2011 50th IEEE Conference on Decision and Control and European Control Conference (CDC-ECC)
Orlando, FL, USA, December 12-15, 2011

Modeling the Transient Behavior of Stochastic Gradient Algorithms

Roger Brockett

Abstract— **We investigate the transient behavior of a class of stochastic gradient algorithms. Unlike the analysis usually applied to stochastic approximation and simulated annealing which focuses on the rate of convergence and the asymptotic limit, we take a more detailed look at the transient behavior with the goal of better understanding how the global structure of the performance measure influences the behavior of the algorithm. For the sake of tractability, we work with a specific class of problems characterized by gradients with easily characterized stationary points. Our prototype involves stochastic algorithms for ordering a numerical list, a problem which is the subject of a recent paper in the condensed matter physics literature, focusing on hysteretic effects in annealing. These authors raise several questions of interest in studying stochastic dynamics which inspired this paper.**

I. INTRODUCTION

Stochastic approximation is a standard tool in learning theory that puts emphasis on getting the precise answer, even if it takes a very long time. It works by reducing the rate of descent over time, setting up a situation where the law of large numbers is applicable. However, it usually happens that at some point, time becomes more important than accuracy. Typically this means that it is necessary to settle for an answer that corresponds to a transient state rather than the equilibrium state, implying that a deeper understanding of the transients would be of value.

In condensed matter physics there are observations relating to nonequilibrium effects that have been difficult to model but now seem to be responding to some new methods. For example, the recent paper by Ling-Nan Zou and Sidney Nagel [1] studies nonequilibrium effects associated with a stochastic sorting algorithm. Writing in reference to hysteretic effects in annealing they say, “We find that sorting can display many features of a glass, even for lists as small as N=5”.

Inspired by their numerical experiments, we look at a similar problem, involving a continuous time gradient flow. It shares important features with their work while lending itself to somewhat more insightful mathematical analysis. As they acknowledge in their work, sorting a moderately sized list of numbers is a routine problem, scarcely meriting a new detailed investigation. The reason for using it as an example here is that in a stochastic context it has some analytical aspects that seem to shed light on more difficult problems.

In [1] the authors consider sorting a list of numbers using a discrete time stochastic algorithm which, at each stage, picks a pair of adjacent elements of the list and either reverses their order or leaves it unchanged based on a probabilistic rule. This rule involves comparing the difference in the “energy” associated with the two states and contains a parameter, which can be thought of as modeling temperature. In some stages of their numerical experiments they vary the temperature capturing some aspects of a simulated annealing algorithm. In this paper we also study list sorting based on extremizing an energy function. We consider several different energy functions, including one that is essentially the same as the one used by Zou and Nagel.

In both physics and in optimization algorithms there is often a local vs. global issue that hinges on the availability of short paths between states. Our model allows for different levels of connectivity using various formulations ranging from the Toda lattice model, which allows only nearest neighbor interactions, to the completely connected double bracket flows. The latter evolves in a set of symmetric matrices with fixed eigenvalues and takes the form

$$\dot{H} = [H, [H, \phi']]$$

where $[A, B] = AB - BA$and ϕ is a differentiable function of H which is to be maximized. Nearest neighbor (Toda) flow is a special case of this. This equation, has a natural stochastic version which, when written as an Itô equation, is

$$dH = [H, [H, N]]dt = \sum [\Omega_{ij}, H]dw_{ij} + \frac{1}{2}[\Omega_{ij}, [\Omega_{ij}, H]]dt$$

In [1] the function ϕ is something like

$$\phi(H) = \beta \operatorname{tr}\left(\mathrm{diag(H)}\right)^2 - \left(\mathrm{diag(SHS^T)}\right)^2 + \mathrm{tr(HN)}$$

with S being a shift (super diagonal ones) matrix. In this case the gradient ascent equation is

$$\dot{H} = [H, [H, \beta \;\mathrm{diag}\;(\mathrm{H - SHS^T}) + \mathrm{N}]]$$

and the equilibria occur when H is diagonal and when $(H - SHS^T) + N]$ has two or more repeated entries on the diagonal. Stability demands that diag H and diag $(H - SHS^T)$ be similarly ordered. A one parameter family that captures a number of interesting features is

$$\dot{H} = [H, [H, N + \alpha \mathrm{diag(H)}]]$$

If α is zero this equation has $n!$ equilibria but only one of these is stable. When α is large it has $n!$ stable equilibria and many unstable equilibria as well.

