e_1 and e_2.

3. Fit a univariate normal distribution, h_1, approximating the conditional distribution of π along the e_1 axis.

Example 2



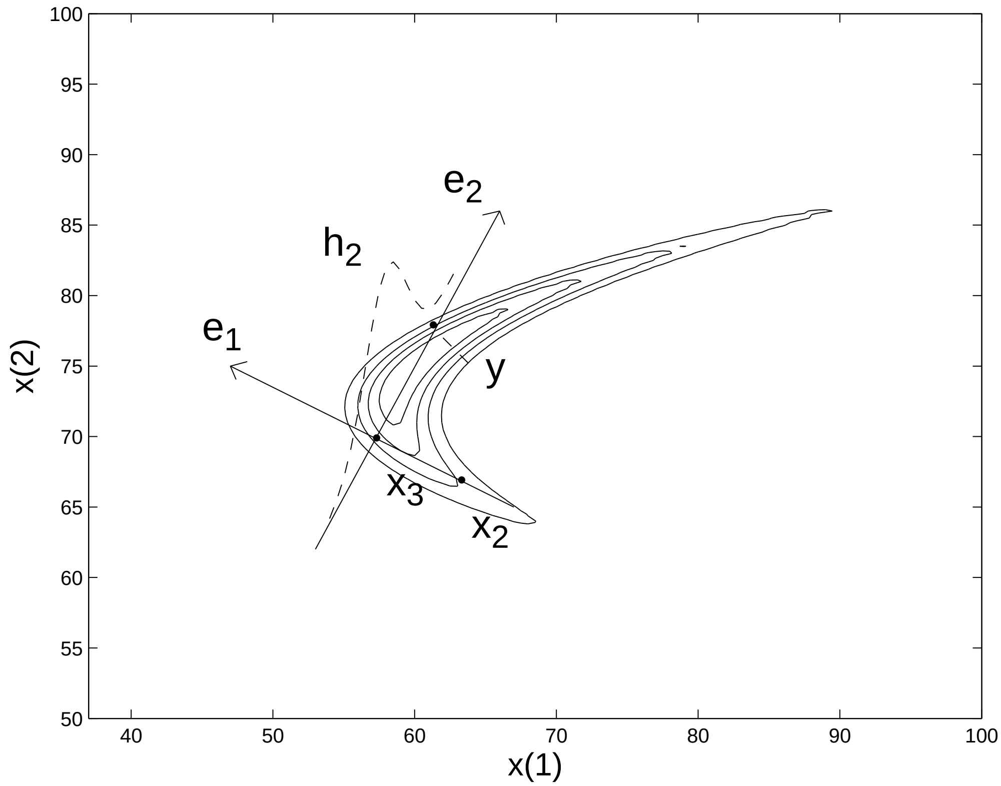
Sampling scheme continued

4. Sample from the conditional distribution, h_1, of π along the e_1 axis to produce the point x_3.

5. Note the density $h_1(x_3)$.

6. Fit a univariate normal distribution, h_2, approximating π along the line through x_3 in direction e_2.

Example 2



Sampling scheme continued

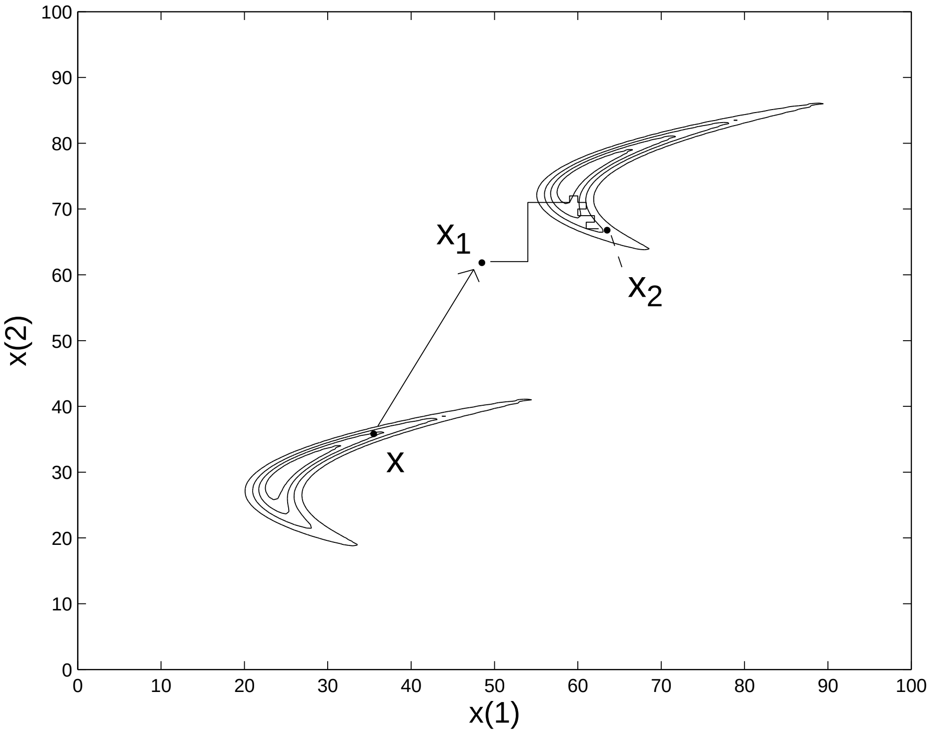
7. Sample the conditional distribution, h_2, of π along the line through x_3 in direction e_2 to produce the proposal y.

