Ensemble Samplers with Affine Invariance

Jonathan Goodman [*] Jonathan Weare[*]

September 30, 2009

Abstract

We propose a family of Markov chain Monte Carlo (MCMC) methods whose performance is unaffected by affine tranformations of space. These algorithms are easy to construct and require little or no additional computational overhead. They should be particulary useful for sampling badly scaled distributions. Computational tests show that the affine invariant methods can be significantly faster than standard MCMC methods on highly skewed distributions.

1 Introduction

Markov chain Monte Carlo (MCMC) sampling methods typically have parameters that need to be adjusted for a specific problem of interest [9] [10] [1]. For

[*] Courant Institute, NYU, 251 Mercer St, New York, 10012

example, a trial step size that works well for a probability density $\pi(x)$, with $x \in \mathbb{R}^n$, may work poorly for the scaled density

$$\pi_\lambda(x) \;=\; \lambda^{-n}\,\pi\left(\lambda x\right) \;, \tag{1}$$

if λ is very large or very small. Christen [2] has recently suggested a simple method whose performance sampling the density π_λ is independent of the value of λ. Inspired by this idea we suggest a family of many particle (ensemble) MCMC samplers with the more general *affine invariance* property. Affine invariance implies that the performance of our method is independent of the aspect ratio in highly anisotropic distributions such as the one depicted in Figure 1.

An affine transformation is an invertible mapping from $\mathbb{R}^n$ to $\mathbb{R}^n$ of the form $y = Ax + b$. If X has probability density $\pi(x)$, then $Y = AX + b$ has density

$$\pi_{A,b}(y) \;=\; \pi_{A,b}(Ax+b) \;\propto\; \pi(x)\;. \tag{2}$$

Consider, for example, the skewed probability density on $\mathbb{R}^2$ pictured in Figure 1:

$$\pi(x) \;\propto\; \exp\left(\frac{-(x_1-x_2)^2}{2\epsilon} \;-\; \frac{(x_1+x_2)^2}{2}\right)\;. \tag{3}$$

Single variable MCMC strategies such as Metropolis or heat bath (Gibbs sampler) [13][10] would be forced to make perturbations of order $\sqrt{\epsilon}$ and would have slow equilibration. A better MCMC sampler for π would use perturbations of order $\sqrt{\epsilon}$ in the $(1,-1)$ direction and perturbations of order one in the $(1,1)$ direction. The $\mathbb{R}^2 \to \mathbb{R}^2$ affine transformation

$$y_1 = \frac{x_1 - x_2}{\sqrt{\epsilon}} , \qquad y_2 = x_1 + x_2$$

turns this challenging sampling problem into the easier problem:

$$\pi_A(y) \ \propto \ e^{-\left(y_1^2 + y_2^2\right)/2} \ . \tag{4}$$

Sampling the well scaled transformed density (4) does not require detailed customization. An affine invariant sampler views the two densities as equally difficult. In particular, the performance of an affine invariant scheme on the skewed density (3) will be independent of ϵ. More generally, if an affine invariant sampler is applied to a non-degenerate multivariate normal $\pi(x) \propto e^{-x^tHx/2}$, the performance is independent of H.

We call an MCMC algorithm *affine invariant* if it has the following property. Suppose that starting point $X(1)$ and initial random number generator seed $\xi(1)$ produces the sequence $X(t)$ $(t = 1, 2, \ldots)$ if the probability density is $\pi(x)$. Now instead apply the MCMC algorithm with the same seed to probability density $\pi_{A,b}(y)$ given by (2) with starting point $Y(1) = AX(1) + b$. The algorithm is affine invariant if the resulting $Y(t)$ satisfy $Y(t) = AX(t) + b$. We are not aware of a practical affine invariant sampler of this form.

In this paper we propose a family of affine invariant *ensemble* samplers.

An ensemble, $\vec{X}$, consists of L *walkers*[1] $X_k \in \mathbb{R}^n$. Since each walker is in $\mathbb{R}^n$, we may think of the ensemble $\vec{X} = (X_1, \ldots, X_L)$ as being in $\mathbb{R}^{nL}$. The target probability density for the ensemble is the one in which the individual walkers are independent and drawn from π, i.e.

$$\Pi(\vec{x}) \; = \; \Pi(x_1, \ldots, x_L) \; = \; \pi(x_1)\,\pi(x_2)\cdots\pi(x_L) \; . \tag{5}$$

An ensemble MCMC algorithm is a Markov chain on the state space of ensembles. Starting with $\vec{X}(1)$, it produces a sequence $\vec{X}(t)$. The ensemble Markov chain can preserve the product density (5) without the individual walker sequences $X_k(t)$ (as functions of t) being independent, or even being Markov. The distribution of $X_k(t+1)$ can and will depend on $X_j(t)$ for $j \neq k$.

We apply an affine transformation to an ensemble by applying it separately to each walker:

$$\vec{X} \; = \; (X_1, \ldots, X_L) \; \xrightarrow{A,b} \; (AX_1 + b, \ldots, AX_L + b) \; = \; (Y_1, \ldots, Y_L) \; = \; \vec{Y} \; . \tag{6}$$

Suppose that $\vec{X}(1) \xrightarrow{A,b} \vec{Y}(1)$ and that $\vec{Y}(t)$ is the sequence produced using $\pi_{A,b}$ in place of π in (5) and the same initial random number generator seed. The ensemble MCMC method is affine invariant if $\vec{X}(t) \xrightarrow{A,b} \vec{Y}(t)$. We will describe the details of the algorithms in Section 2.

Our ensemble methods are motivated in part by the Nelder Mead [11]

[1] Here x_k is walker k in an ensemble of L walkers. This is inconsistent with (3) and (4), where x_1 was the first component of $x \in \mathbb{R}^2$.

