Proceedings of the 28th Conference
on Decision and Control
Tampa, Florida • December 1989

ON STOCHASTIC STABILIZATION[1]

Roger W. Brockett

Division of Applied Sciences
Harvard University
Cambridge, MA 02138

Abstract

In some settings the deliberate introduction of noise can have a beneficial effect on the rate of convergence of numerical algorithms for finding the minima of functions. This is an important aspect of simulated annealing. However, there are situations which, in contrast to simulated annealing, do not involve reducing the noise level with time and which are also of interest in a computational and control context. In this paper we consider problems of this type. The effect of the addition of a noise term to the solution of a differential equation is investigated via properties of the associated Fokker-Plank operator. In the last section of the paper we indicate the relevance of this to our earlier paper on the descent equation $\dot{H} = [H,[H,N]]$.

1. Introduction

If we are given a function ϕ which takes on real values and if we wish to find the value of the argument of ϕ for which it takes on its minimum, an obvious thing to do is to set $\dot{x} = -\nabla \phi(x)$ and integrate this equation. This method can fail to give satisfactory results for several reasons. The most commonly cited problem is that the descent leads to a local minimum which is not a global one. A second difficulty also observed in numerical work is that the descent proceeds very slowly because of unfortunate topological features such as ridges or the existence of one or more saddle points for ϕ, near which $\nabla\phi$ is quite small. If it happens, as in our recent work [1,2], that the function ϕ is defined on a manifold which, by its very nature, demands that any differentiable function defined on it has a number of saddle points, then this second difficulty can be quite significant.

In order to correct the problem of convergent to a local but not global minimum the advocates of simulated annealing recommend the addition of a time dependent noise term so as to obtain

$$\dot{x} = -\nabla \phi + \varepsilon(t)n(t)$$

with n a stationary stochastic process and ε a deterministic function which goes to zero as t goes to infinity. Problems which arise in this context include the question of how to choose ε so as to get a satisfactory rate of convergence without significantly degrading the accuracy of the final answer.

The main point of this paper is to investigate a sample problem for which the addition of a noise term is i) useful in avoiding slowdown at saddle points or ii) useful in achieving stochastic stability in settings where topological considerations exclude the possibility of obtaining a control law which makes the desired equilibrium point asymptotically stable in the large. The particular problem which we study is a disguised form of the matrix equation $\dot{H} = [H,[H,N]]$ whose properties are investigated in [1-3].

Consider a deterministic equation

$$\dot{x} = f(x)$$

Suppose that x belongs to a compact Riemannian manifold and that $x(0)$ is a random variable distributed uniformly with respect to the Riemannian volume on M. In this case we can, in principle, evaluate $\mathcal{E}T_\epsilon$ the expected value of the time it takes for x to arrive within ε of some stable equilibrium point. On the other hand, if we add white noise to the equation to get an Itô equation of the form

$$dx = -\left(f(x) - \frac{1}{2}\frac{\partial g}{\partial x}g\right)dt + \alpha g(x)\,dw$$

then the probability density satisfies

$$\frac{\partial \rho}{\partial t} = -L\rho + \frac{\alpha^2}{2}\frac{\partial^2 \rho}{\partial x^2}$$

and we can ask if, in some probabilistic sense, this improves the rate of convergence. Since these calculations cannot be done explicitly, it seems worthwhile to do some numerical work.

2. The Finite State Case

Consider a finite state continuous time jump process $x(t)$ taking on real number values in the set $s_1, s_2, \ldots, s_n$. Letting $p_i(t)$ be the probability that $x(t) = s_i$ we assume that the probabilities evolve according to

$$\dot{p} = Ap + uBp$$

where u is a parameter (such as the noise level) which we can adjust. Our goal is to minimize a function which is obtained by balancing a desire that $p(t)$ should reach its steady state quickly and a desire that in the steady state there should be very little spread in the values of x, e.g. that

$$\begin{aligned} \sigma &= \mathcal{E}(x - \mathcal{E}x)^2 \\ &= (\mathcal{E}x^2) - (\mathcal{E}x)^2 \\ &= \Sigma(p_i s_i)^2 - (\Sigma p_i s_i)^2 \end{aligned}$$

should be small. Since p belongs to a compact set and since η has a lower bound, σ will have an infinum and it is not unreasonable to we expect it to have a minimum if u belongs to a compact set.

