

# A MODIFICATION OF NEAL'S ALGORITHM FOR A CONTINUOUS STATE SPACE AND AN APPLICATION TO THE FOKKER-PLANCK EQUATION.

KYLE MARSHALL

SUPERVISOR: DR. KEVIN LIN

ABSTRACT. The Metropolis-Hastings algorithm generates correlated samples from a target distribution by constructing a Markov chain which has as its stationary distribution the desired target distribution. One property of this algorithm is that it creates reversible Markov chains. As a result, reversible chains are often used in Monte Carlo simulations. Reversible Markov chains also have the added benefit of being easier to deal with analytically. However, Neal proposes that one should not restrict to only dealing with reversible Markov chains, by proposing an algorithm for which reversible Markov chains can be made into non-reversible Markov chains in a process which will not increase the asymptotic variance. The non-reversible chains work by avoiding the "backtracking" that causes Markov chains to remain in one position for too long. In the body of this paper, we extend Neal's algorithm from a discrete to a continuous state space and then test the algorithm on an application to the Fokker-Planck equation.

# Introduction.

For simplicity, we will first suppose  $\Omega$  is a finite (or countable) state space. A Markov chain is a sequence of random variables  $X_1, X_2, \dots$  defined on  $\Omega$  for which the future states depend only on the present state. This condition, called the Markov property, can be expressed symbolically in the form

$$P(X_{t+1} = x_{t+1} | X_1 = x_1, X_2 = x_2, \dots, X_t = x_t) = P(X_{t+1} = x_{t+1} | X_t = x_t)$$

and we will use the following notation for the transition probability

$$P(X_{t+1} = y | X_t = x) = T(x, y).$$

To simplify matters further, any Markov chain we discuss will be irreducible, meaning that for any two elements  $i, j \in \Omega$ ,  $P(X_t = i | X_s = j) > 0$  for some  $t, s$ . We will deal with only time-homogenous Markov chains, which is to say that the probability distribution of the Markov chain is fixed, and independent of time. In such a case, every Markov chain has a convenient matrix representation, where the entry in the  $i^{\text{th}}$  row and  $j^{\text{th}}$  column is the transition probability  $T(i, j)$ . A vector  $\pi$  is called a stationary (or invariant) distribution if  $\pi \geq 0$ , its entries sum to 1 and

$$\pi(j) = \sum_{i \in \Omega} \pi(i)T(i, j).$$

A common practice is to use reversible Markov chains, which are Markov chains which satisfy the condition called "detailed balance", which says that for each  $i, j \in \Omega$ ,  $\pi(i)T(i, j) = \pi(j)T(j, i)$ . Since

$$\sum_{i \in \Omega} \pi(i)T(i, j) = \sum_{i \in \Omega} \pi(j)T(j, i) = \pi(j) \sum_{i \in \Omega} \pi(j) = \pi(j)$$

we see that if a Markov chain is reversible with respect to  $\pi$ , then  $\pi$  is automatically a stationary distribution. Furthermore, it is the unique stationary distribution.

Markov chain Monte Carlo (MCMC) is a technique which uses Markov chains to draw random samples from a given target distribution by using a Markov chain which has the target distribution as its stationary distribution. MCMC methods arise naturally as a method for estimating the expected value of an observable with respect to a probability distribution. For instance, if  $\pi$  is the stationary distribution of a given Markov chain, and we can obtain a large number of random samples  $X_1, X_2, \dots, X_N$ , then the ergodic theorem tells us that the averages converge to the expectation of  $\pi$ . This technique can be modified to computer integrals in the following way. Suppose that we have a complicated function  $f$  which we wish to integrate over some region,  $\Omega$ . If we can decompose  $f$  into the product of another function  $\varphi$  and a probability density function  $\rho$ , then

$$\int_{\Omega} f = \int_{\Omega} \varphi \cdot \rho = E_{\rho}[\varphi].$$

If we can obtain samples from  $\rho$ , say  $x_1, x_2, \dots, x_N$ , then

$$\int_{\Omega} f \sim \frac{1}{N} \sum_{i=1}^N \varphi(x_i).$$

EXAMPLE 1. In the case where the state space is discrete, integration is replaced by summation. Suppose that we wish to find the value of