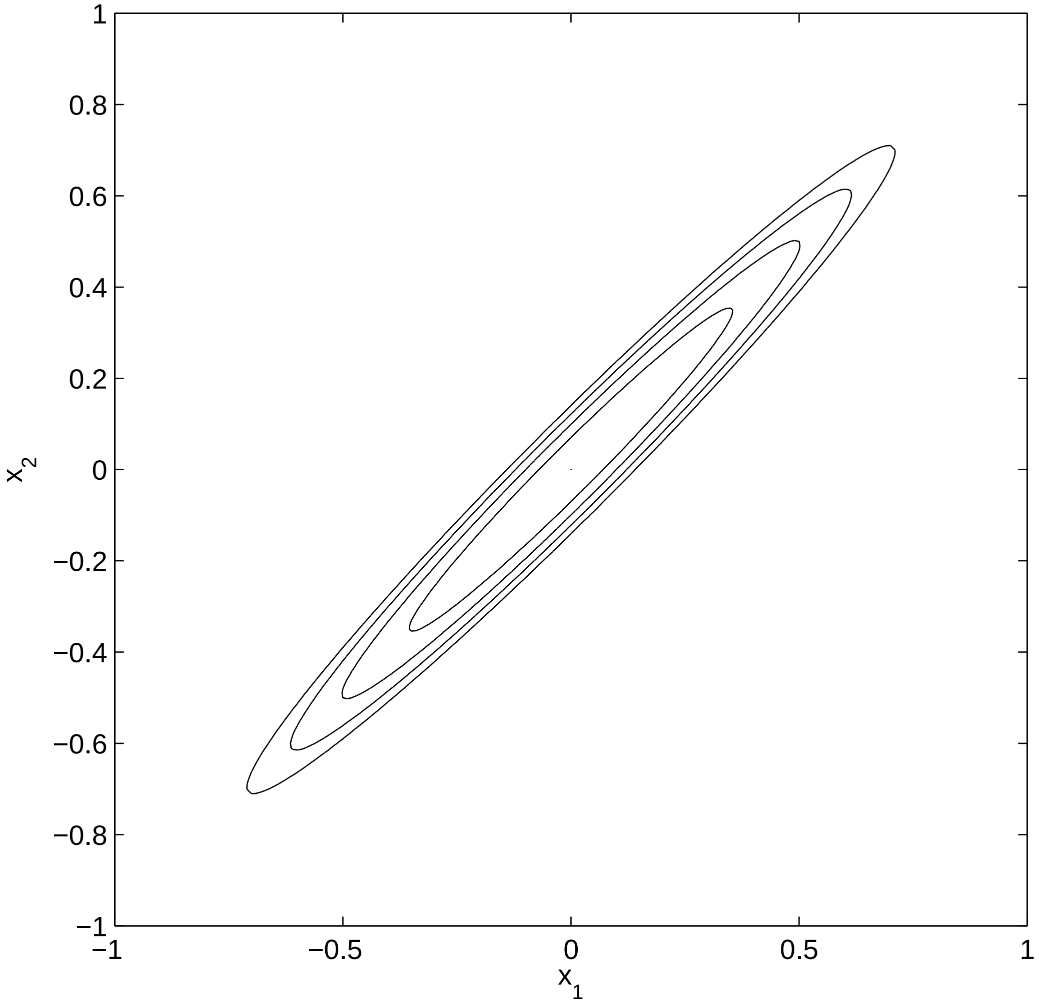

Figure 1: Contours of the Gaussian density defined in expression (3).

simplex algorithm for solving deterministic optimization problems. Many in the optimization community attribute its robust convergence to the fact that it is affine invariant. Applying the Nelder Mead algorithm to the ill conditioned optimization problem for the function (3) in Figure 1 is exactly equivalent to applying it to the easier problem of optimizing the well scaled function (4). This is not the case for non-invariant methods such as gradient descent [6].

The Nelder-Mead symplex optimization scheme evolves many copies of the system toward a local minimum (in our terminology: many walkers in an ensemble). A new position for any one copy is suggested by an affine invariant transformation which is constructed using the current positions of the other copies of the system. Similarly, our Monte Carlo method moves one walker using a proposal generated with the help of other walkers in the ensemble. The details of the construction of our ensemble MCMC schemes are given in the next section.

An additional illustration of the power of affine invariance was pointed out to us by our colleague Jeff Cheeger. Suppose we wish to sample X uniformly in a convex body, K (a bounded convex set with non-empty interior). A theorem of Fritz John (see [8]) states that there is a number r depending only on the dimension, and an affine transformation such that $\widetilde{K} = AK + b$ is well conditioned in the sense that $B_1 \subseteq \widetilde{K}$ and $\widetilde{K} \subseteq B_r$, where B_r is the ball of radius r centered at the origin. An affine invariant sampling method should, therefore, be uniformly effective over all the convex bodies of a given dimension regardless of their shape.

After a discussion of the integrated autocorrelation time as a means of comparing our ensemble methods with single particle methods in Section 3 we present the results of several numerical tests in Section 4. The first of our test distributions is a difficult 2 dimensional problem which illustrates the advantages and disadvantages of our scheme. In the second example we use our schemes to sample from a 101 dimensional approximation to the invariant measure of stochastic partial differential equation. In both cases the affine invariant methods significantly outperform the single site Metropolis scheme. Finally, in Section 5 we give a very brief discussion of the method used to compute the integrated autocorrelation times of the algorithms.

2 Construction

As mentioned in the introduction, our ensemble Markov chain is evolved by moving one walker at time. We consider one step of the ensemble Markov chain $\vec{X}(t) \to \vec{X}(t+1)$ to consist of one cycle through all L walkers in the ensemble. This is expressed in pseudo code as

```
for $k = 1, \ldots, L$
 {
   update $X_k(t) \to X_k(t+1)$
 }
```

Each walker X_k is updated using the current positions of all of the other walkers in the ensemble. The other walkers (besides X_k) form the *complementary ensemble*

$$\vec{X}_{[k]}(t) = \{X_1(t+1), \ldots, X_{k-1}(t+1), X_{k+1}(t), \ldots, X_L(t)\}\ .$$

Let $\mu(d\widetilde{x}_k, x_k \mid \vec{x}_{[k]})$ be the transition kernel for moving walker X_k. The notation means that for each $x_k \in \mathbb{R}^n$ and $\vec{x}_{[k]} \in \mathbb{R}^{(L-1)n}$, the measure $\mu(\cdot, x_k \mid \vec{x}_{[k]})$ is the probability measure for $X_k(t+1)$, if $X_k(t) = x_k$ and $\vec{X}_{[k]}(t) = \vec{x}_{[k]}$.