Of course the matrix $A + uB$, being an infinitesimal generator, has zero as an eigenvalue. We let p_∞ denote the corresponding equilibrium probability vector. Taking the rate of decay of the slowest mode as a measure of the speed of convergence and taking the variance as a measure of the quality of convergence we would be led to the consideration of combinations of the form

[1]This work was supported in part by the the U.S. Army Research Office under grant DAAL03-86-K-0171 and in part by the National Science Foundation under grant CDR-85-00108.

$$\eta = f(\lambda_1) + [\Sigma(p_i s_i)^2 - (\Sigma p_i s_i)^2]$$

The λ which shows up here is the real part (necessarily negative) of the eigenvalue of $A + uB$ which is "the next to the largest". Because 0 is an eigenvalue, λ measures the gap between zero and the rest of the spectrum. Its fundamental role in this kind of problem has been observed and studied by others [4].

The problems that we are investigating here involve Fokker-Planck equations which correspond to Itô equations of the form

$$dx = f(x)\, dt + \alpha\, dw$$

In the scalar case we get

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x} f(x)\rho + \frac{\alpha^2}{2}\frac{\partial^2 \rho}{\partial x^2}$$

There is a large literature on how to derive finite state models for such equations. The second order term presents no particular difficulties but we want to point out one aspect of the first order operator.

For smooth functions f and rho the equation

$$\frac{\partial \rho}{\partial t} = -\frac{\partial}{\partial x} f(x)\rho$$

will evolve in such a way as to keep $\int \rho dx$ constant. This insures that the probability is appropriately normalized. If we want the p_i to generate an approximation of ρ, we should insure that p satisfies

$$\dot{p} = Ap$$

with A having columns which sum to one and which are nonnegative off the diagonal. This can be accomplished by approximating $-\partial f\rho/\partial x$ by a forward difference when f is negative and a backward difference when f is positive

$$-\frac{\partial}{\partial x} f\rho \approx \begin{cases} h^{-1}[f(x-h)\rho(x-h) - f(x)\rho(x)] & \text{if } f(x) > 0 \\ h^{-1}[f(x)\rho(x) - f(x+h)\rho(x+h)] & \text{if } f(x) < 0 \end{cases}$$

If we associate a p_i with each equilibrium point and if we use this scheme, we get a matrix A which has the properties of an infinitesimal generator.

3. The Perturbation Formula

Given

$$\dot{p} = Ap + uBp$$

with A and B square matrices whose diagonals are nonpositive, whose off-diagonals are nonnegative and whose columns sum to zero, by the Peron-Frobenius theorem, 0 is an eigenvalue of A and there exists eigenvector p_∞ with nonnegative entries, summing to one, such that $A + uBp_\infty = 0$. If $A + uB$ is irreducible, there is only one choice for p_∞. We investigate how p_∞ depends on u for

$$(A + uB)p_\infty = 0$$

To do so, we introduce the vector

$$c^T = (1, 1, \ldots, 1)$$

and look for v such that $c^T v = 0$ (so that $p_\infty + c^T v$ is a probability vector) and such that

$$\phi(u) = (A + uB)(p_\infty(0) + uv(u))$$

We may as well consider u to be small and to require the right-hand side to vanish to first order in u. This would be completely routine if A were invertible but it cannot be so we need a special device. Note that $Ap_\infty = 0$ and at $u = 0$

$$\frac{\partial p}{\partial u} = Bp_\infty + Av(0)$$

Because of the form of A and B, $c^T A = c^T B = 0$. Thus if

$$0 = Bp_\infty + Av(0)$$

then

$$(A + cc^T)v(0) + Bp_\infty = 0$$

Conversely, if $(A + c^T c)v(0) + Bp_\infty = 0$, then a premultiplication by c^T shows that $cv = 0$ and $Bp_\infty + Av(0) = 0$. However, if A is irreducible, then $(A + cc^T)$ is invertible and we may write

$$v(0) = (A + cc^T)Bp_\infty$$

thus obtaining, to first order, the effect of u on the steady state probability p_∞.