After this introduction we proceed in steps.

A: We recall some properties of a well studied nonlinear

This work was supported in part by the Army Research Office,under the aMURI award N0. N00014-01-1-0859. The views expressed in this paper are those of the author alone and not of the sponsor

Roger Brockett is with School of Engineering and Applied Sciences, Harvard University , Cambridge MA USA, 02138 brockett@seas.harvard.edu

978-1-61284-799-3/11/$26.00 ©2011 IEEE

flow having many equilibria.
B: We describe a family of particularly simple gradient flows.
C: We introduce a stochastic version and compute expectations, variances, etc.
D: We describe a related Markov chain model for the non equilibrium behavior.

The author is grateful to Dr. James Phillips of Rutgers University for calling his attention to the work of Zou and Nagel and subsequent correspondence.

II. Gradient Flows with many Local Maxima

To give a better understanding of the flows being discussed it may be helpful to provide some further background material. Define a linear operator mapping square matrices into square matrices

$$ad_H(X) = [H, X] \text{ or } \mathrm{ad_H}(\cdot) = [\mathrm{H}, \cdot\ ]$$

This operator has eigenvalues that are the pairwise differences of the eigenvalues of H. Its null space is the set of matrices that commute with H. Let ad_H^{-1} denote the Moore-Penrose inverse of ad_H relative to the matrix inner product $\mathrm{tr}(A^TB)$.

Define $Sym(\Lambda)$ to be the set of all real symmetric matrices with eigenvalues $\Lambda = \{\lambda_1, \lambda_2, ..., \lambda_n\}$, all distinct. The so-called normal metric on $Sym(\Lambda)$ is a Riemannian metric defined on the space of symmetric matrices with a given set of (distinct) eigenvalues. The differential description is

$$(ds)^2 = \langle ad_H^{-1}(dH), ad_H^{-1}(dH)\rangle$$

It is to be noted that ad_H is one-to-one and onto as a map of the tangent space of $Sym(\Lambda)$ into itself. Given a function ϕ on a Riemannian manifold with metric G, the gradient ascent equation is $\dot{x} = G^{-1}\phi'$. In our context $G^{-1}(\cdot) = [H, [H, \cdot]]$ so a function $\phi(H)$ gives rise to the gradient assent equation $\dot{H} = [H, [H, \phi']]$.

One way to think about this metric involves the idea that any two symmetric matrices with the same set of eigenvalues, say H_1 and H_2, are related by $H_1 = \Theta H_2 \Theta^T$ for some orthogonal matrix Θ. If the eigenvalues of H_1, and hence those of H_2, are distinct then the matrix Θ that relates them is almost unique; it is unique to within a multiplication on the right by a diagonal matrix with diagonals that are ± 1. Of course any orthogonal matrix can be written as $\Theta = e^\Omega$ for some skew symmetric matrix Ω. Let $||\Omega||$ denote the square root of the sum of the squares of the entries of Ω; i.e., its Frobenius norm. Then if $||\Omega||$ is small $\Theta = e^\Omega$ is close to the identity and $H_2 = \Theta H_1 \Theta^T$ is close to H_1. In fact, if we define the distance between H_1 and H_2 to be the smallest value of $||\Omega||$ consistent with $e^\Omega H_1 e^{-\Omega} = H_2$ then this distance is the same as the distance measure coming from the Riemannian metric defined above.

Observe that for

$$H(\theta) = \exp\begin{bmatrix} 0 & \theta \\ -\theta & 0 \end{bmatrix}\begin{bmatrix} a & 0 \\ 0 & b \end{bmatrix} exp\begin{bmatrix} 0 & -\theta \\ \theta & 0 \end{bmatrix}$$

We have

$$H(\theta) = \begin{bmatrix} \frac{a+b}{2} + \frac{a-b}{2}\cos 2\theta & \frac{b-a}{2}\sin 2\theta \\ \frac{b-a}{2}\sin 2\theta & \frac{a+b}{2} - \frac{a-b}{2}\cos 2\theta \end{bmatrix}$$

with θ parametrizing the path by arc length. So, by this measure, two diagonal matrices in $Sym(\Lambda)$ which differ by virtue of a transposition of two diagonal entries are $\pi/\sqrt{2}$ units apart.