Our single walker moves are based on *partial resampling* (see [13] [10]). This states that the transformation $\vec{X}(t) \to \vec{X}(t+1)$ preserves the joint distribution Π if the single walker moves $X_k(t) \to X_k(t+1)$ preserve the conditional distribution of x_k given $X_{[k]}$. For our Π (which makes walkers independent), this is the same as saying that $\mu(\cdot, \cdot \mid \vec{x}_{[k]})$ preserves π for all $\vec{x}_{[k]}$, or (somewhat informally)

$$\pi(\widetilde{x}_k)d\widetilde{x}_k \ = \ \int_{\mathbb{R}^n} \mu(d\widetilde{x}_k, x_k \mid \vec{x}_{[k]})\pi(x_k)\, dx_k\ .$$

As usual, this condition is achieved using detailed balance. We use a trial distribution to propose a new value of X_k and then accept or reject this move using the appropriate Metropolis Hastings rule [13][10]. Our motivation is that the distribution of the walkers in the complementary ensemble carries useful information about the density π. This gives an automatic way to adapt the trial move to the target density. Christen [2] uses an ensemble of 2 walkers to generate scale invariant trial moves using the relative positions of the walkers.

The simplest (and best on the Rosenbrock test problem in Section 4) move of this kind that we have found is the *stretch move.* In a stretch move, we move walker X_k using one complementary walker $X_j \in \vec{X}_{[k]}(t)$ (i.e. $j \neq k$). The move consists of a proposal of the form (see Figure 2):

$$X_k(t) \;\rightarrow\; Y = X_j \,+\, Z\left(X_k(t) - X_j\right)\ . \tag{7}$$

The stretch move defined in expression (7) is similar to what is referred to as the "walk move" in [2] though the stretch move is affine invariant while the walk move of [2] is not. As pointed out in [2], if the density g of the scaling variable Z satisfies the symmetry condition

$$g\left(\frac{1}{z}\right) \;=\; z\,g(z)\ , \tag{8}$$

then the move (7) is symmetric in the sense that (in the usual informal way Metropolis is discussed)

$$\Pr\left(X_k(t) \;\rightarrow\; Y\right) \;=\; \Pr\left(Y \;\rightarrow\; X_k(t)\right).$$

The particular distribution we use is the one suggested in [2]

$$g(z) \;\propto\; \begin{cases} \frac{1}{\sqrt{z}}, & \text{if } z \in [\frac{1}{a}, a], \\ 0, & \text{otherwise.} \end{cases} \tag{9}$$

where the parameter $a > 1$ can be adjusted to improve performance.

To find the appropriate acceptance probability for this move we again appeal to partial resampling. Notice that the proposal value Y lies on the ray

$$\{y \in \mathbb{R}^n : \, y - X_j = \lambda \, (X_k(t) - X_j) \, , \; \lambda > 0\} \, .$$

The conditional density of π along this ray is proportional to

$$\|y - X_j\|^{n-1} \, \pi(y).$$

Since the proposal in (7) is symmetric, partial resampling then implies that if we accept the move $X_k(t+1) = Y$ with probability

$$\min\left\{1, \frac{\|Y - X_j\|^{n-1} \, \pi(Y)}{\|X_k(t) - X_j\|^{n-1} \, \pi(X_k(t))}\right\} \; = \; \min\left\{1, Z^{n-1} \, \frac{\pi(Y)}{\pi(X_k(t))}\right\}$$

and set $X_k(t+1) = X_k(t)$ otherwise, the resulting Markov chain satisfies detailed balance.

The stretch move, and the walk and replacement moves below, define irreducible Markov chains on the space of *general* ensembles. An ensemble is general if there is no lower dimensional hyperplane (dim $< n$) that contains all the walkers in the ensemble. The space of general ensembles is $\mathcal{G} \subset \mathbb{R}^{nL}$. For $L \geq n + 1$, a condition we always assume, almost every ensemble (with respect to Π) is general. Therefore, sampling Π restricted to $\mathcal{G}$ is (almost) the same as sampling Π on all of $\mathbb{R}^{nL}$. It is clear that if $\vec{X}(1) \in \mathcal{G}$, then almost surely $\vec{X}(t) \in \mathcal{G}$ for $t = 2, 3, \ldots$. We assume that $\vec{X}(1)$ is general. It is clear that any general ensemble can be transformed to any other general ensemble by a finite sequence of stretch moves.

The operation $\vec{X}(t) \to \vec{X}(t+1)$ using one stretch move per walker is given by:

```
for $k = 1, \ldots, L$
 {
   choose $X_j \in \vec{X}_{[k]}(t)$ at random
   generate $Y = X_j + Z(X_k(t) - X_j)$, all $Z$ choices independent
   accept, set $X_k(t+1) = Y$, with probability (7)
   otherwise reject, set $X_k(t+1) = X_k(t)$
 }
```

We offer two alternative affine invariant methods. The first, which we call the *walk* move, is illustrated in Figure 3. A walk move begins by choosing a subset S of the walkers in $\vec{X}_{[k]}(t)$. It is necessary that $|S| \geq 2$, and that the choice of S is independent of $X_k(t)$.

The center of mass of this subset is

$$\overline{X}_S \;=\; \frac{1}{|S|} \sum_{X_m \in S} X_m \; .$$

Let Z_m be independent mean zero variance σ^2 normals, and define the trial walk step by

$$W \ = \sum_{X_m \in S} Z_m \left(X_m - \overline{X}_S \right) \ . \tag{10}$$

The proposed trial move is $X_k(t) \to X_k(t) + W$. The random variable (10) is symmetric in that

$$\Pr(X \to X + W = Y) \ = \ \Pr(Y \to Y - W = X) \ .$$

Therefore, we insure detailed balance by accepting the move $X_x(t) \to X_k(t) + W$ with the Metropolis acceptance probability

$$\min \left\{ 1, \ \frac{\pi\left((X_k(t) + W\right)}{\pi\left((X_k(t)\right)} \ \right\} \ .$$

The walk move ensemble Monte Carlo method just described clearly is affine invariant in the sense discussed above. In the invariant density $\Pi(\vec{x})$ given by (5), the covariance matrix for W satisfies (an easy check)

$$\operatorname{cov}\left[W\right] \ \propto \ \operatorname{cov}_\pi[X] \ .$$

The constant of proportionality depends on σ^2 and $|S|$. If π is highly skewed in the fashion of Figure 1, then the distribution of the proposed moves will have the same skewness.

Finally, we propose a variant of the walk move called the *replacement move.* Suppose $\pi_S(x \mid S)$ is an estimate of $\pi(x)$ using the sub-ensemble $S \subset X_{[k]}(t)$.