8. Note the density $h_2(y)$.

9. Combine the conditional densities to obtain

$$h(x_2, y) = h_1(x_3)\, h_2(y).$$

Results for Example 2



Implementing this scheme for distributions in $\Re^2$, $\Re^4$ and $\Re^6$ led to earlier remarks on high-dimensional problems:

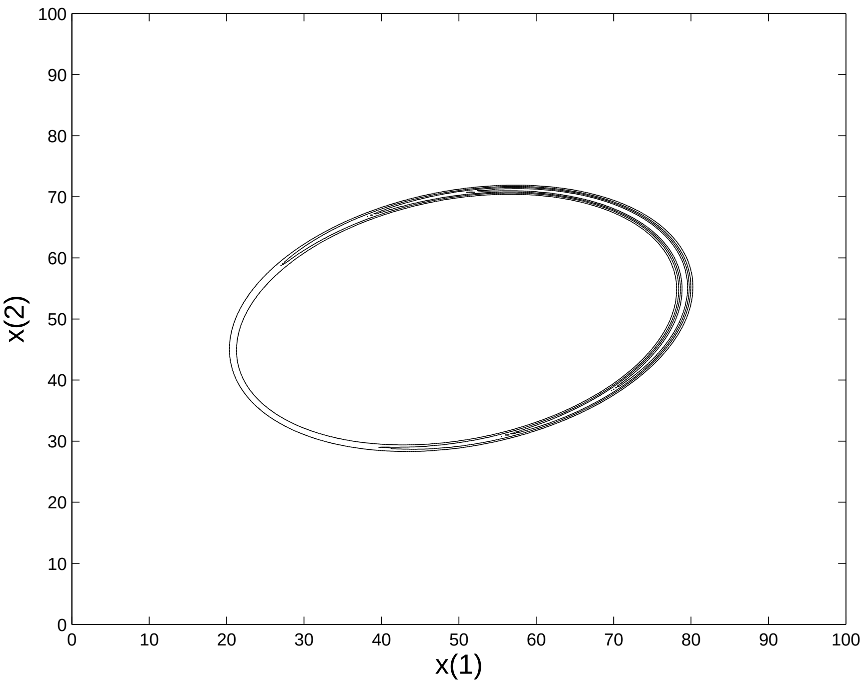
If modes of π occupy a very small volume of the state space, adapted updates offer a large efficiency gain.

Adaptation needs longer MCMC sequences to find modal states if modes exhibit curvature and high correlations.

Oddly shaped “basins of attraction” can imply a forward and return step do not come back to the original mode.

7. Distributions with “thin” support

The scheme for constructing a sample from $h(x_2, y)$ is of use in its own right to sample within a single mode.

