Scenes from Exclusive-Or: Back Propagation is Sensitive to Initial Conditions

John F. Kolen & Jordan B. Pollack
Laboratory for Artificial Intelligence Research
Computer and Information Science Department
The Ohio State University
Columbus, Ohio 43210, USA
kolen-j@cis.ohio-state.edu, pollack@cis.ohio-state.edu

This paper explores the effect of initial weight selection on feed-forward networks learning simple functions with the back-propagation technique. We first demonstrate, through the use of Monte Carlo techniques, that the magnitude of the initial condition vector (in weight space) is a very significant parameter in convergence time variability. In order to further understand this result, additional deterministic experiments were performed. The results of these experiments demonstrate the extreme sensitivity of back propagation to initial weight configuration.

Back Propagation (Rumelhart, Hinton, & Williams, 1986) is the network training method of choice for many cognitive modeling projects, and for good reason. Like other weak methods, it is simple to implement, faster than many other "general" approaches, well-tested by the field, and easy to mold (with domain knowledge encoded in the learning environment) into very specific and efficient algorithms.

Rumelhart et al. made a confident statement: for many tasks, "the network rarely gets stuck in poor local minima that are significantly worse than the global minima."(p. 536) According to them, initial weights of exactly 0 cannot be used, since symmetries in the environment are not sufficient to break symmetries in initial weights. Since their paper was published, the convention in the field has been to choose initial weights with a uniform distribution between plus and minus ρ, usually set to 0.5 or less.

The convergence claim was based solely upon their empirical experience with the back propagation technique. Since then, Minsky & Papert (1988) have argued that there exists no proof of convergence for the technique, and several researchers (Judd 1988; Blum and Rivest 1988; Kolen 1988) have found that the convergence time must be related to the difficulty of the problem, otherwise an unsolved computer science question ($P \stackrel{?}{=} NP$) would finally be answered. We do not wish to make claims about convergence of the technique in the limit (with vanishing step-size), or the relationship between task and performance, but wish to talk about a pervasive behavior of the technique which has gone unnoticed for several years: the sensitivity of back propagation to initial conditions.

Initially, we performed empirical studies to determine the effect of learning rate, momentum rate, and the range of initial weights on t-convergence (Kolen and Goel, 1989). We use the term *t–convergence* to refer to whether or not a network, starting at a precise initial configuration, could learn to separate the input patterns (correct outputs above or below .5) within t epochs. The experiment consisted of training a 2-2-1 network on exclusive-or while varying three independent variables in 114 combinations: learning rate, η, equal to 1.0 or 2.0; momentum rate, α, equal to 0.0, 0.5, or 0.9; and initial weight range, ρ, equal to 0.1 to 0.9 in 0.1 increments, and 1.0 to 10.0 in 1.0 increments. Each combination of parameters was used to initialize and train a number of networks.[1] Figure 1 plots the percentage of t-convergent (where t = 50,000 epochs of 4 presentations) initial conditions for the 2-2-1 network trained on

[1]Numbers ranged from 8 to 8355, depending on availability of computational resources. Those data points calculated with small samples were usually surrounded by data points with larger samples.

the exclusive-or problem. From the figure we thus conclude the choice of $\rho \leq 0.5$ is more than a convenient symmetry-breaking default, but is quite necessary to obtain low levels of nonconvergent behavior.

Why do networks exhibit the behavior illustrated in Figure 1? While some might argue that very high initial weights (i.e. $\rho > 10$) lead to very long convergence times since the derivative of the semi-linear sigmoid function is effectively zero for large weights, this does not explain the fact that when ρ is between 2 and 4, the non-t-convergence rate varies from 5 to 50 percent.