A replacement move seeks to replace $X_k(t)$ with an independent sample from $\pi_S(x \mid S)$. The probability of an $x \to y$ proposal is $\pi(x)\pi_S(y \mid S)$, and the probability of a $y \to x$ proposal is $\pi(y)\pi_S(x \mid S)$. It is crucial here, as always, that S is the same in both expressions. If $P_{x\to y}$ is the probability of accepting an $x \to y$ proposal, detailed balance is the formula

$$\pi(x)\pi_S(y \mid S)P_{x\to y} \;=\; \pi(y)\pi_S(x \mid S)P_{y\to x}\ .$$

The usual reasoning suggests that we accept an $x \to y$ proposal with probability

$$P_{x\to y} \;=\; \min\left\{\,1,\ \frac{\pi(y)}{\pi_S(y \mid S)}\cdot\frac{\pi_S(x \mid S)}{\pi(x)}\right\}\ . \tag{11}$$

In the case of a Gaussian π, one can easily modify the proposal used in the walk move to define a density $\pi_S(x \mid S)$ that is an accurate approximation to π if L and $|S|$ are large. This is harder if π is not Gaussian. We have not done computational tests of this method yet.

3 Evaluating ensemble sampling methods

We need criteria that will allow us to compare the ensemble methods above to standard *single particle* methods. Most Monte Carlo is done for the purpose of estimating the expected value of something:

$$A \;=\; E_\pi\left[\,f(X)\,\right] \;=\; \int_{\mathbb{R}^n} f(x)\pi(x)\,dx\ , \tag{12}$$

where π is the target density and f is some function of interest.[2] Suppose $X(t)$, for $t = 1, 2, \ldots, T_s$, are the successive states of a single particle MCMC sampler for π. The standard single particle MCMC estimator for A is

$$\widehat{A}_s \;=\; \frac{1}{T_s} \sum_{t=1}^{T_s} f(X(t)) \; . \tag{13}$$

An ensemble method generates a random path of the ensemble Markov chain $\vec{X}(t) = (X_1(t), \ldots, X_L(t))$ with invariant distribution Π given by (5). Let T_e be the length of the ensemble chain. The natural ensemble estimator for A is

$$\widehat{A}_e \;=\; \frac{1}{T_e} \sum_{t=1}^{T_e} \left(\frac{1}{L} \sum_{k=1}^{L} f(X_k(t) \right) \; . \tag{14}$$

When $T_s = LT_e$, the two methods do about the same amount of work, depending on the complexity of the individual samplers.

For practical Monte Carlo, the accuracy of an estimator is given by the asymptotic behavior of its variance in the limit of long chains [13][10]. For large T_s we have

$$\mathrm{var}\left[\widehat{A}_s\right] \;\approx\; \frac{\mathrm{var}_\pi\left[f(X)\right]}{T_s/\tau_s} \quad , \tag{15}$$

where τ_s is the *integrated autocorrelation time* given by

$$\tau_s \;=\; \sum_{t=-\infty}^{\infty} \frac{C_s(t)}{C_s(0)} \; , \tag{16}$$

[2] The text [9] makes a persuasive case for making this the definition: *Monte Carlo* means using random numbers to estimate some number that itself is not random. Generating random samples for their own sakes is *simulation*.

and the lag t *autocovariance* function is

$$C_s(t) \;=\; \lim_{t'\to\infty} \mathrm{cov}\left[\, f(X(t'+t))\,,\, f(X(t'))\,\right] \;. \tag{17}$$

We estimate τ_s from the time series $f(X(t))$ using a shareware procedure called `Acor` [14] that uses a variant (described below) of the self consistent window method of [7].

Define the ensemble average as

$$F(\vec{x}) \;=\; \frac{1}{L}\sum_{k=1}^{L} f(x_k) \;.$$

Then (14) is

$$\widehat{A}_e \;=\; \frac{1}{T_e}\sum_{t=1}^{T_e} F(\vec{X}(t)) \;.$$

The analogous definitions of the autocovariance and integrated autocorrelation time for the ensemble MCMC method are:

$$\tau_e \;=\; \sum_{t=-\infty}^{\infty} \frac{C_e(t)}{C_e(0)} \;,$$

with

$$C_e(t) \;=\; \lim_{t'\to\infty} \mathrm{cov}\left[\, F(\vec{X}(t'+t))\,,\, F(\vec{X}(t'))\,\right] \;.$$

Given the obvious relation (Π in (5) makes the walkers in the ensemble independent)

$$\mathrm{var}_{\Pi}\left[F(\vec{X})\right] \;=\; \frac{1}{L}\mathrm{var}_{\pi}\left[f(X)\right] \;,$$

the ensemble analogue of (15) is

$$\mathrm{var}\left[\widehat{A}_e\right] \approx \frac{\mathrm{var}_\pi\left[f(X)\right]}{LT_e/\tau_e} \quad .$$

The conclusion of this discussion is that, in our view, a sensible way to compare single particle and ensemble Monte Carlo is to compare τ_s to τ_e. This compares the variance of two estimators that use a similar amount of work. Comparing variances is preferred to other possibilities such as comparing the mixing times of the two chains [4] for two reasons. First, the autocorrelation time may be estimated directly from Monte Carlo data. It seems to be a serious challenge to measure other mixing rates from Monte Carlo data (see, however, [5] for estimating the spectral gap). Second, the autocorrelation time, not the mixing rate, determines the large time error of the Monte Carlo estimator. Practical Monte Carlo calculations that are not in this large time regime have no accuracy.

Of course, we could take as our ensemble method one in which each $X_k(t)$ is an independent copy of a single Markov chain sampling π. The reader can easily convince herself or himself that in this case $\tau_e = \tau_s$ exactly. Thus such an ensemble method with $T_e = LT_s$ would have exactly the same large time variance as the single particle method. Furthermore with $T_e = LT_s$ the two chains would require exactly the same computation effort to generate. The two methods would therefore be indistinguishable in the long time limit.