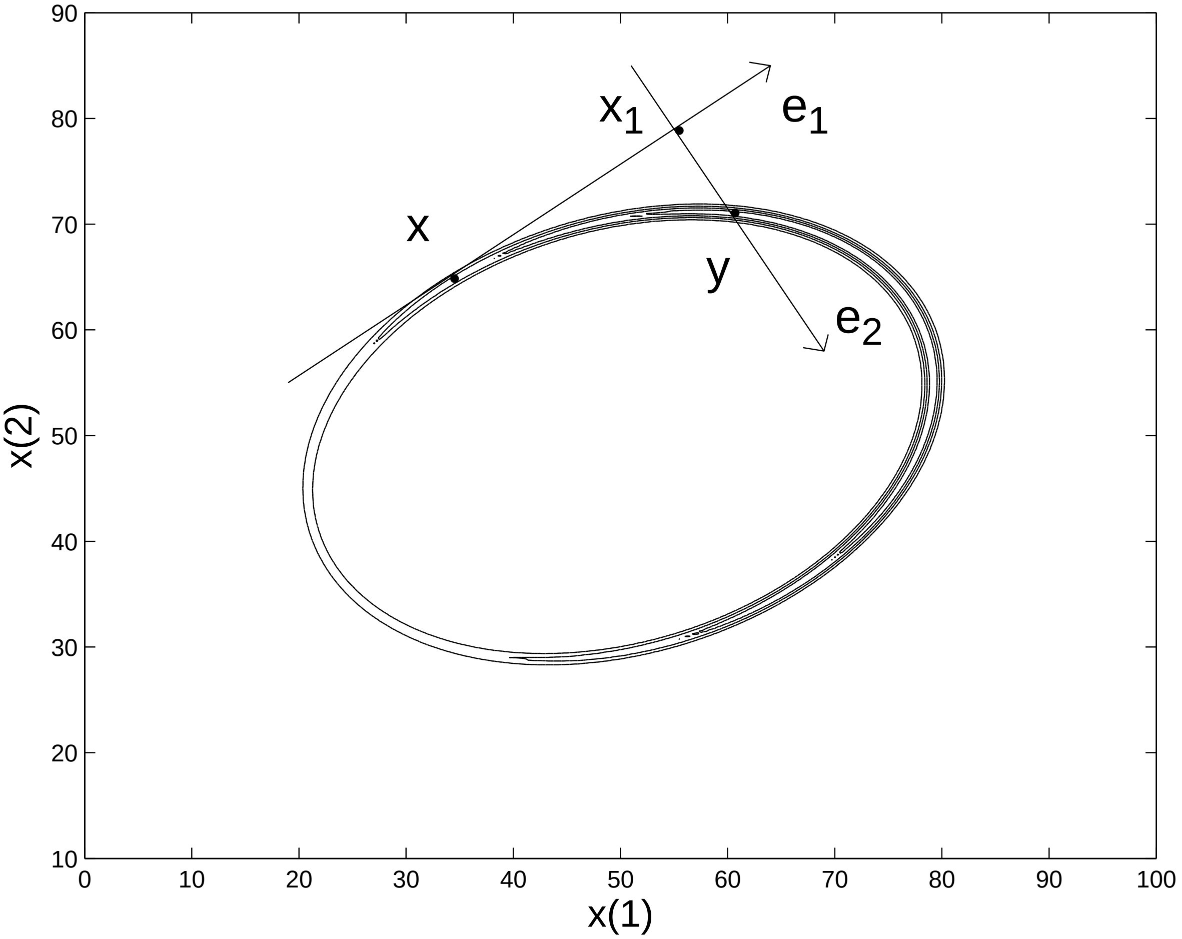


If elements of x are almost deterministically related, π is nearly confined to a lower dimensional sub-space of $\Re^n$.

M-H updates are hard to achieve and, even then, only very small steps are likely to be accepted.

N.B. You don’t know the distribution looks like this, all you can do is calculate the value $\pi(x)$ for chosen x values.