We now observe that in certain important cases the solution of $\dot{H} = [H, [H, N]]$ is simply a reparametrization of a geodesic.

Lemma: If $H(0)$ and N are symmetric and if for $\Omega = [H(0), N]$ we have $[H(0), \Omega] = k_1 N$ and $[N, \Omega] = k_2 H(0)$ then there is a rescaling of time, $t \mapsto \alpha(t)$ such that

$$H(t) = e^{\Omega\alpha(t)} H(0) e^{-\Omega\alpha(t)}$$

satisfies the equation $\dot{H} = [H, [H, N]]$

Proof: Let $[H(0), N] = \Omega$. The derivative of $H(t)$ is

$$\frac{d}{dt} e^{\Omega\alpha} H(0) e^{-\Omega\alpha} = \dot{\alpha} e^{\Omega\alpha}[\Omega, H(0)] e^{-\Omega\alpha}$$

Some manipulation shows that with this assumption about the form of the solution, the differential equation for H is equivalent to

$$\dot{\alpha}[\Omega, H(0)] = [H(0), [H(0), e^{-\Omega\alpha} N e^{\Omega\alpha}]]$$

Using the hypothesis we see that

$$e^{\Omega\alpha} N e^{-\Omega\alpha} = rH(0) + sN$$

Thus

$$[H(0), e^{-\Omega\alpha} N e^{\Omega\alpha}] = u\Omega$$

and hence that the right-hand side is proportional to the left. This completes the proof.

As an example, if N = diag (a,b) and $H(0) = [0, 1; 1, 0]$ the flow $\dot{H} = [H, [H, N]]$ the equation for α is

$$\dot{\alpha} = \frac{(b-a)}{2}\sin 2\alpha$$

which has the solution

$$\alpha(t) = \tan^{-1}\left(\tan\left(\theta\left(0\right)\right) e^{(b-a)t}\right)$$

The distance between diagonal matrices related by a cyclic permutations is also easily computed. For example, the distance between

$$\begin{bmatrix} a & 0 & 0 \\ 0 & b & 0 \\ 0 & 0 & c \end{bmatrix} \text{ and } \begin{bmatrix} b & 0 & 0 \\ 0 & c & 0 \\ 0 & 0 & a \end{bmatrix}$$

is $2\sqrt{2}\pi/3$, about one-third larger than the distance between the identity and a transposition.

In an earlier paper [2] we investigated the behavior of the solutions of the equation in symmetric matrices

$$\dot{H} = [H, [H, N]]$$

Bloch [3] developed this further and subsequently there has developed considerable further work, e.g., [4]. Under the assumption that the eigenvalues of $H(0)$ and N are symmetric with distinct eigenvalues, It was shown that this equation flows to the value of H that maximizes the trHN. More generally, if $\phi(H)$ is any differentiable function of H and if we write $d\phi(H)/dH = \phi'$, then, as we just observed, the flow

$$\dot{H} = [H, [H, \phi']]$$

is the gradient ascent equation relative to the normal metric.

In the case of the function $\phi(H) = \mathrm{tr(HN)}$ with N having distinct eigenvalues, there are exactly $n!$ values of H where the gradient vanishes but only one of these is a local maximum and thus a global maximum. However, this is an exceptional situation not shared by most choices of ϕ. Letting diag(H) denote the diagonal of H, the function $\phi(H) = \mathrm{tr(HN + \alpha Hdiag(H))}$ may have as many as $n!$ local maxima if α is sufficiently large. Other choices for ϕ having many extrema have been explored in connection with the assignment problem in Brockett and Wong [5] and Wong [6].