$$\sum_{n=1}^{\infty} \frac{1}{n^2} = \frac{\pi^2}{6} \approx 1.6449.$$

One way in which we can do this is by writing

$$\frac{1}{n^2} = \frac{2^n}{n^2} \cdot \frac{1}{2^n}.$$

The sum of the reciprocals of the powers of 2 forms a probability distribution  $\rho$  over our state space  $\{1, 2, 3, \dots\}$  and if we let  $\varphi(n) = \frac{2^n}{n^2}$ , then we have

$$\sum_{n=1}^{\infty} \frac{1}{n^2} \approx \frac{1}{N} \sum_{i=1}^N \varphi(x_i)$$

where  $x_i$  are distributed according to  $\rho$  and  $N$  is the number of samples. Taking  $N = 50000$ , this method yielded

$$\frac{1}{N} \sum_{i=1}^N \varphi(x_i) = 1.6409.$$

As expected, the result is close to the real value, with the error in the range of  $1/\sqrt{50000} \approx .0045$ . ■

The above example is an example of Monte Carlo integration, but has not yet incorporated the use of Markov chains. In our example, the distribution  $\rho$  was easy to obtain independent samples from. Often, the distribution we are interested in is not so simple, and drawing independent samples can be somewhat difficult. With MCMC, we liberate ourselves from having to draw independent samples by drawing a large number of correlated samples which approximate the target distribution. Since the samples we draw from MCMC will not be independent, we can no longer apply the law of large numbers to to prove convergence of the expectation. However, there is an ergodic theorem which guarantees the convergence of the correlated samples to the expected value of the stationary distribution.

Detailed balance for a continuous state space is defined analogously to the discrete case. On a continuous state space, transition probabilities are replaced by a transition kernel  $P(x, A)$ , which is a conditional distribution function that represents the probability of moving from  $x$  into the set  $A$ . We can express any transition kernel in the form

$$P(x, dy) = p(x, y)dy + r(x)\delta_x(dy)$$

where  $p(x, x) = 0$ ,  $\delta_x(dy) = 1$  if  $x \in dy$  and 0 otherwise, and  $r(x) = 1 - \int_{y \neq x} p(x, y)dy$  is the probability that the chain remains at  $x$ . The form above can be thought of as a decomposition of the transition kernel into its continuous and discrete components. Now, if

we suppose that  $p(x, y)$  satisfies detailed balance, then

$$\begin{aligned}
\int P(x, A)\pi(x)dx &= \int \left[ \int_A p(x, y)dy \right] \pi(x)dx + \int r(x)\delta_x(A)\pi(x)dx \\
&= \int_A \left[ \int p(x, y)\pi(x)dx \right] dy + \int_A r(x)\pi(x)dx \\
&= \int_A \left[ \int p(y, x)\pi(y)dx \right] dy + \int_A r(x)\pi(x)dx \\
&= \int_A (1 - r(y))\pi(y)dy + \int_A r(x)\pi(x)dx \\
&= \int_A \pi(y)dy
\end{aligned}$$

and so  $\pi$  is the stationary distribution. For the purposes of this paper,  $r(x) = 0$

The Metropolis Hastings (M-H) algorithm is a way of creating a Markov chain which has  $\pi$  as the desired target distribution, without sampling from  $\pi$  directly [1]. Suppose that we have a candidate-generating function  $q(x, y)$  (which plays the role of  $p(x, y)$  above). Ideally, this density will satisfy the detailed balance condition above, in which case we will know that  $\pi$  is the stationary distribution. However, this is not often the situation, in which case we may have

$$\pi(x)q(x, y) > \pi(y)q(y, x) \text{ for some } x, y.$$

If we want to balance this, we impose a function  $\alpha(x, y)$  which will act as the probability that the move is made. In other words, transitions from  $x$  to  $y$  occur too frequently, and so  $\alpha(x, y)$  acts to balance the equation by occasionally rejecting such moves. We have

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

and so  $\alpha(x, y) = \pi(y)q(y, x)/\pi(x)q(x, y)$ . Since  $\alpha$  is a probability,  $\alpha \leq 1$  and so if we define

$$\alpha(x, y) = \begin{cases} \min \left[ \frac{\pi(y)q(y, x)}{\pi(x)q(x, y)}, 1 \right] & \text{if } \pi(x)q(x, y) > 0 \\ 1 & \text{otherwise} \end{cases}$$

then the Markov chain satisfies detailed balance and thus leaves  $\pi$  as the stationary distribution. To apply M-H, first a proposal distribution,  $q$ , must be chosen. Then, a Markov chain is run according to the following rules:

```

Draw  $x_0$  from initial distribution  $\mu$ 
While  $t < N$ 
  Let  $x' \sim q(x_t, \cdot)$ 
  Let  $\alpha \sim U(0, 1)$ 
  If  $\alpha < \min[\pi(y)q(y, x)/\pi(x)q(x, y), 1]$ 
     $x_{t+1} = x'$ , increment  $t$  by one
  Else
     $x_{t+1} = x_t$ , increment  $t$  by one
end