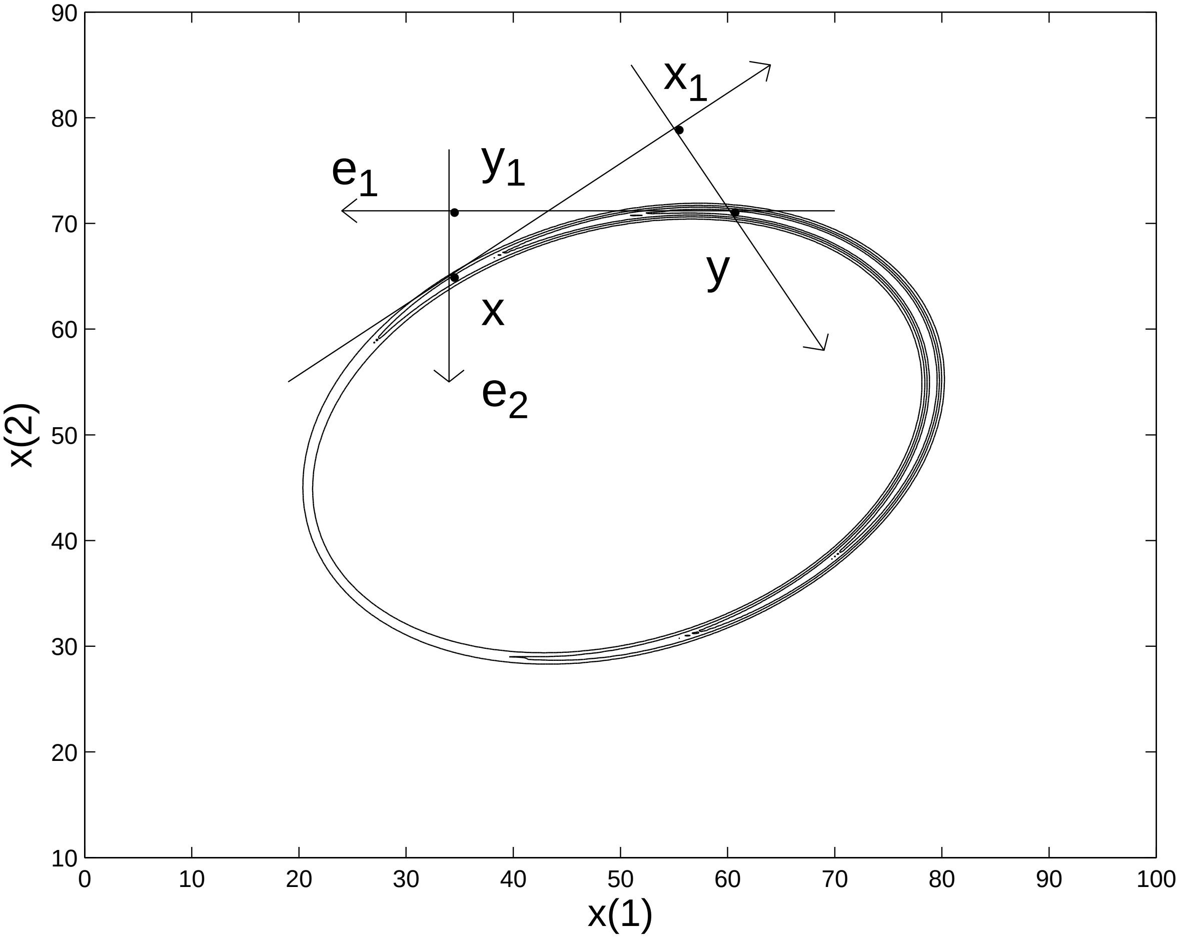
Distributions with thin support



The “$h(x_2, y)$” construction gives the whole proposal step.

An initial series of M-H steps along the e_2 direction can be included, to get nearer to the thin region of support. Like the previous $g(x_1, x_2)$, the probability of this path does not appear in $\alpha(x, y)$.