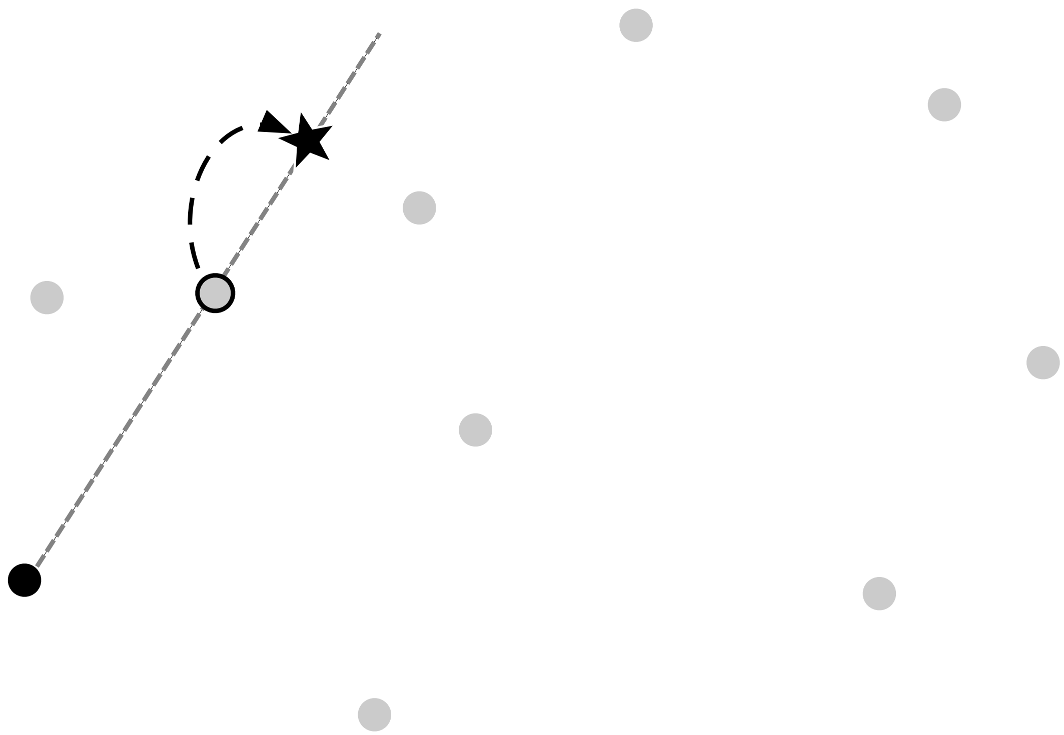

Figure 2: A stretch move. The light dots represent the walkers not participating in this move. The dot with the dark border represents X_k and the dark dot represents X_j. The thick dashed arrow connects X_k to the proposed new location, Y, marked by a dark star. The proposal is generated by stretching along the grey dashed line connecting X_j to X_k.

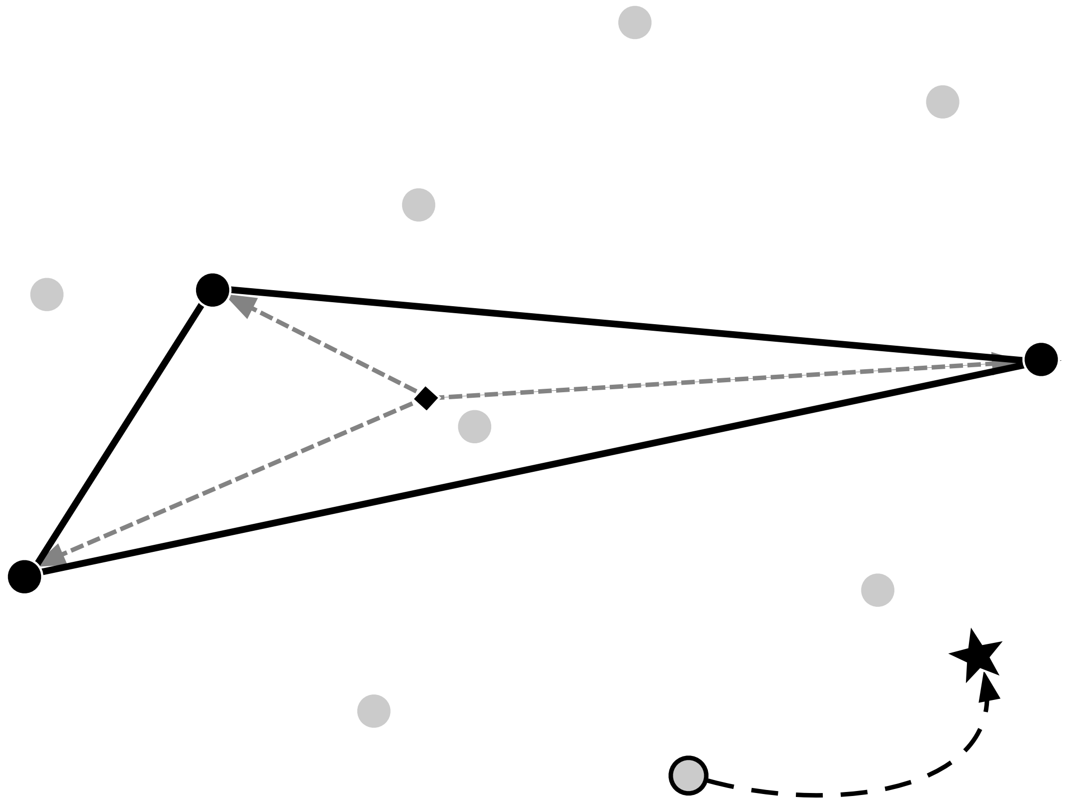

Figure 3: A Walk move. The dots represent the ensemble of particles. The dark ones represent the walkers in $\vec{X}_S$. The dot with the dark border represents X_k. The black dashed arrow connects X_k to the proposed position Y, marked by a dark star. The proposed perturbation has covariance equal to the sample covariance of the three dark dots. The perturbation is generated by summing random multiples of the dashed grey arrows. The black diamond represents the sample mean $\overline{X}_S$.

4 Computational tests

In this section we present and discuss the results of computational experiments to determine the effectiveness of our ensemble methods relative to a standard single particle Markov chain Monte Carlo method. The MCMC method that we choose for comparison is the single site Metropolis scheme in which one cycles through the coordinates of $X(t)$ perturbing a single coordinate at a time and accepting or rejecting that perturbation with the appropriate Metropolis acceptance probability before moving on to the next coordinate. For the perturbations in the Metropolis scheme we choose Gaussian random variables. All user defined parameters are chosen (by trial and error) to optimize performance (in terms of the integrated autocorrelation times). In all cases this results in an acceptance rate close to 30%. For the purpose of discussion, we first present results from tests on a difficult 2-dimensional example. The second example is a 101-dimensional, badly scaled distribution which highlights the advantages of our scheme.