4. An Example

Consider the stochastic equation

$$dx = -\cos x\, dt + \alpha\, dw$$

Because the related deterministic equation

$$\dot{x} = -\cos x$$

has a periodic right-hand side we can regard this as an equation defined on a circle. The deterministic equation has two equilibrium points; one at $x = \pi/2$ and one at $x = 3\pi/2$. The former is unstable and the latter is stable. The Fokker-Planck equation associated with the stochastic equation is

$$\frac{\partial \phi(t,x)}{dt} = \frac{\partial}{\partial x}\cos x \rho(t,x) + \frac{\alpha^2}{2}\frac{\partial \rho}{\partial x^2}$$

It is not difficult to verify that $e^{-2(\sin x)/\alpha^2}$ is annihilated by the operator on the right-hand side and so for

$$N = \int_0^{2\pi} e^{-2(\sin x)/\alpha^2} dx$$

we see that $N^{-1}e^{-2(\sin x)/\alpha^2}$ is the steady state density associated with this equation. Near $x = 3\pi/2$ we have $-\sin(3\pi/2 + \delta) = \cos\delta \approx 1 = \delta^2/2$ and so

$$\rho_\infty \approx N^{-1}e^{-2/\alpha^2} \cdot e^{-\delta^2/\alpha^2}$$

A linearization of the stochastic equation near the same point yields

$$d\delta = -\delta\, dt + \alpha\, dw$$

which has a steady state variance of $\mathcal{E}\delta^2 = \alpha^2/2$; this is fully consistent with the actual solution.

The operator

$$L = \frac{\partial}{\partial x}\cos x + \frac{\alpha^2}{2}\frac{\partial^2}{\partial x^2}$$

defined on a suitable subset of the space of all functions periodic of period 2π has one zero eigenvalue, corresponding to the eigenfunction $e^{-2(\sin x)/\alpha^2}$, and all remaining eigenvalues have real parts which are negative. This operator is not self-adjoint as written but it is similar to a self-adjoint operator and so its spectrum is real. If the first order term was absent, then the spectrum would be $\{0, -\alpha^2/2, -\alpha^2/2, -4\alpha^2/2, -4\alpha^2/2, \ldots\}$. On the other hand, if α is zero, then we have to solve

$$\frac{d}{dx}\cos x\rho(x) = \lambda\rho(x) \quad ; \quad \rho(0) = \rho(2\pi)$$

but this has no solution because $\int_{-G+\pi/2}^{\epsilon+\pi/2} 1/\cos\lambda\, d\lambda$ diverges.

5. Numerical Experiments

Because the Fokker-Planck equation cannot be solved explicitly it is worthwhile to use numerical work to investigate the trade-off between the speed of response and the accuracy with which the zero is ultimately found. We discretize the circle as 50 segments of length $2\pi/50 \approx .125$. In this way we get a Fokker-Planck equation of the form

$$\dot{p} = (A + uB)p$$

with u being the diffusion constant (α^2 in the above context).

We see that L is the sum of an operator which comes from $\frac{\partial}{\partial x}\cos x$ and which we approximate as indicated above and an operator which comes from $\frac{\partial^2}{\partial x^2}$ and which we approximate by a symmetric circulant matrix of the form

$$D = 64\begin{bmatrix} -2 & 1 & & 1 \\ 1 & -2 & \ddots & \\ & \ddots & \ddots & 1 \\ 1 & & 1 & -2 \end{bmatrix}$$

The results of a numerical integrations of $\dot{p} = Lp$ for $\alpha = \sqrt{2}$ are shown in Figure 1. Measurement of the asymptotic rate of convergence of the probability vector indicates that the nonzero eigenvalue with the largest real part decays at the rate of $\approx \exp -1.3t$.

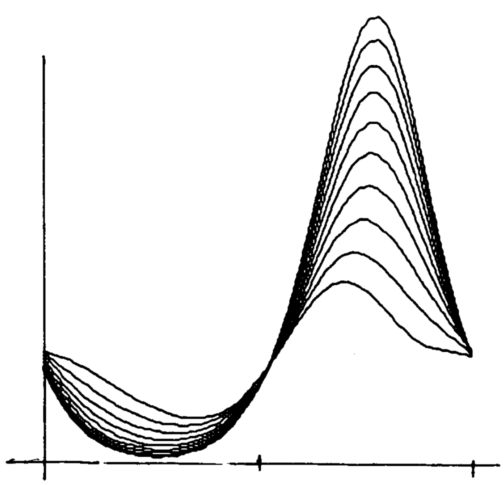

Figure 1. The numerical approximation of the time dependent solution of $\frac{\partial\rho}{\partial t} = \frac{\partial}{\partial x}\cos x\rho + \frac{\partial^2\rho}{\partial x^2}\rho$

6. The Double Commutator Equation

One reason for investigating this question for $\dot{\theta} = \cos\theta$ is that the equation

$$\dot{H} = [H, [H, N]] \quad ; \quad H(0) = H^T(0)$$

whose properties as an analog computer are discussed in [2], has, for typical values of N and $H(0)$, $n!$ equilibrium points, exactly one of which is stable. Even the modest simulation results reported [2] deal with a system having 5040 equilibrium points and thus slowdown near saddle points can be a problem.