Distributions with thin support

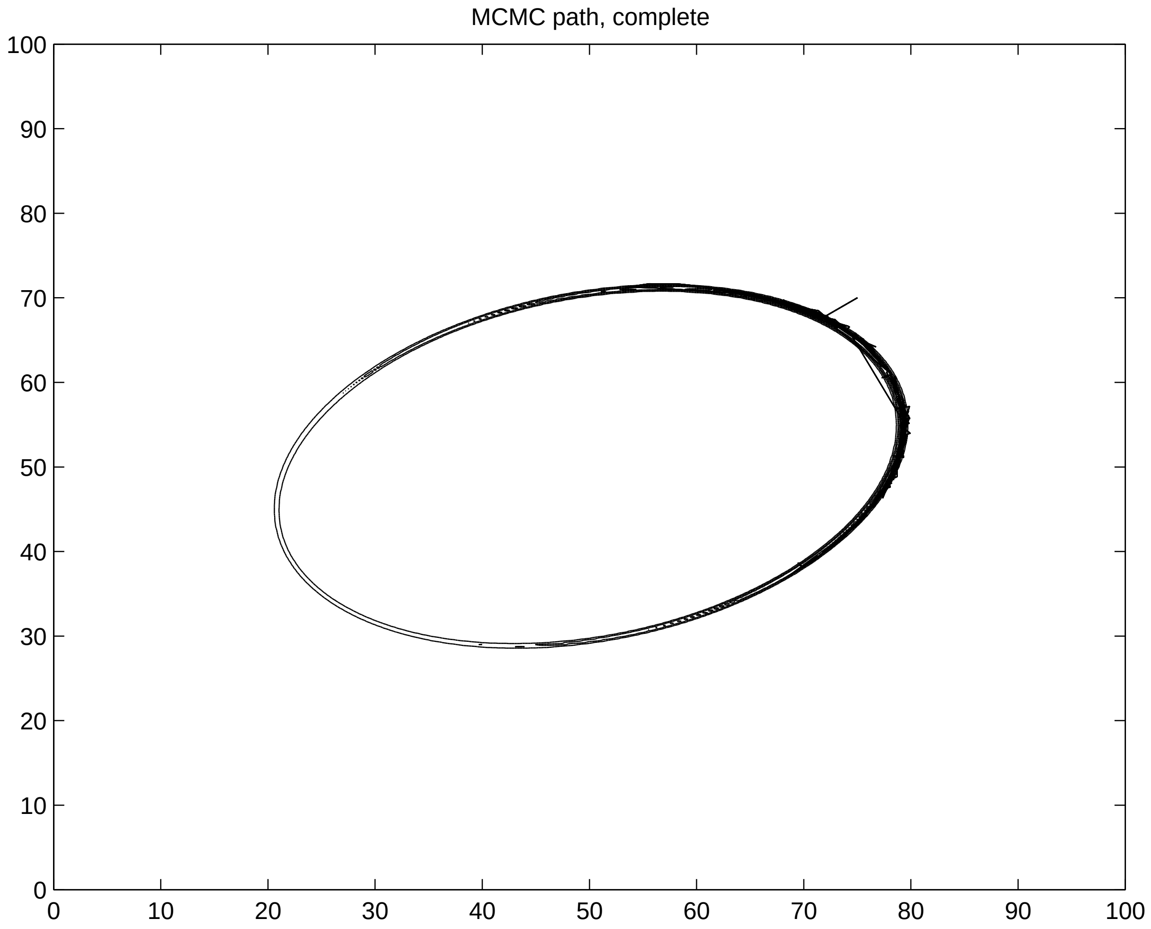


In the reverse step, the directions for the sequence of updates are fixed, as is the displacement from y to x.

The acceptance probability is

$$\alpha(x,y) = \min\{1, \frac{\pi(y)\ h_1(y,y_1)\,h_2(y_1,x)}{\pi(x)\ h_1(x,x_1)\,h_2(x_1,y)}\}.$$