4.1 The Rosenbrock density.

In this subsection we present numerical tests on the Rosenbrock density, which is given by[3]

$$\pi(x_1, x_2) \propto \exp\left(-\frac{100({x_2} - {x_1}^2)^2 + (1 - x_1)^2}{20}\right). \tag{18}$$

[3] To avoid confusion with earlier notation, in the rest of this section (x_1, x_2) represents an arbitrary point in $\mathbb{R}^2$.

Contours of the Rosenbrock density are shown in Figure 4. Though only 2-dimensional, this is a difficult density to sample efficiently as it exhibits the scaling and degeneracy issues that we have discussed throughout the paper. Further the Rosenbrock density has the feature that there is not a single affine transformation that can remove these problems. Thus in some sense this density is designed to cause difficulties for our affine invariant estimators. Of course its degeneracy will cause problems for the single particle estimator and we will see that the affine invariant schemes are still superior.

Tables 1 and 2 present results for the functionals $f(x_1, x_2) = x_1$ and $f(x_1, x_2) = x_2$ respectively. The times should be multiplied by 1000 because we subsampled every Markov chain by 1000. In both cases, the best ensemble sampler has an autocorrelation time about ten times smaller than that of isotropic Metropolis. The walk move method with $|S| = 3$ has autocorrelation times a little more than twice as long as the stretch move method. All estimates come from runs of length $T_s = 10^{11}$ and $T_e = T_s/L$. In all cases we estimate the autocorrelation time using the `Acor` procedure [14].

To simulate the effect of $L = \infty$ (infinite ensemble size), we generate the complementary X_j used to move X_k by independent sampling of the Rosenbrock density (18). For a single step, this is exactly the same as the finite L ensemble method. The difference comes in possible correlations between steps. With finite L, it is possible that at time $t = 1$ we take $j = 4$ for $k = 5$ (i.e. use $X_4(1)$ to help move $X_5(1)$, and then use $j = 4$ for $k = 5$ again at the next time $t = 2$. Presumably, possibilities like this become unimportant as $L \to \infty$.

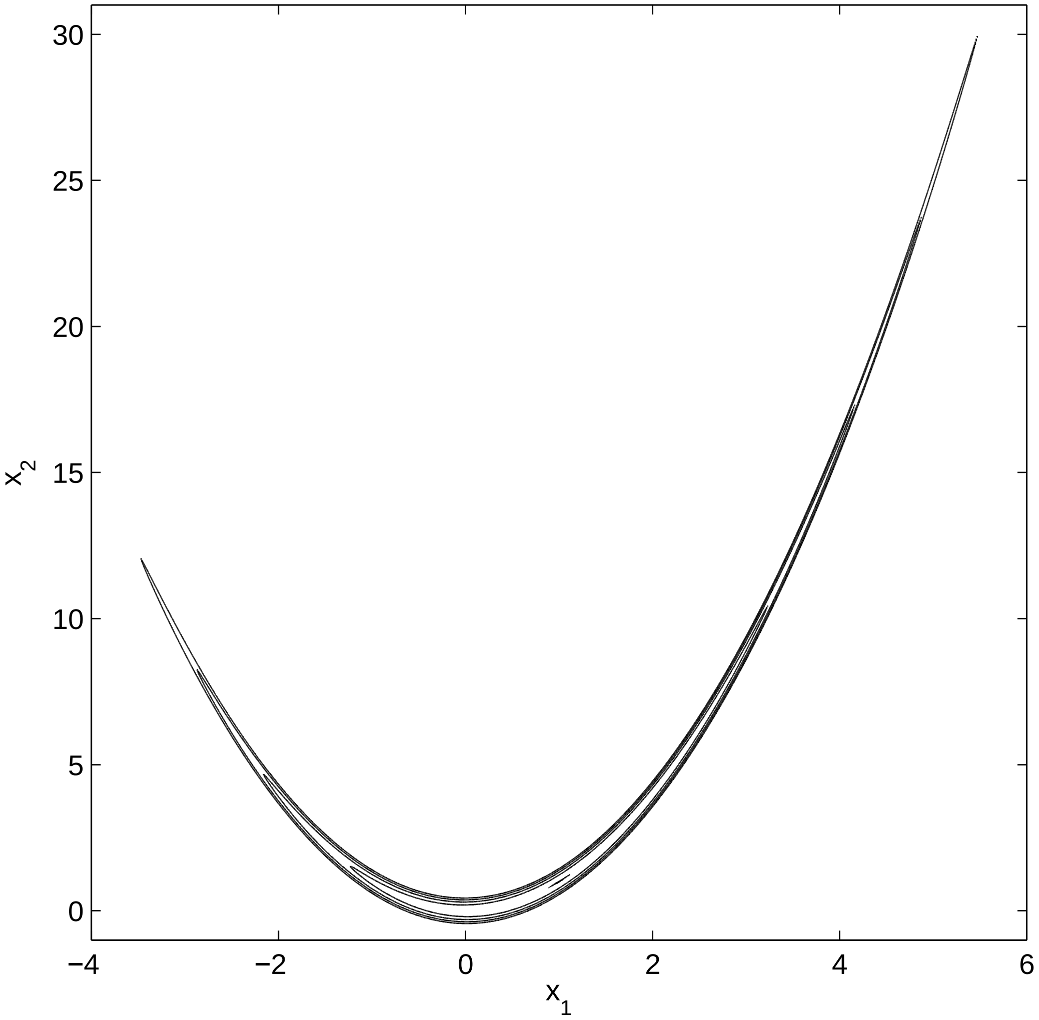

Figure 4: Contours of the Rosenbrock density.

x_1 auto-correlation times $(\times 10^{-3})$

	ensemble size			
method	1	10	100	∞
Metropolis	163	–	–	–
Stretch moves	–	19.4	8.06	8.71
Walk moves, $\lvert S\rvert = 3$	–	46.4	19.8	18.6

Table 1: Auto-correlation times (multiplied by 10^{-3}) with $f(x_1, x_2) = x_1$ for single particle isotropic Metropolis and the chains generated by the two ensemble methods. The ensemble methods with ensemble size $L = \infty$ generate complementary walkers by exact sampling of the Rosenbrock density. The per-step cost of the methods are roughly equivalent on this problem.

We sample the Rosenbrock density using the fact that the marginal of X is Gaussian, and the conditional density of Y given X also is Gaussian.