```

EXAMPLE 2. As an illustration of the Metropolis-Hastings algorithm, consider a Markov chain on a finite state space  $\{1, 2, 4, \dots, 10\}$  with proposal distributions given by the matrix

$$Q = \begin{bmatrix} .3 & .2 & .15 & .05 & .05 & .05 & .05 & .05 & .05 & .05 \\ .05 & .3 & .2 & .15 & .05 & .05 & .05 & .05 & .05 & .05 \\ .05 & .05 & .3 & .2 & .15 & .05 & .05 & .05 & .05 & .05 \\ .05 & .05 & .05 & .3 & .2 & .15 & .05 & .05 & .05 & .05 \\ .05 & .05 & .05 & .05 & .3 & .2 & .15 & .05 & .05 & .05 \\ .05 & .05 & .05 & .05 & .05 & .3 & .2 & .15 & .05 & .05 \\ .05 & .05 & .05 & .05 & .05 & .05 & .3 & .2 & .15 & .05 \\ .05 & .05 & .05 & .05 & .05 & .05 & .05 & .3 & .2 & .15 \\ .15 & .05 & .05 & .05 & .05 & .05 & .05 & .05 & .3 & .2 \\ .2 & .15 & .05 & .05 & .05 & .05 & .05 & .05 & .05 & .3 \end{bmatrix}$$

where the transition probability of moving to  $j$  given that the chain is currently at the state  $i$  is given by  $Q_{ij}$ . Our target distribution will be

$$P = [.01 \ .02 \ .04 \ .08 \ .35 \ .35 \ .08 \ .04 \ .02 \ .01].$$

then after running the Markov chain described in the Metropolis-Hastings algorithm for 10000 steps, we obtain the following distribution, which is compared to the distribution of  $P$  in figure 1. After only 10000 steps, the distribution of the correlated samples is close to the desired distribution.

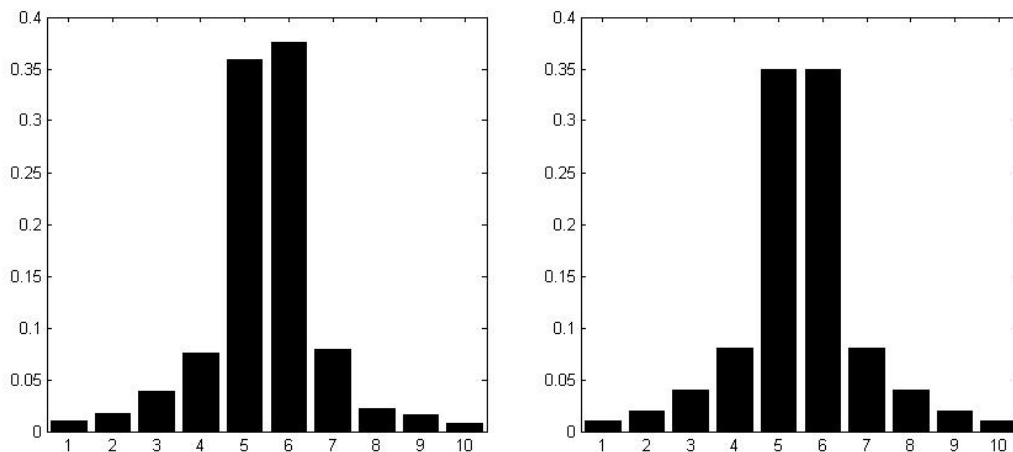


FIGURE 1. Distribution obtained using Metropolis-Hastings vs. Target Distribution

■

The idea behind M-H is that we can sample from a complicated density by sampling from densities from which samples can be easily drawn. An example of the use of M-H in the case of a continuous state space will be given in next section.