Distributions with thin support

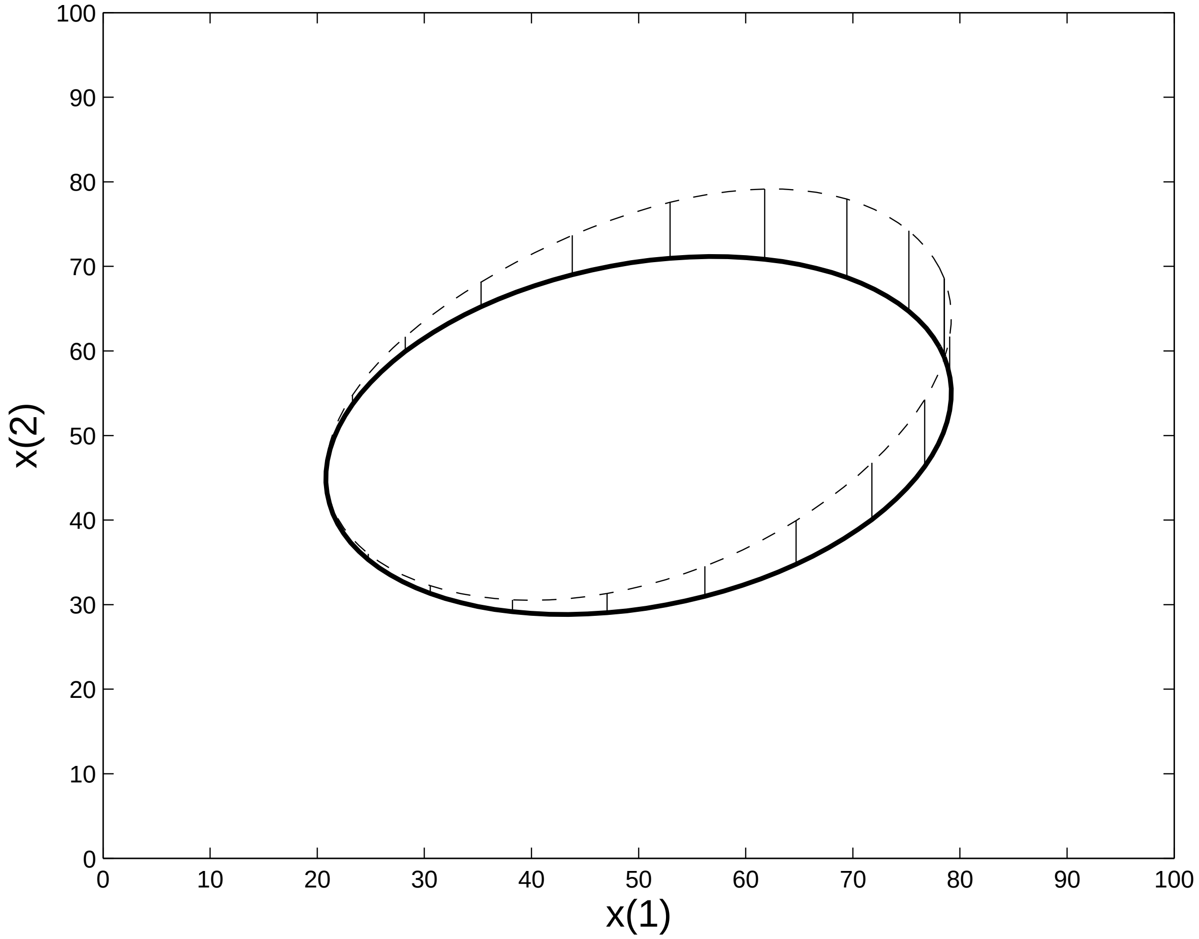


The path from a sequence of 1000 adaptive updates.

The acceptance rate for these proposals was 42%.

8. Distributions on manifolds

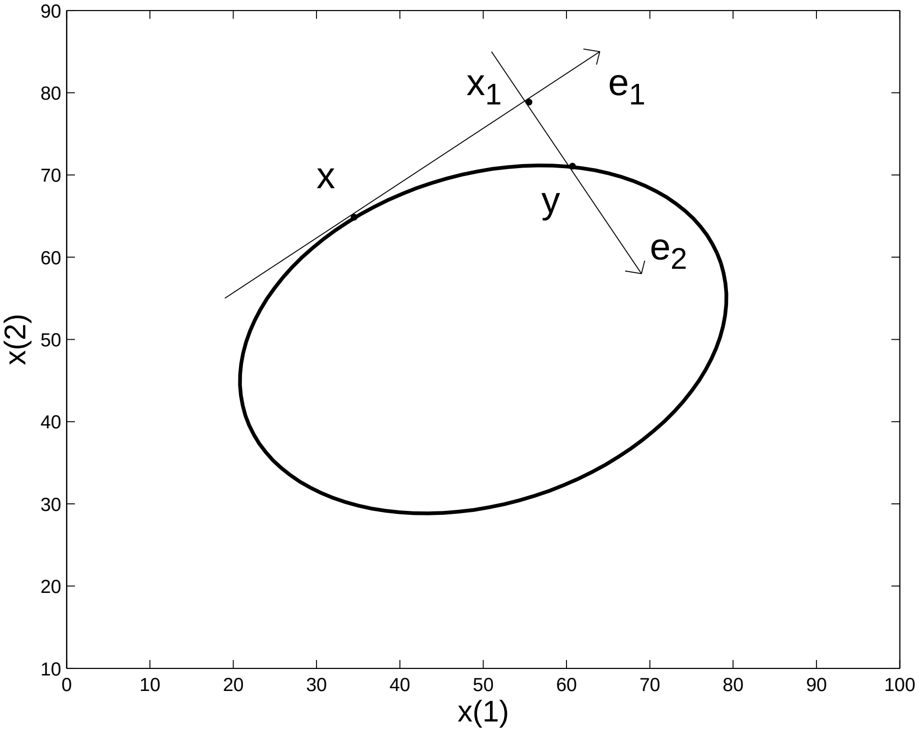
In the limit as the support of π decreases, we reach the special case where π is supported on a manifold, a surface within $\Re^n$ of dimension less than n.



The preceding method converges to a limiting form —

or view this directly as a M-H sampler on the manifold.

Distributions on manifolds



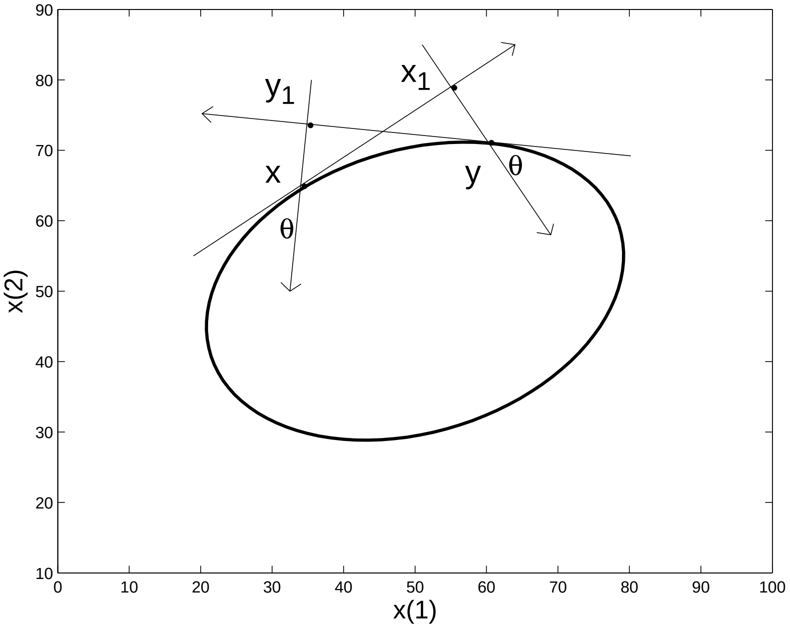
The second step is to “find the manifold” in direction e_2.