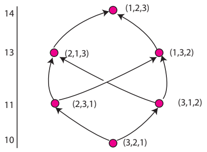

Fig. 1. Let $n = 3$. Showing the 3! equilibria identified with the corresponding permutations, the value of tr(HN), and some possible trajectories joining the equilibria

A simple modification of this model allows for an adjustment of the levels of connectivity. Let π be a projection mapping skew-symmetric matrices into themselves; i.e., $\pi(\pi(\cdot)) = \pi(\cdot)$. Then the modified equation

$$\dot{H} = [H, \pi([H, \phi'])]$$

also evolves with unchanging spectrum. It is easy to verify that

$$\frac{d}{dt}\mathrm{tr(HN)} = \mathrm{tr}\left([\mathrm{H},\mathrm{N}]\pi([\mathrm{H},\mathrm{N}]\right) = \mathrm{tr}\left((\pi([\mathrm{H},\mathrm{N}])^2\right)$$

If π is the identity map then of course we see that trHN is weakly monotone increasing and only fails to increase if H and N commute. For arbitrary projection trHN is still weakly monotone but now fails to increase when $\pi([H, N])$ is zero. If, for example, π projects onto the skew-symmetric tridiagonal matrices then we can say that there is only nearest neighbor connectivity; if it projects onto a skew-symmetric band matrix of width five there is greater connectivity, etc.

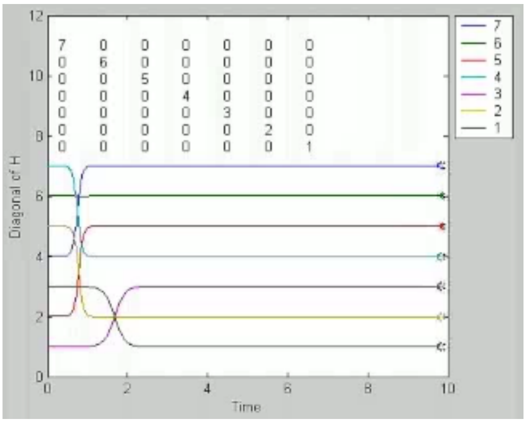

Fig. 2. Showing the final value of H and the evolution of the diagonal terms of the solution of the full double bracket equation for a particular initial condition. Here $n = 7$ and

III. The Tridiagonal Case

We begin with an exploration of a special case of the equation $\dot{H} = [H, [H, N]]$ as treated in [7]. In the section N is the diagonal matrix $N = \mathrm{diag}\ (1, 2, ..., \mathrm{n})$. In this case the tridiagonal matrices are an invariant manifold for $\dot{H} = [H, [H, N]]$. It is common to use the notation

$$H = \begin{bmatrix} b_1 & a_1 & 0 & 0 & 0 & ... & 0 & y_n \\ a_1 & b_2 & a_2 & 0 & 0 & ... & 0 & 0 \\ 0 & b_2 & b_3 & a_3 & ... & ... & 0 & 0 \\ ... & ... & ... & ... & ... & ... & ... & ... \\ 0 & 0 & ... & 0 & 0 & ... & b_{n-1} & a_{n-1} \\ a_n & 0 & ... & 0 & 0 & ... & a_{n-1} & b_n \end{bmatrix}$$

The resulting equations coincide with the equations proposed by Toda, as recast by Flascka. These equations are a special case of the Lax form, widely studied in the theory of integrable systems. As has been observed, this equation flows without changing the eigenvalues of H. Except for initial conditions corresponding to a set of measure zero, the resting state will be the equilibrium point $H = \mathrm{diag}\ (\lambda_1, \lambda_2, ..., \lambda_\mathrm{n})$ where $\lambda_1 > \lambda_2 >, ..., > \lambda_n)$. That is to say, this equation acts to find and sort the eigenvalues.

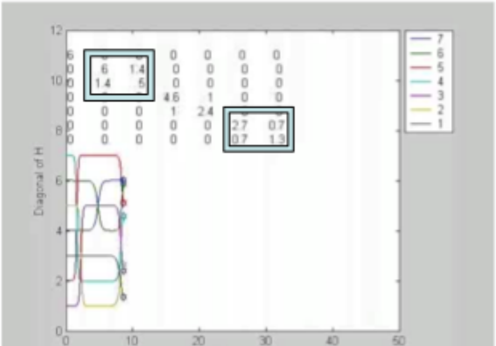

Fig. 3. Showing a portion of the solution of the tridiagonal situation as a pair of transpositions are occurring. The traces show the values of the various diagonal entries. The matrix shows the value of H at the moment the simulation is stopped. The boxes superimposed on the matrix highlight some of the entries taking part in a transposition.