## Fokker-Planck Equation.

Suppose we have a system of particles represented on the real line  $\mathbb{R}$ , where the density of the particles at time  $t$  is given by  $p(x, t)$ . Then for an interval  $(a, b) \subset \mathbb{R}$ , we can find the rate of change of the particles within that interval,

$$\frac{d}{dt} \int_a^b p(x, t) dx = \int_a^b \frac{\partial}{\partial t} p(x, t) dx.$$

If we denote by  $j(x, t)$  the current through the point  $x$  at time  $t$ , then we can also express the change in the quantity of particles in  $(a, b)$  by  $-j(b, t) + j(a, t)$ . Averaging over the interval,

we have

$$\frac{1}{b-a} \int_a^b \frac{\partial}{\partial t} p(x, t) dx = -\frac{j(b, t) - j(a, t)}{b-a}$$

and taking  $b \rightarrow a$ , we obtain the continuity equation

$$\frac{\partial p}{\partial t} = -\frac{\partial j}{\partial x}.$$

More generally, if the particles in the system behave according to the Ornstein-Uhlenbeck stochastic differential equation [4],

$$dx_t = -v'(x_t)dt + \varepsilon dW_t,$$

where  $v$  is a differentiable function,  $\varepsilon$  is a constant and  $W_t$  is a Wiener process [2], then we can show that

$$j(x, t) = -v'(x)p(x, t) - \frac{\varepsilon^2}{2} \frac{\partial}{\partial x} p(x, t).$$

Taking the derivative with respect to  $x$  and using the continuity equation above, we obtain the Fokker-Planck (F-P) equation

$$\frac{\partial p}{\partial t} = \underbrace{\frac{\partial}{\partial x}(v' \cdot p)}_{\text{drift}} + \underbrace{\frac{\varepsilon^2}{2} \frac{\partial^2 p}{\partial x^2}}_{\text{diffusion}}.$$

The stationary distribution can be found by setting  $\frac{\partial p}{\partial t} = 0$  and solving for  $p$ . The general solution of the F-P equation is given by

$$p(x) = C e^{-v(x) \frac{2}{\varepsilon^2}}.$$

and since we want  $p$  to be a probability density, we require  $C = \left( \int e^{-v(x) \frac{2}{\varepsilon^2}} \right)^{-1}$ .

The discrete-time version of the Ornstein-Uhlenbeck SDE takes the form

$$x_{t+\Delta t} = x_t - v'(x_t)\Delta t + \varepsilon\sqrt{\Delta t}N_t$$

where  $N_t \sim N(0, 1)$ . We can therefore think of this process as a Markov chain, where transitions are made according to  $x_{t+1} \sim N(x_t - v'(x_t)\Delta t, \epsilon^2\Delta t)$ . Although sampling from the discrete time version will not give the exact stationary solution to the continuous time F-P equation, it will be fairly close, with the accuracy increasing as  $\Delta t \rightarrow 0$ .

The interpretation of the F-P equation is that there are particles in a potential  $v(x)$  which move toward a minimum energy configuration according to

$$x'(t) = -v'(x(t)),$$

while simultaneously being affected by "white noise", random disturbances in the position of the particle. The diffusion term is also known as the heat equation, and has solution

$$f(x, t) = \frac{1}{(2\pi Dt)^{1/2}} e^{-\frac{x^2}{2Dt}}$$

under the initial condition  $f(x, 0) = \delta_0$ , where  $\delta_0$  is the unit mass at  $(x, 0)$ . The solution is recognizable as the  $N(0, Dt)$ , a normally distributed random variable with mean 0 and variance  $Dt$ .

EXAMPLE 3. If we set  $v(x) = x^2$ ,  $\Delta t = 1/100$ , and  $\epsilon = 2$ , then we would expect M-H to give us a stationary distribution which would closely resemble the standard normal distribution. Our proposal distribution will be the  $N(x_t - v'(x_t), \epsilon^2\Delta t)$  distribution, which is the transition kernel associated with the discretized version of the Ornstein-Uhlenbeck SDE in the discussion above. Running a Markov chain for 1000000 steps yielded the following:

It is worth noting that the M-H method only rejected .008% of the proposals (80 out of 1 million). Therefore, the Markov chain without M-H would have also been close to the target distribution. If we make  $v(x) = (x - 1)^2(x + 1)^2$ , then the rejection rate increases to 8291 or .8%, and we compare the results of direct sampling against M-H update sampling in figure 3.