Thus, we decided to utilize a more deterministic approach for eliciting the structure of initial conditions giving rise to t-convergence. Unfortunately, most networks have many weights, and thus many dimensions in initial-condition space. We can, however, examine 2-dimensional slices through the space in great detail. A slice is specified by an origin and two orthogonal directions (the X and Y axes). In the figures below, we vary the initial weights regularly throughout the plane formed by the axes (with the origin in the lower left-hand corner) and collect the results of running back-propagation to a particular time limit for each initial condition. The map is displayed with grey-level linearly related to time of convergence: black meaning not t-convergent and white representing the fastest convergence time in the picture. Figure 2 is a schematic representation of the networks used in this and the following experiment. The numbers on the links and in the nodes will be used for identification purposes. Figures 3 through 11 show several interesting "slices" of the the initial condition space for 2-2-1 networks trained on **exclusive-or**. Each slice is compactly identified by its 9-dimensional weight vector and associated learning/momentum rates. For instance, the vector (-3+2+7-4X+5-2-6Y) describes a network with an initial weight of -0.3 between the left hidden unit and the left input unit. Likewise, "+5" in the sixth position represents an initial bias of 0.5 to the right hidden unit. The letters "X" and "Y" indicate that the corresponding weight is varied along the X- or Y-axis from -10.0 to +10.0 in steps of 0.1. All the figures in this paper contain the results of 40,000 runs of back-propagation (*i.e.* 200 pixels by 200 pixels) for up to 200 epochs (where an epoch consists of 4 training examples).

Figures 12 and 13 present a closer look at the sensitivity of back-propagation to initial conditions. These figures zoom into a complex region of Figure 11; the captions list the location of the origin and step size used to generate each picture.

Sensitivity behavior can also be demonstrated with even simpler functions. Take the case of a 2-2-1 network learning the **or** function. Figure 14 shows the effect of learning "or" on networks (+5+5-1X+5-1Y+3-1) and varying weights 4 (X-axis) and 7 (Y-axis) from -20.0 to 20.0 in steps of 0.2. Figure 15 shows the same region, except that it partitions the display according to equivalent solution networks after t-convergence (200 epoch limit), rather than the time to convergence. Two networks are considered equivalent[2] if their weights have the same sign. Since there are 9 weights, there are 512 (2^9) possible network equivalence classes. Figures 16 through 25 show successive zooms into the central swirl identified by the XY coordinate of the lower-left corner and pixel step size. After 200 iterations, the resulting networks could be partitioned into 37 (both convergent and nonconvergent) classes. Obviously, the smooth behavior of the t-convergence plots can be deceiving, since two initial conditions, arbitrarily alike, can obtain quite different final network configuration.

Note the triangles appearing in Figures 19, 21, 23 and the mosaic in Figure 25 corresponding to the area which did not converge in 200 iterations in Figure 24. The triangular

[2]For rendering purposes only. It is extremely difficult to know precisely the equivalence classes of solutions, so we approximated.

boundaries are similar to fractal structures generated under iterated function systems (Barnsley 1988): in this case, the iterated function is the back propagation learning method. We propose that these fractal-like boundaries arise in back-propagation due to the existence of multiple solutions (attractors), the non-zero learning parameters, and the non-linear deterministic nature of the gradient descent approach. When more than one hidden unit is utilized, or when an environment has internal symmetry or is very underconstrained, then there will be multiple attractors corresponding to the large number of hidden-unit permutations which form equivalence classes of functionality. As the number of solutions available to the gradient descent method increases, the more complicated the non-local interactions between them. This explains the puzzling result that several researchers have noted, that as more hidden units are added, instead of speeding up, back-propagation slows down (e.g. Lippman and Gold, 1987). Rather than a hill-climbing metaphor with local peaks to get stuck on, we should instead think of a many-body metaphor: The existence of many bodies does not imply that a particle will take a simple path to land on one. From this view, we see that Rumelhart et al.'s claim of back-propagation usually converging is due to a very tight focus inside the "eye of the storm".