In iteration described in [1] a pair of nearest neighbors is picked and possibly interchanged, meaning that only nearest neighbors are transposed. In the tridiagonal case equations all interactions are nearest neighbor relations as well. Figures 3 and 4 show a solution of these equations in the case $n = 7$. Notice that the solution spends most of its time close to an equilibrium point. Even more, over the interval shown it spends most of its time close to an unstable equilibrium point.

Using the standard identities

$$\sum h_{ij}^2 = \sum \lambda_i^2 \; ; \; \sum h_{ii} = \sum \lambda_i$$

we can develop a bound on how large the off-diagonal terms can be. Minimizing the sum of the squares of the diagonal terms of H, subject to the constraint on their sum yields

$$\sum h_{ii}^2 \geq \frac{1}{n}\left(\sum \lambda_i\right)^2 \implies \sum_{i \neq j} h_{ij}^2 \leq \sum \lambda_i^2 - \frac{1}{n}\left(\sum \lambda_i\right)^2$$

The right-hand side can be thought as being n times the "sample variance" of the eigenvalues. By including the constraint tr$HN = m$ we can sharpen this to

$$\sum_{i \neq j} h_{ij}^2 \leq \mathrm{trH}^2 - \frac{(\mathrm{trH})^2}{\mathrm{trI}^2} - \frac{(\mathrm{trH\hat{N}})^2}{\mathrm{tr\hat{N}}^2}$$

where $\hat{N} = N - (\mathrm{trN})\mathrm{I}$. The more closely clustered the eigenvalues and the closer trHN is to an extreme value, the tighter the bound on the off-diagonal terms. This bound is, of course, valid for the full double bracket flow, not just the tridiagonal flow.

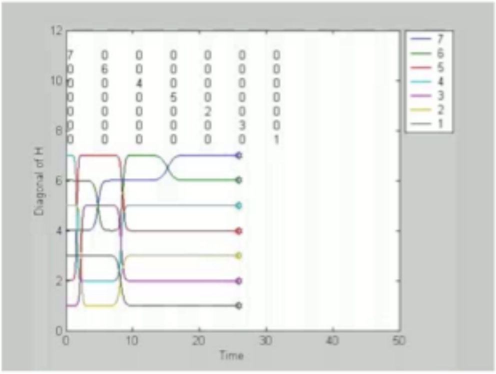

Fig. 4. Showing trajectories of the Toda sorter with a prolonged meta stable period. The decrease in connectivity, as compared to the simulation shown in figure 2, results in a correspondingly slower evolution of the process.

In comparing the flow produced by a tridiagonal H with the flow produced with a full symmetric matrix it is of interest to recall the theorem of Cayley relating the minimum number of transpositions t needed to generate a particular permutation of n objects to the number of cycles c in the permutation. It asserts that $t = n - c$. In the case of the tridiagonal flow being used as an illustration here, the number of cycles associated with permuting (7,6,5,4,3,2,1) to (4,6,2,7,1,5,3) is 4 an so at least 3 permutations are required. The fully connected flow illustrated in figure 2 generates just three but the tridiagonal sorter is less efficient and uses nine.

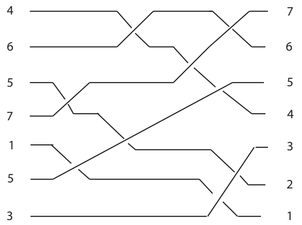

Fig. 5. Showing the "braid" corresponding to the trajectories of the diagonals of the tridiagonal flow. Time flows from left to right. There is a total of the nine transpositions. Our convention is that when h_{ii} crosses h_{jj} the larger of the two passes over the smaller.

IV. Stochastic Gradients

The density equation is a basic tool in statistical mechanics. It has a variety of names, forms and interpretations but here we want to call attention to the way it is used in the literature on NMR. In this setting it is used to describe the statistical properties of a many particle quantum system which is not completely isolated from the environment because of interactions with unmodeled dynamics. The unmodeled dynamics interact with the system of primary interest contributing thermal noise shaped by the natural frequencies of the unmodeled dynamics. These stochastic effects are modeled as additive terms in the density equation. In the NMR literature these are often called Lindblad terms.