If there is more than one intersection, the only concern is whether the return move finds the original x.

Higher dimensions:

If the manifold is an m-dimensional surface in $\Re^n$, then e_1 is the m-dimensional tangent plane at x and e_2 is the $(n-m)$-dimensional space $\perp e_1$.

Distributions on manifolds



Acceptance probability:

Denote the density of x_1 given x by $h_1(x, x_1)$.

The position of y is determined by x_1.

This transformation has a scale change of $\sin(\theta)$, so the density on the manifold is $h(x, y) = h_1(x, x_1)\ \sin(\theta)$.

Similarly $h(y, x) = h_1(y, y_1)\ \sin(\theta)$.

Thus, the acceptance probability for the move x to y is

$$\alpha(x, y) = \min\{1, \frac{\pi(y)\ \ h_1(y, y_1)}{\pi(x)\ \ h_1(x, x_1)}\}.$$

9. Conclusions

It *is* possible to create a M-H algorithm where an initial proposal can be updated to improve its likelihood under π before the accept/reject decision is taken.

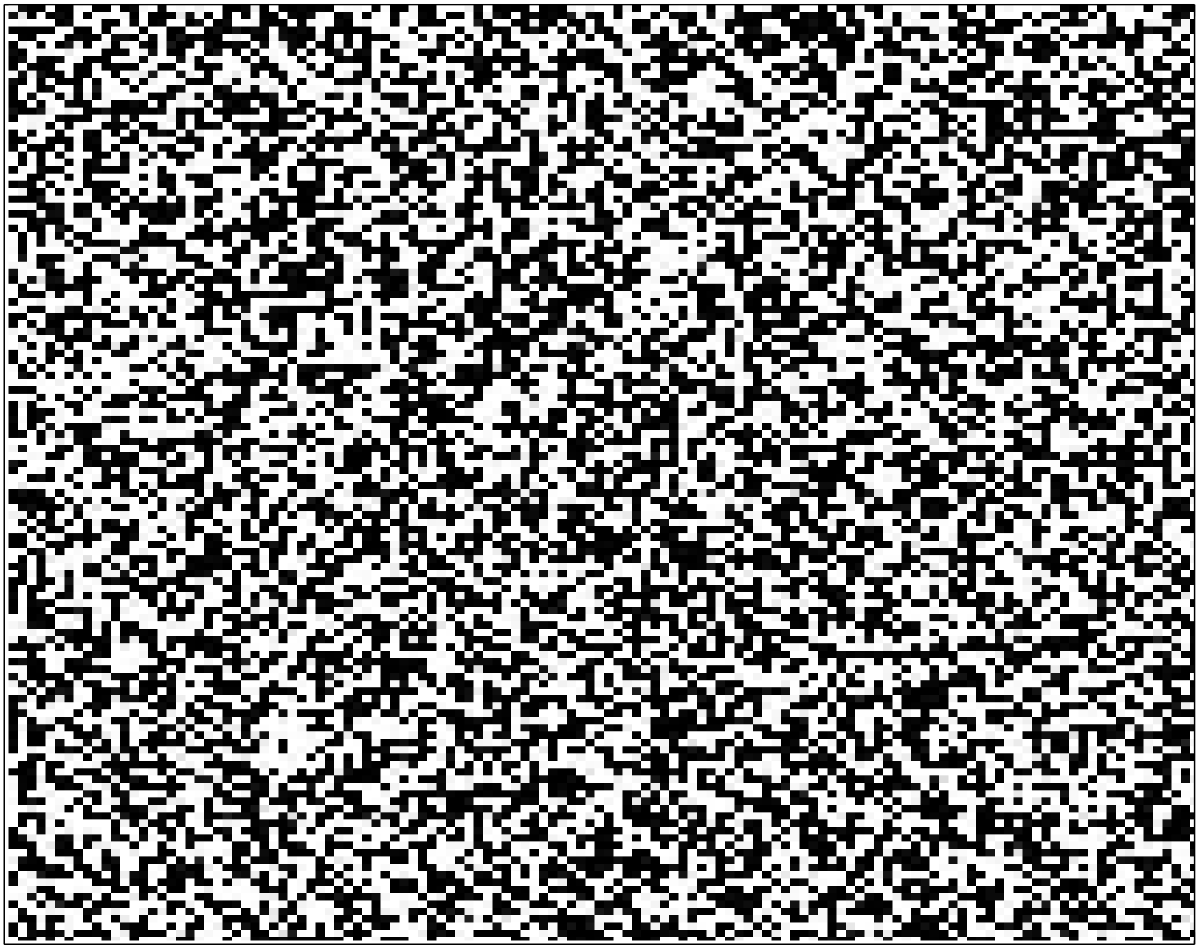
An element of our proposed method can overcome difficulties in providing M-H steps within pathological distributions, either with very thin support or confined to a manifold.