The emergence of chaotic phenomena in neural network models was certainly anticipated by both Kurten (1989) and Huberman (1987), carefully avoided (through the choice of symmetric weights) by Hopfield (1982), but usually disregarded by connectionists, including Rumelhart et al., despite the common knowledge that non-linearity is what enables these models to perform non-trivial computations in the first place.

What does this mean to the back-propagation community? From an engineering applications standpoint, where only the solution matters, nothing at all. From a cognitive science standpoint, when making claims that a back-propagation learning experiment is relevant to psychological data, **the initial conditions for the network need to be precisely specified or made publicly available.**

What about the future of back-propagation? We remain neutral on the issue of its ultimate convergence, but our result points to a few directions for improved methods. Since the slowdown occurs as a result of global influences of multiple solutions, an algorithm for first factoring the symmetry out of both network and training environment (e.g. domain knowledge) may be helpful. Furthermore, it may also turn out that search methods which harness "strange attractors" ergodically guaranteed to come arbitrarily close to some subset of solutions might work better than methods based on strict gradient descent. Thus, rather than abandoning back-propagation, we view this result as yet further impetus to discover how to exploit the information-creative aspects of non-linear dynamical systems for future models of cognition (Pollack 1989).

Acknowledgments

This work was partially supported by the Office of Naval Research. Substantial free use of over 200 Sun workstations was generously provided by our department.

References

M. Barnsley. 1988. *Fractals Everywhere*, Academic Press, San Diego, CA. 1988.

A. Blum and R. Rivest. 1988. Training a 3-node Neural Network is NP-Complete. *Proceedings of IEEE Conference on Neural Information Processing Systems*, Denver, Colorado, 1988.

J. J. Hopfield. 1982. Neural Networks and Physical Systems with Emergent Collective Computational Abilities. *Proceedings US National Academy of Science* **79**:2554-2558.

B. A. Huberman and T. Hogg. 1987. Phase Transitions and Artificial Intelligence. *Artificial Intelligence*, **33**, 155-172.

S. Judd. 1988. Learning in Networks is Hard. *Journal of Complexity* **4**:177-192.

J. Kolen. 1988. Faster Learning Through a Probabilistic Approximation Algorithm. *Proceedings of the Second IEEE International Conference on Neural Networks*, San Diego, California, pp. I:449-454.

J. Kolen and A. Goel. 1989. Learning in Parallel Distributed Processing Networks: Computational Complexity and Information Content. (Tech. Rep., 89-JK-LEARNING). Columbus, Ohio: The Ohio State University, Laboratory for AI Research.

K. E. Kurten and J. W. Clark. 1986. Chaos in Neural Networks. *Physics Letters*, **114A**, 413-418.

R. P. Lippman and B. Gold. 1987. Neural Classifiers Useful for Speech Recognition. In *1st International Conference on Neural Networks* ,IEEE, pp. IV:417-426.

M. L. Minsky and S. A. Papert. 1988. *Perceptrons*. Cambridge, MA: MIT Press.

J. B. Pollack. 1989. Implications of Recursive Auto Associative Memories. In *Advances in Neural Information Processing Systems*. (ed. D. Touretzky) pp 527-536, Morgan Kaufman, San Mateo.

D. E. Rumelhart, G. E. Hinton, and R. J. Williams. 1986. Learning Representation by Back-Propagating Errors. *Nature* **323**:533-536.