The density equation is written in terms of a Hermitean matrix having trace 1. In the physics literature it is denoted by the letter ρ but to be consistent with our earlier notation, we use the letter H. The basic density equation takes the form $\dot{H} = [H_0, H]$ with H_0 being the Hamiltonian of the system of interest. If H has point spectrum the solutions evolve in such a way as to keep the eigenvalues constant. The stochastic version can be described using an Itô equation of the form

$$dH = [H_0, H] + \sum [\Omega_i, H] dw_i + \frac{1}{2} \sum [\Omega_i, [\Omega_i, H]] dt$$

with Ω skew-Hermitean. The last two terms on the right model the heat bath. This solution of this equation also evolves with unchanging eigenvalues. The corresponding equation for the expectation of the density is

$$\frac{d}{dt} \mathcal{E} H = [H_0, \mathcal{E} H] + \frac{1}{2} \sum [\Omega_i, [\Omega_i, \mathcal{E} H]]$$

The stochastic equation describes a flow on a manifold of dimension $n(n-1)/2$ imbedded a euclidean space of dimension $n(n+1)/2$. The operation of taking expectations involves taking convex combinations of paths lying in the manifold but because the manifold is not a convex subset of the larger space these linear combinations need not be points in $Sym(\Lambda)$. Thus the equation for the expectation of H will not evolve in $Sym(\Lambda)$ and $\mathcal{E} H$ need not have the same eigenvalues as H. See figure 6. In fact, it is clear that for typical values of Ω the equation for the expectation will decay to $(1/n)I$ which matches $H(0)$ only to the extent that it has the same trace.

We adopt a similar model for our noisy gradient flow. In n dimensions there are $n(n-1)/2$ 2-planes generated by selecting elements two at a time from an orthonormal basis. If we want isotropic noise, in the sense that when H is perturbed to $\Theta H\Theta^T$ the choice of Θ should not favor any element of the orthogonal group over any other, then we need this many independent noise terms, each Ω_{ij} of the form $e_ie_j^T - e_je_i^T$. In this case the Lindblad terms take the form

$$\sum[\Omega_{ij},H]dw_{ij} - \frac{1}{2}\left(nH - \mathrm{tr(H)I}\right)dt$$

Putting these terms into the full nonlinear gradient equation, gives

$$dH = [H,[H,\phi']]dt + \sum \alpha[\Omega_{ij},H]dw_{ij} - \alpha^2\frac{nH - I\mathrm{trH}}{2}$$

and linearizing this about an equilibrium point for which $H = \mathrm{diag}(\lambda_1,\lambda_2,...,\lambda_\mathrm{n})$ and $\phi'(H) = \mathrm{diag(m_1,m_2,...,m_n)}$ we get a set of $n(n-1)/2$ decoupled equations for the off-diagonal elements

$$dh_{ij} = a_{ij}h_{ij}dt + \alpha b_{ij}dw_{ij} - (\alpha^2/2)h_{ij}dt$$

where $a_{ij} = -(\lambda_i - \lambda_j)(m_i - m_j)$ and $b_{ij} = (\lambda_i - \lambda_j)$. If all the a_{ij} are negative, as would be the case if the equilibrium point is a local maximum then we can compute the steady state variances

$$\mathcal{E}h_{ij}^2 = -\frac{\alpha^2(\lambda_i - \lambda_j)^2}{-(\lambda_i - \lambda_j)(m_i - m_j) + \alpha^2/2}$$

but if the equilibrium in question is not a local maximum the linearized model does not have a steady state. With the present choice of stochastic terms, the equation for the expected value of H takes the form

$$\frac{d}{dt}\mathcal{E}H = \mathcal{E}[H,[H,\phi']] + \alpha^2\left(n\mathcal{E}H - \mathrm{tr(H)I}\right)$$

Notice that the Itô term has the effect of adding damping to the equation for the expected value and that the equation for the expected value does not evolve in the isospectral manifold.

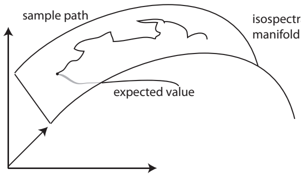

Fig. 6. Showing a sample path in the isospectral manifold and the solution of the equation for the expected value evolving off the isospectral manifold.