Applications to explore further:

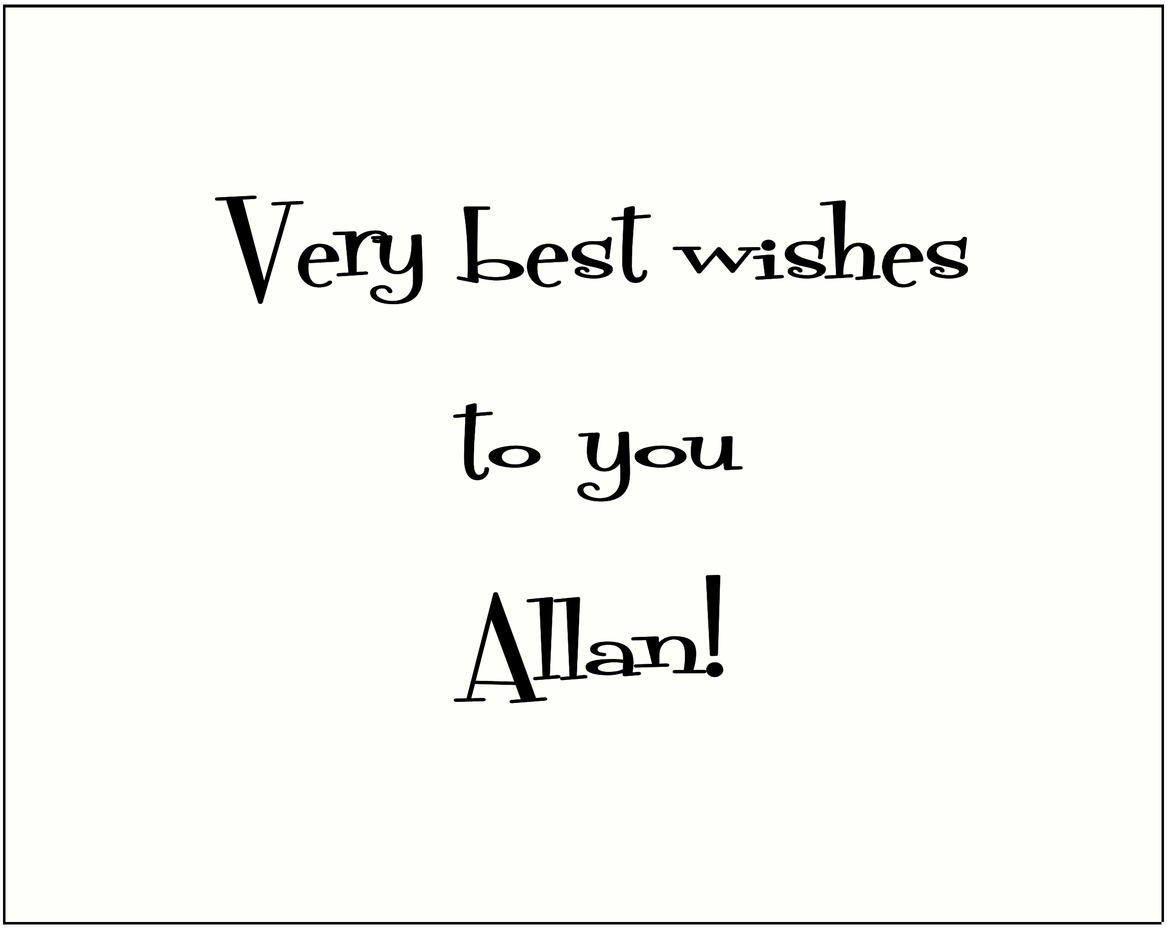
jumps between dimensions in a variable dimension state space,

many other interesting problems . . .

Postscript: An Image Example



An Image Example



Very best wishes

to you

Allan!