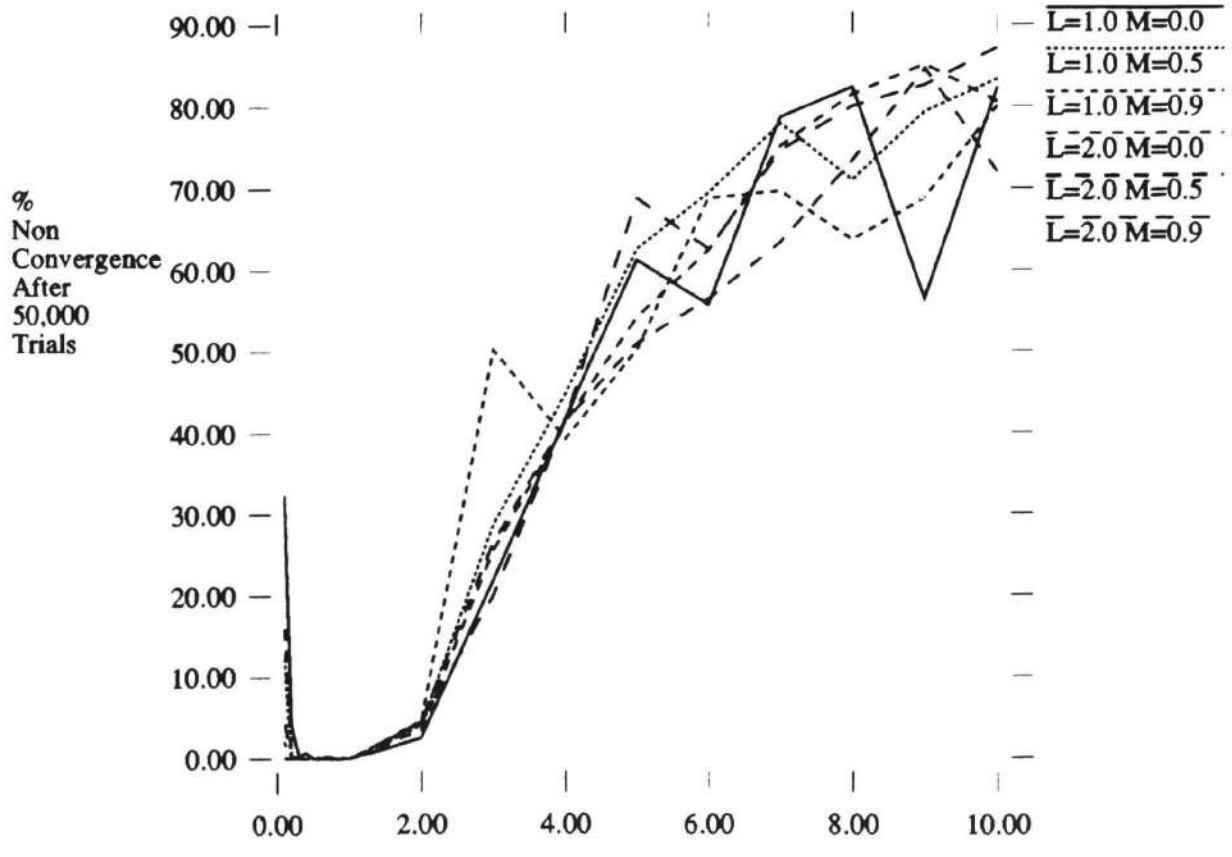

Figure 1: Percentage T-Convergence vs. Initial Weight Range

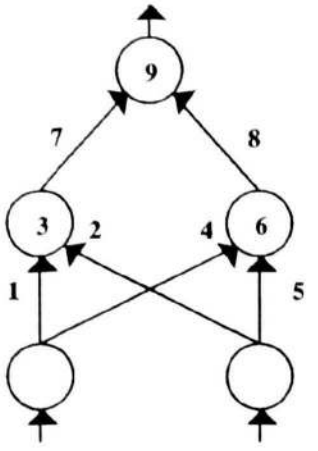

Figure 2 : Schematic Network

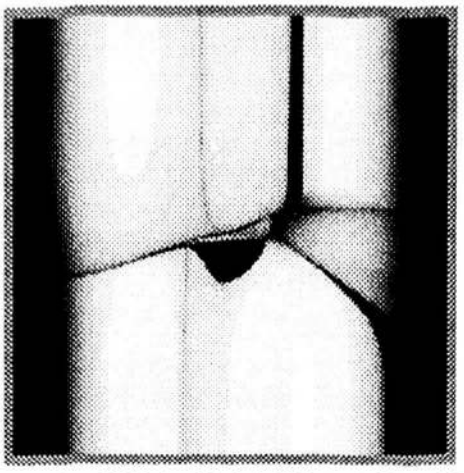

Figure 3 : (-5-3+3+6Y-1-6+7X) η=3.25 α=0.40

Figure 4 : (+4-7+6+0-3Y+1X+1) η=2.75 α=0.00

Figure 5 : (-5+5+1-6+3XY+8+3) η=2.75 α=0.80