It happens that there is an explicit formula for the steady state solution of the Fokker-Planck equation associated with our stochastic descent equation. It takes the form

$$\rho_{\mathrm{ss}}(H) = \frac{1}{Z}e^{-2\phi(H)/\alpha^2)}$$

where Z is the constant necessary to normalize ρ. (See, e.g., [8] where this example is treated and put in a broader context.) This implies that in steady state the surface(s) of constant probability coincide with surface(s) of constant ϕ and, consequently, for small α the density is peaked about the minimum value of $\phi(H)$. This argument is an essential part of the verification that simulated annealing works in a continuous time, continuous space setting. In particular, if the maximum value of ϕ occurs when if $\phi'(H)$ is diagonal then with high probability the solution is nearly diagonal when α is small and the system is in steady state.

What the above analysis suggests can be summarized with the help of figure 7. Notice that for values of α such that $\alpha^2(n\mathcal{E}H - \mathrm{tr(H)I})$ dominates $\mathcal{E}[H,[H,\phi']]$ the expected value of H is close to $(\mathrm{tr(H)/n})\mathrm{I}$. For small values of α the flow is close to the deterministic flow and consequently it is close to the isospectral manifold and, when in equilibrium, is close to being diagonal. For such α the term $\mathcal{E}[H,[H,\phi']]$ is nearly diagonal as is H. Thus we see that if ϕ has multiple local maxima, as we reduce α from a large value, more equilibria will appear, consistent with the fact that at $\alpha = 0$ all the equilibria associated with the gradient flow will be present.

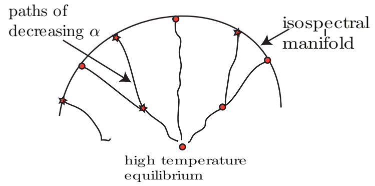

Fig. 7. Schematic showing the effect of reducing α on the creation of additional equilibria as α goes to zero and the isospectral manifold is approached.

V. A Markov Process Approximation

As one sees from the simulations, the sample paths spend relatively long periods of time near the equilibria of the equation $\dot{H} = [H,[H,\phi'(H)]]$. This is the case even if the equilibria are not stable. A more quantitative study of this is partially summarized in Figure 8. The vertical lines denote moments in time, with time advancing from left to right. On the left of each line is shown an ordering of the integers one through seven. On the right of each line is the numerical value of the eigenvalue of the linearized system with each entry being associated with the nearest neighbor transposition that would interchange the two integers bracketing it on the left. It may be observed that the order in which the transpositions occur is related to the relative sizes of the unstable eigenvalues. For example, for the initial ordering, the largest unstable eigenvalue is 5, leading to the transposition of 7 and 2. The next largest is 4, corresponding to an interchange of 5 and 1, etc.

Because, as we have seen, the trajectories spend most of their time near equilibria, we now turn our attention to the possibility of approximating the behavior of the system by a continuous time Markov chain whose states are the various equilibria, and whose transition probabilities are to be estimated on the basis of the properties of the solutions near

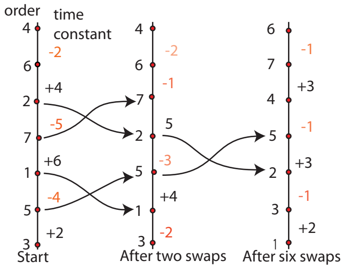

Fig. 8. Diagramming the sequence of transpositions of the numbers 1,2,...,7 determined by the Toda flow of the earlier section. The numbers labeled as time constants refer to the negative of the eigenvalue of the linearized gradient flow (the Hessian) which goes along with the particular transposition indicated. Large negative eigenvalues generate rapid transitions.

equilibrium points. A closer examination of the trajectories generated by the Toda reinforces the intuitive idea that there should be a relationship between the "degree of instability" of a particular eigen-direction and the time which elapses before the solution leaves that equilibrium point traveling in that eigen-direction. We propose a quantitative version defining a Markov chain whose transition probabilities are proportional to the size of the unstable eigenvalues. We identify the states of the Markov chain with the $n!$ possible diagonal forms of H. The transition rate associated with a transition which is not a simple transposition is taken to be zero. The transition rate associated with the interchange of the diagonal elements i and j is take to be zero if the interchange results in a decrease of trHN and is otherwise taken to be

$$a_{ij} = (n_i - n_j)(h_{ii} - h_{jj})$$

A glance at Figure 4 shows that after about ten units of time there are three unstable eigenvalues, all equal to +1, and yet the observed transition times associated with these modes are widely separated, reinforcing the idea that a probabilistic description is appropriate.