Finally, we offer a tentative explanation of the fact that stretch moves are better than walk moves for the Rosenbrock function. The walk step, W, is chosen using three points as in Figure 3, see (10). If the three points are close to X_k, the covariance of W will be skewed in the same direction of the probability density near X_k. If one or more of the X_m are far from X_k, the simplex formed by the X_m will have the wrong shape. In contrast, the stretch move only requires that we choose one point X_j in the same region as X_k. This suggests that it might be desirable to use proposals which depend on clusters of near by particles. We have been unable to find such a method that is at the same time reasonably quick and has the Markov property, and is even approximately affine invariant. The replacement move may have promise in this regard.

x_2 auto-correlation times ($\times 10^{-3}$)

	ensemble size			
method	1	10	100	∞
Metropolis	322	–	–	–
Stretch moves	–	67.0	18.4	23.5
Walk moves, $\lvert S\rvert = 3$	–	68.0	44.2	47.1

Table 2: Auto-correlation times (multiplied by 10^{-3}) with $f(x_1, x_2) = x_2$ for single particle isotropic Metropolis and the chains generated by the two ensemble methods. The ensemble methods with ensemble size $L = \infty$ generate complementary walkers by exact sampling of the Rosenbrock density. The per-step cost of the methods are roughly equivalent on this problem.

4.2 The invariant measure of an SPDE.

In our second example we attempt to generate samples of the infinite dimensional measure on continuous functions of $[0, 1]$ defined formally by

$$\exp\left(-\int_0^1 \frac{1}{2}u_x(x)^2 + V(u(x))\,dx\right) \tag{19}$$

where V represents the double well potential

$$V(u) = (1 - u^2)^2.$$

This measure is the invariant distribution of the stochastic Allen Cahn equation

$$u_t = u_{xx} - V'(u) + \sqrt{2}\,\eta \tag{20}$$

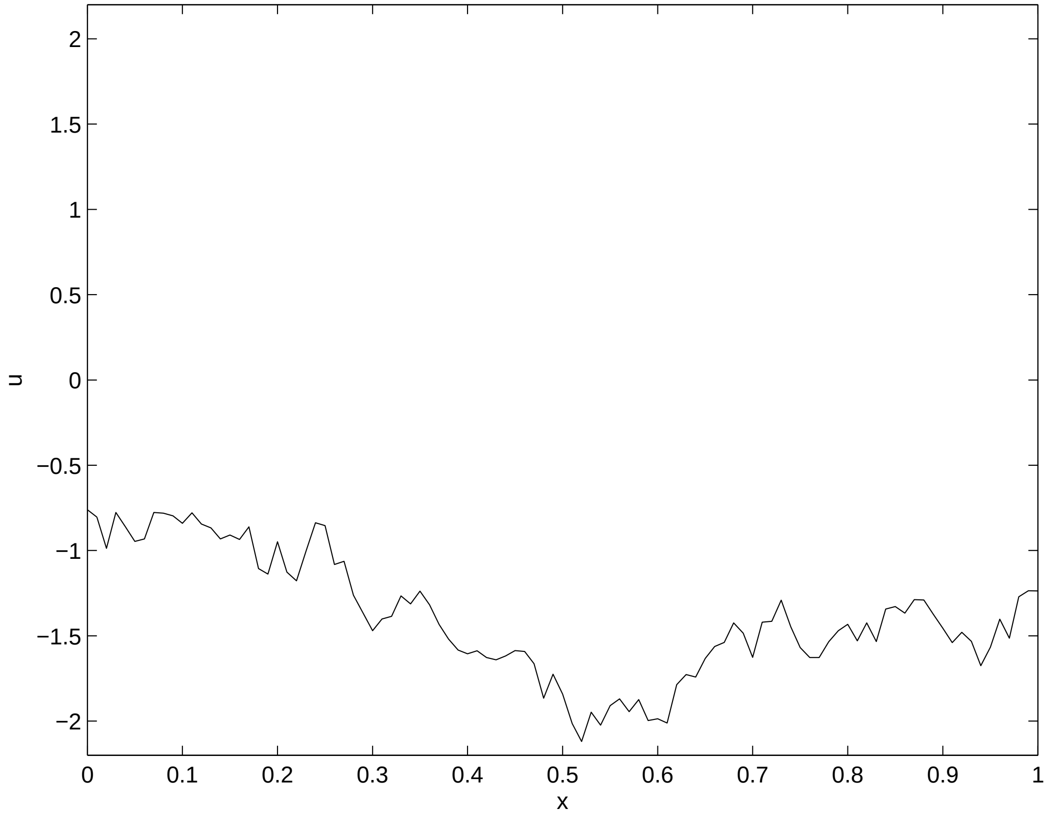

Figure 5: Sample path generated according to π in (21).

with free boundary condition at $x = 0$ and $x = 1$ (see [3, 12]). In these equations η is a space time white noise. Samples of this measure tend to resemble rough horizontal lines found either near 1 or near -1 (see Figure 5).

In order to sample from this distribution (or approximately sample from it) one must first discretize the integral in (19). The finite dimensional distribution can then be sampled by Markov chain Monte Carlo. We use the discretization

$$\begin{aligned}&\pi(u(0), u(h), u(2h)\ldots, u(1)) = \\ &\qquad \exp\left(-\sum_{i=0}^{N-1} \frac{1}{2h}\left(u((i+1)h) - u(ih)\right)^2 + \frac{h}{2}\left(V(u((i+1)h) + u(ih)\right)\right) \qquad (21)\end{aligned}$$

where N is a large integer and $h = \frac{1}{N}$. This distribution can be seen to converge to (19) in an appropriate sense as $N \to \infty$. In our experiments we choose $N = 100$. Notice that the first term in (21) strongly couples neighboring values of u in the discretization while the entire path roughly samples from the double well represented by the second term in (21).