■

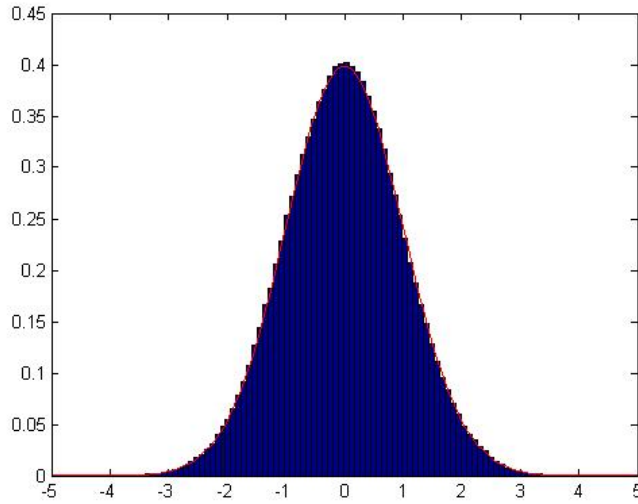


FIGURE 2. Normal distribution and distribution obtained from the F-P equation with M-H.

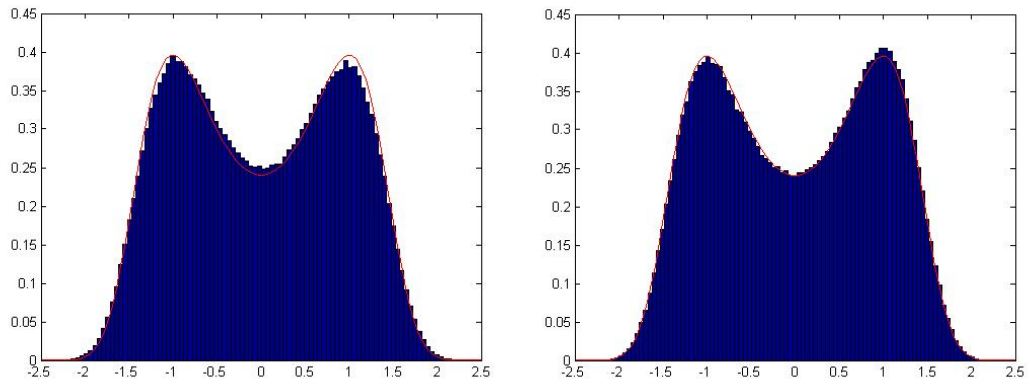


FIGURE 3. Distribution obtained by direct sampling vs. M-H update sampling

The M-H method is not without its limitations. Recall that the acceptance rate  $\alpha$  is governed by the formula

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

where  $y$  is generated according to the proposal distribution  $q(x, \cdot)$ . If we have a well shaped stationary distribution (a well in the stationary distribution corresponds to a peak in the potential function  $v(x)$ ), as in figure 4, and a gaussian shaped proposal distribution, then the chain may have trouble traveling between the higher density areas. We could always use a different proposal distribution, but recall that in the discussion of the F-P equation above,

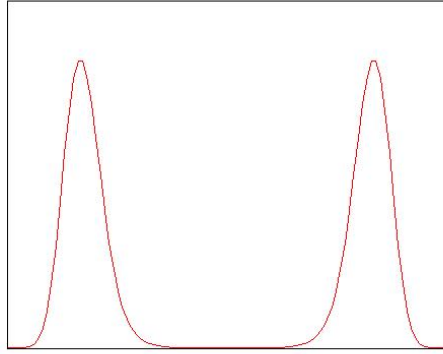


FIGURE 4. Example of a distribution with a well

the discrete-time stochastic differential equation gives rise to a Markov chain governed by  $x_{t+1} = N(x_t - v'(x_t)\Delta t, \epsilon^2\Delta t)$  so it is natural to choose a gaussian in this case. Under these circumstances, we will test Neal's algorithm and determine whether the non-reversible chain it makes is more efficient than the reversible one which arises from a straightforward application of the M-H algorithm.

## Neal's Algorithm.

Neal proposes an algorithm with the intention of eliminating the "backtracking" of the Markov chain like we see in a situation with a well. The algorithm takes what normally would be a reversible Markov chain and makes it non-reversible while preserving the stationary distribution. He proves [3] that his algorithm at the very least does not increase the asymptotic variance, and shows that in many cases it actually decreases the asymptotic variance. This indicates that the algorithm may be more efficient or more capable of dealing with situations such as the well problem more effectively than a standard application of M-H.