Local maxima lead to stable equilibria and these need to be treated differently from those that are unstable; for stable equilibria the role of the noise is critical because without noise the trajectory will never leave the equilibrium point. In order to get a useful estimate for the transition times it would be helpful to know the probability distribution for the exit time associated with a suitable open set around the equilibrium point. Recall the equation from section 2

$$\dot{\theta} = \frac{(b-a)}{2}\sin 2\theta$$

which describes gradient flow leading to the transposition of diagonal elements a and b

If $a > b$ then the equilibrium solution $\theta = 0$ is asymptotically stable and the equilibrium solution $\theta = \pi/2$ is unstable. Expressed in terms of θ, the stochastic equation

$$dh_{ij} = a_{ij}h_{ij}dt + \alpha b_{ij}dw_{ij} - (\alpha^2/2)h_{ij}dt$$

is simply

$$d\theta = \frac{(b-a)}{2}\sin 2\theta dt + \alpha dw$$

Assuming that $\theta(0) = 0$, we would like to know the probability distribution of the first time θ leaves the interval $(-\pi, \pi)$.

Recall that for the simpler process governed by $d\theta = \alpha dw$; $x(0) = 0$ the probability density for the first exit time from $(-r, r)$ is the Levy distribution $\rho(T) = (r/\alpha\sqrt{2\pi T^3})e^{r^2/2\alpha^2 T}$, making the most likely exit time $r^2/3\alpha^2$. (See [9].) If we use this as a crude approximation to our situation then we would set the rate of transition from an stable equilibrium state to a second equilibrium state differing only in that λ_i and λ_j have been interchanged on the diagonal of H as

$$a_{ij} = \frac{3\alpha^2}{2(\lambda_i - \lambda_j)\pi^2}$$

The theory of large deviations, and more specifically, the Wentzell-Freidlin theory provides a method to estimate the expected time required to leave a region about a stable equilibrium [10]. To interchange two diagonal elements of H, say a and b, the corresponding value of the off diagonal element must reach the value $h = |a - b|/2$. This is the barrier that must be crossed. A basic part of this analysis in the case of an equation such as $dx = f(x)dt + g(x)dw$ is to solve a minimum energy transfer for the deterministic equation $\dot{x} = f(x) + g(x)u$.

VI. Acknowledgments

The author thanks Anthony Bloch for helpful comments on an earlier draft.

References

[1] Ling-Nan Zou and Sidney Nagel, " Glassy Dynamics in Thermally Activated List Sorting", *Physics Review Letters* , paper No. 257201, 25 June, 2010.

[2] R. W. Brockett, "Dynamical Systems That Sort Lists, Diagonalize Matrices and Solve Linear Programming Problems," *Linear Algebra and its applications*, Vol 146 (1991) pp. 79-91. (also *Proceedings of the 1988 IEEE Conference on Decision and Control*, (1988) pp. 799-803.)

[3] A.M. Bloch, "Steepest descent, linear programming and Hamiltonian flows," *Contemporary Mathematics,* A.M.S. 114 (1990), 77-88

[4] U. Helmke and J. B. Moore, *Optimization and Dynamical Systems,* Springer-Verlag, London, 1996.

[5] R. W. Brockett and Wing Wong, "A Gradient Flow for the Assignment Problem," *Progress in System and Control Theory* (G. Conte and B. Wyman, eds.) Birkhäuser, 1991, pp.

[6] W. S. Wong, "Matrix representation and gradient flows for NP-hard problems," *Journal of Optimization Theory and Applications* Volume 87, Number 1, pp. 197-220.

[7] A. M. Bloch, R. W. Brockett and T. Ratiu, "A New Formulation of the Generalized Toda Lattice Equations and their Fixed Point Analysis via the Moment Map," *Bulletin of the American Mathematical Society*, Vol. 23, No 2 (1990) pp. 477-485.

[8] R. W. Brockett, "Notes on Stochastic Processes on Manifolds," *Systems and Control in the Twenty-First Century*, (C. Byrnes, et al., eds.) pg 75-101, Birkhäuser, Boston, 1997.

[9] Øksendal, Bernt K. *Stochastic Differential Equations: An Introduction with Applications* Springer, Berlin, 2003.

[10] Amir Dembo and Ofer Zeitouni, *Large Deviations Techniques and Applications*