In [3] we show that in the two by two case this equation can be reduced to

$$\dot{\theta} = \sin\theta$$

We indicate briefly how one can see the present paper in the more general context. The H-equation evolves in such a way as to keep H symmetric and to keep its eigenvalues constant. The set of real symmetric matrices with a fixed set of distinct eigenvalues admits the structure of an $n(n-1)/2$-dimensional compact, connected manifold. In fact, it is identifiable with the proper orthogonal group acting on $\mathbf{E}^n$ modulo the 2^n element discrete subgroup consisting of diagonal matrices whose diagonal values are ± 1. Denote this so-called hyperoctahedral group by D. Because S0(n) has a natural Riemannian metric, the coset space $S0(n)/D$ can be thought of as a Riemannian manifold with its metric structure coming from that of $S0(n)$. The correspondence between Θ and H is given in terms of the diagonal matrix of eigenvalues of H, say Λ, and is given by

$$\Theta \approx \Theta^T\Lambda\Theta$$

Because $D_1\Lambda D_1 = \Lambda$ for any D_1 in D, one can show that if the diagonal entries of Γ are distinct this establishes a smooth invertible map between $S0(n)/D$ and the set of symmetric matrices with a fixed set of eigenvalues.

A key point of [1] is that $\dot{H} = [H, [H, N]]$ can be viewed as a gradient flow associated with the function $\phi(H) = \text{tr}(HN)$ and the Riemannian metric on H alluded to above.

Now consider the possibility of speeding up the convergence of H by adding a suitable noise term. Of course there are an infinity of possibilities. We consider just one; a natural one in that it is as "isotropic" as possible. Recall that on any Riemannian manifold there is an analog of the Laplacian, called the Laplace-Beltrumi operator. Like the Laplacian in $\mathbf{E}^n$, this operator occurs as the right-hand side of a Fokker-Planck equation associated with a stochastic differential equation (which we write in central difference form)

$$dH = H\,d\Omega - d\Omega H$$

where Ω is a skew-symmetric matrix of Wiener processes. For this equation the density $\rho(t, H)$ satisfies

$$\frac{\partial\rho(t,H)}{\partial t} = \frac{1}{2}L\rho(t,H)$$

with L being the Laplace-Beltrumi operator.

Of course, what we really want to consider is the equation which results when we add this noise term to the gradient equation

$$\dot{H} = \nabla\,\text{tr}(HN)$$

In this case we get

$$dH = \nabla \mathrm{tr}(HN)dt + \varepsilon(H d\Omega - d\Omega H)$$

where ε is the noise level.

This equation has an associated density $\rho(t, H)$. If it were not for the fact that the gradient term we present, the invariant density on H would coincide with the Riemannian volume $SO(n)/D$. It is not difficult to show that the effect of the drift term in the present context is to yield an invariant density which is proportional to $e^{-\mathrm{tr}(NH)}$.

References

[1] R. W. Brockett, "Least Squares Matching Problems," *Linear Algebra and Its Applications*, 1989. (to appear)

[2] R. W. Brockett, "Dynamical Systems That Sort Lists, Diagonalize Matrices and Solve Linear Programming Problems," *Proceedings of of the 1988 IEEE Conference on Decision and Control*, December 1988.

[3] R. W. Brockett, "Smooth Dynamical Systems Which Realize Arithmetic and Logical Operations," in *Three Decades of Mathematical System Theory. A Collection of Surveys at the Occasion of the Fiftieth Birthday of Jan C. Willems.* Lecture Notes in Control and Information Sciences Vol. 135 (H. Nijmeijer and J. M. Schumacher, eds.) Springer Verlag, Berlin, 1989. (to appear)

[4] Richard Holley and Daniel Stroock, "Simulated Annealing via Sobolev Inequalities," to appear.