The first step is a technique called *expanding the chain*. If we have a reversible Markov chain defined on  $\Omega$  with transition probabilities  $T(x, y)$ , then we can generate a non-reversible chain which avoids backtracking in the following way. First, we consider a Markov chain

on a larger state space,  $\Omega' = \{(x, y) \mid T(x, y) > 0\}$ . A transition in this expanded chain is determined by applying the following two steps in order.

(1) Swap the components of state:  $(x, y) \rightarrow (y, x)$

(2) Replace the second component with a new value,  $y^*$ , sampled from its conditional distribution (under  $\pi'$ ) given the value of the first component.

Computing the stationary distribution of the expanded chain, we have

$$\begin{aligned}
 \pi'((x, y)) &= \int_{\Omega \times \Omega} \pi'((w, z))T'((w, z), (x, y))dzdw \\
 &= \int_{\Omega} \pi'((w, x))T'((w, x), (x, y))dw \\
 &= \int_{\Omega} \pi'((w, x))T(x, y)dw \\
 &= T(x, y) \int_{\Omega} \pi'((w, x))dw \\
 &= T(x, y) \int_{\Omega} \pi(w)T(w, x)dw \\
 &= T(x, y)\pi(x) = T(y, x)\pi(y)
 \end{aligned}$$

where the last equality comes from the reversibility of the original chain. Now if we take the marginal distributions of either the first or second component we will recover the stationary distribution of the original chain. Thus, the idea is to create a sequence of steps  $(x_1, x_2), (x_2, x_3), (x_3, x_4) \dots$  and then retrieve the desired distribution  $\pi$  by considering only the first or second component. The intuitive interpretation of this expanded state space is that instead of viewing the elements of  $\Omega$  as the states, we view the "arrows" between the elements of  $\Omega$  as the states. The second step is to replace step (2) above with a sample from a distribution which decreases the probability of staying in the same state, but leaves  $\pi'$  invariant. Neal's implementation of the algorithm deals with the the case when the state space  $\Omega$  is discrete. In this case, he uses a form of Gibb's sampling due to Liu [3] which proposes  $y^*$  with probability  $\pi(y^* \mid x)/(1 - \pi(y \mid x))$  if  $y \neq y^*$  and sets the proposal probability of the current state  $y$  to zero.

To incorporate this algorithm into our discussion of the F-P equation, we need to adapt his method to the case when we have a continuous state space. The choice of update in step (2) above will be the following. If  $x_{t-1} < x_t$ , then we propose  $x' \sim q(x_t, \cdot \mid x' > x_t)$  and if  $x_{t-1} > x_t$  then we propose  $x' \sim q(x_t, \cdot \mid x' < x_t)$ . We then use M-H in order to maintain the invariant distribution  $\pi'$ . While this will make the Markov chain on the expanded state space reversible, when we restrict to one of the components of state the Markov chain we obtain will not be reversible except in degenerate cases. Recall, that  $v(x)$  is the potential function and let  $\eta(x, y)$  be the transition probability of going from  $x$  to  $y$  of a gaussian with mean  $x$  and variance  $\epsilon^2 \Delta t$ . If the current state is  $(x, y)$ , with  $x < y$  and  $z$  is proposed according to our updated distribution, then the M-H acceptance ratio is given by

$$\begin{aligned} \alpha &= \frac{\pi'((y, z))q((y, z), (y, x))}{\pi'((y, x))q((y, x), (y, z))} \\ &= \frac{\pi(y)\eta(y - v'(y)\Delta t, z) \left( \frac{\eta(y - v'(y)\Delta t, x)}{\int_y^\infty \eta(y - v'(y)\Delta t, s)ds} \right)}{\pi(y)\eta(y - v'(y)\Delta t, x) \left( \frac{\eta(y - v'(y)\Delta t, z)}{\int_{-\infty}^y \eta(y - v'(y)\Delta t, s)ds} \right)} \\ &= \frac{\int_{-\infty}^y \eta(y - v'(y)\Delta t, s)ds}{\int_y^\infty \eta(y - v'(y)\Delta t, s)ds}. \end{aligned}$$

If  $x > y$ , then the ratio just becomes the reciprocal, and so the acceptance rate is given by taking the minimum of this ratio and the number 1.