For this problem we compare the auto correlation time for the function

$$f(u(0), u(h), \ldots, u(1)) = \sum_{i=0}^{N-1} \frac{h}{2}\left(u((i+1)h) + u(ih)\right) \qquad (22)$$

which is the trapezoidal rule approximation of the integral

$$\int_0^1 u(x)\, dx.$$

As before we use $|S| = 3$ for the walk step and $T_e = T_s/L$ where $T_s = 10^{11}$ and $L = 102$. As with most MCMC schemes that employ global moves (moves of many or all components at a time), we expect the performance to decrease somewhat as one considers larger and larger problems. However, as the integrated auto correlation times reported in Table 3 indicate, the walk

f **auto-correlation times** $(\times 10^{-3})$

	ensemble size
method	102
Metropolis	80
Stretch moves	5.2
Walk moves, $\lvert S \rvert = 3$	1.4

Table 3: Auto-correlation times with f given in (22) for single particle Metropolis and the chains generated by the two ensemble methods. Note that in terms of CPU time in our implementation, the Metropolis scheme is about 5 times more costly per step than the other two methods. We have not adjusted these autocorrelation times to incorperate the extra computational requirements of the Metropolis scheme.

move outperforms single site Metropolis by more than a factor of 50 on this relatively high dimensional problem. Note that in terms of CPU time in our implementation, the Metropolis scheme is about 5 times more costly per step than the other two methods tested. We have not adjusted the autocorrelation times in Table 3 to incorperate the extra computational requirements of the Metropolis scheme.

5 Software

Most of the software used here is available on the web [14]. We have taken care to supply documentation and test programs, and to create easy general user interfaces. The user needs only to supply procedures in `C` or `C++` that evaluate $\pi(x)$ and $f(x)$, and one that supplies the starting ensemble $\vec{X}(1)$. We appreciate feedback on user experiences.

The `Acor` program for estimating τ uses a self consistent window strategy related to that of [7] to estimate (17) and (16). Suppose the problem is to estimate the autocorrelation time for a time series, $f^{(0)}(t)$, and to get an error bar for its mean, $\overline{f}$. The old self consistent window estimate of τ (see (16) and [13]) is

$$\widehat{\tau}^{(0)} \;=\; \min\left\{ s \;\middle|\; 1 + 2\sum_{1\leq t\leq Ms} \frac{\widehat{C}^{(0)}(t)}{\widehat{C}^{(0)}(0)} = s \right\} \;, \tag{23}$$

where $\widehat{C}(t)$ is the estimated autocovariance function

$$\widehat{C}^{(0)}(t) \;=\; \frac{1}{T-t}\sum_{t'=1}^{T-t} \left(f^{(0)}(t') - \overline{f}\right)\left(f^{(0)}(t+t') - \overline{f}\right) \;. \tag{24}$$

The window size is taken to be $M = 10$ in computations reported here. An efficient implementation would use an FFT to compute the estimated autocovariance function. The overall running time would be $O(T\ln(T))$.

The new `Acor` program uses a trick that avoids the FFT and has an $O(T)$ running time. It computes the quantities $\widehat{C}^{(0)}(t)$ for $t = 0, \ldots, R$. We used $R = 10$ in the computations presented here. If (23) indicates that $M\widehat{\tau} > R$, we restart after a pairwise reduction

$$f^{(k+1)}(t) \;=\; \frac{1}{2}\left(f^{(k)}(2t) \;+\; f^{(k)}(2t+1)\right) \;.$$

The new time series is half as long as the old one and its autocorrelation time is shorter. Repeating the above steps (24) and (23) successively for $k = 1, 2, ...$ gives an overall $O(T)$ work bound. Of course, the (sample) mean of the time series $f^{(k)}(t)$ is the same $\overline{f}$ for each k. So the error bar is the same too. Eventually we should come to a k where (23) is satisfied for $s \leq R$. If not, the procedure reports failure. The most likely cause is that the original time series is too short relative to its autocorrelation time.

6 Conclusions

We have presented a family of many particle ensemble Markov chain Monte Carlo schemes with an affine invariance property. Such samplers are uniformly effective on problems that can be rescaled by affine transformations to be well conditioned. All Gaussian distributions and convex bodies have this property. Numerical tests indicate that even on much more general distributions our methods can offer significant performance improvements over standard single particle methods. The computational cost of our methods over standard single particle schemes is negligible.

7 Acknowledgements

Our work grew out of discussions of a talk by Fox at the 2009 SIAM meeting on Computational Science. We are grateful to our colleagues Jeff Cheeger, Esteban Tabak, and Eric Vanden-Eijnden for useful comments and suggestions. Both authors were supported in part by the Applied Mathematical Sciences Program of the U.S. Department of Energy under Contract DEFG0200ER25053

References

[1] Andrieu, C. and Thoms, J., “A tutorial on adaptive MCMC,” Stat. Comput, 18, pp. 343-373, 2008.

[2] Christen, J., “A general purpose scale-independent MCMC algorithm,” Comunicacin Tcnica PE/CIMAT, I-07-16, 2007

[3] Da Prato, G. and Zabczyk, J., *Ergodicity for infinite dimensional systems*, London Mathematical Society Lecture Note Series 229, Cambridge University Press, Cambridge, 1996.

[4] Diaconis, P. and Saloff-Coste, L., “What do we know about the Metropolis algorithm?” J. Comput. System. Sci, 57(1), pp. 20-36 27th annual ACM Symposium on the Toelry of Computing (STOC95).

[5] Gade, K., PhD Thesis, Computer Science Department, Courant Institute of Mathematical Sciences, New York University, 2008. `http://www.stat.unc.edu/faculty/cji/Sokal.pdf`

[6] Gill, P.E., Murray, W., and Wright, M.H., *Practical Optimization*, Academic Press, 1982.

[7] Goodman, J. and Sokal, A., “Multigrid Monte Carlo method. Conceptual foundations,” Phys. Rev. D, 40(6), pp. 2035 – 2071 1989.

[8] John, F., “Extremum problems with inequalities as subsidiary conditions,” in *Studies and Essays Presented to R. Courant on his 60th Birth-*

day, January 8, 1948, Interscience Publishers, New York, 1948, pp. 187 – 204.

[9] Kalos, M. and Whitlock, P., *Monte Carlo Methods*, Wiley-VCH, Weinheim, 2008.

[10] Liu, J.S., *Monte Carlo Strategies in Scientific Computing*, Springer-Verlag, New York, LLC, 2001.

[11] Nelder, J. A. and Mead, R., "A Simplex Method for Function Minimization", Computer Journal, 7, pp. 308 - 313, 1965.

[12] Reznikoff, M. and Vanden-Eijnden, E., "Invariant measures of stochastic partial differential equations," C.R. Math. Acad. Sci. Paris, 340(4), pp. 305–308 2005.

[13] Sokal, A., "Monte Carlo Methods in Statistical Mechanics: Foundations and New Algorithms."

[14] `http://www.math.nyu.edu/faculty/goodman/software/`