Implementing this algorithm is fairly straightforward. The choices  $v(x) = (x - 2)^2(x + 2)^2$  and  $\epsilon = 2$  give the graph in figure 4, and we chose  $\Delta t = 1/100$  for the length of the time step. To determine the relative efficiency, we ran a test with the goal to determine which process was best at clearing the well. We started the chain at  $x = -2$  (the local maximum of the left hump) and counted the number of steps it took for the chain to be greater or equal to 2 (the local maximum of the right hump). In the table below,  $t$  is the total number of steps, including rejections.

TABLE 1. Non-Reversible Markov chain Test Data

t	71094	12819	70277	20403	5138	5888	4474	12599	11655	71096	1156
reject	18110	3257	17904	5192	1266	1487	1129	3245	2916	18112	275
reject %	25.47	25.41	25.48	25.45	24.64	25.25	25.23	25.76	25.05	25.48	23.79

TABLE 2. Reversible Markov Chain Test Data

t	170127	175177	226743	36243	80934	686868	28970	273429	263488	170127	175177
reject	7037	7291	9330	1470	3313	28115	1173	11400	11035	7037	7291
reject %	4.14	4.16	4.11	4.06	4.09	4.09	4.05	4.17	4.19	4.14	4.16

The preliminary tests indicate that the non-reversible chain is faster when it comes to clearing the well, however the rejection rate is also much higher. Earlier, we had used the M-H algorithm to approximate the distribution of a normal random variable with mean 0 and variance 1. When we used the non-reversible algorithm to make the same approximation, we obtained the figure below, but the rejection rate was significantly higher, at 6%.

## Conclusion.

Using Neal’s algorithm, we constructed a non-reversible Markov chain which is in some ways superior to the reversible Markov chain which is generated using the M-H algorithm. Since the rejection rate is much higher in the non-reversible case, it is still questionable

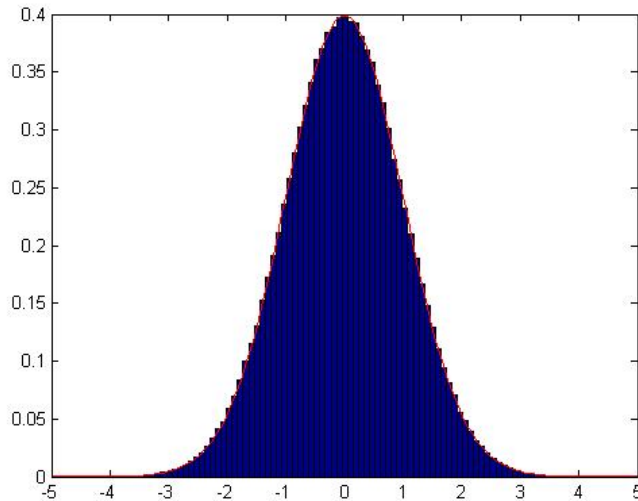


FIGURE 5. Normal distribution approximation obtained from using the non-reversible chain.

whether this method is actually more efficient for most MCMC scenarios. When the distribution is known to have a well, such as in the case we examined, the non-reversible algorithm of Neal causes the Markov chain to avoid remaining in islands of high probability and so may be better suited for this purpose. This algorithm may also prove useful when the shape of the distribution is not known. In the future, I hope to elaborate on this project and include more interesting examples of non-reversible chains and determine convergence properties of the non-reversible chain. I would like to thank Dr. Kevin Lin for proposing this topic and volunteering his time toward guiding me through each and every stage. I would also like to thank Dr. Robert Indik for approving this project, and the NSF for funding me through a VIGRE grant awarded to the University of Arizona Mathematics Department.

## REFERENCES

- [1] Chib, S., Greenberg, E., "Understanding the Metropolis-Hastings Algorithm", *The American Statistician*, Vol. 49 (1995), pp. 327-335
- [2] Evans, L.C., "An Introduction to Stochastic Differential Equations", <http://math.berkeley.edu/~evans/SDE.course.pdf>
- [3] Neal, R.M., "Improving asymptotic variance of MCMC estimators: Non-reversible chains are better", University of Toronto Computer Science Department Technical Report No. 0406 (2004)
- [4] Gareth O. Roberts and Richard L. Tweedie, "Exponential Convergence of Langevin Distributions and Their Discrete Approximations," *Bernoulli*, Vol. 2 (1996), pp. 341-363