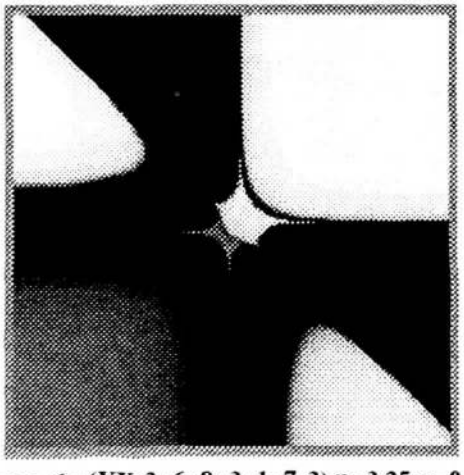

Figure 6 : (YX-3+6+8+3+1+7-3) η=3.25 α=0.00

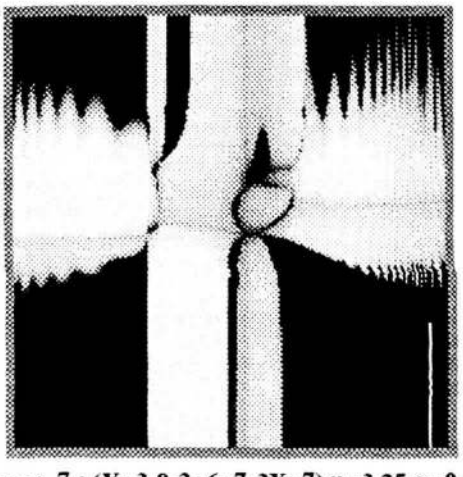

Figure 7 : (Y+3-9-2+6+7-3X+7) η=3.25 α=0.60

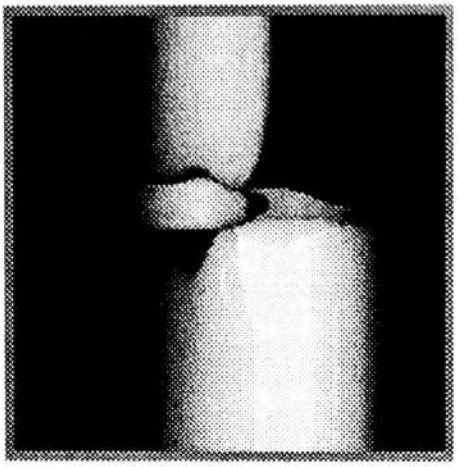

Figure 8 : (-6-4XY-6-6+9-4-9) η=3.00 α=0.50

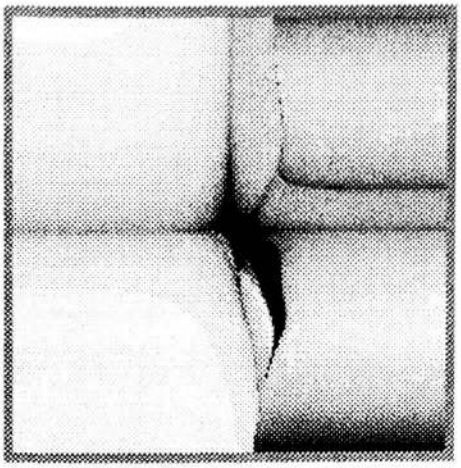

Figure 9 : (-2+1+9-1X-3+8Y-4) η=2.75 α=0.20

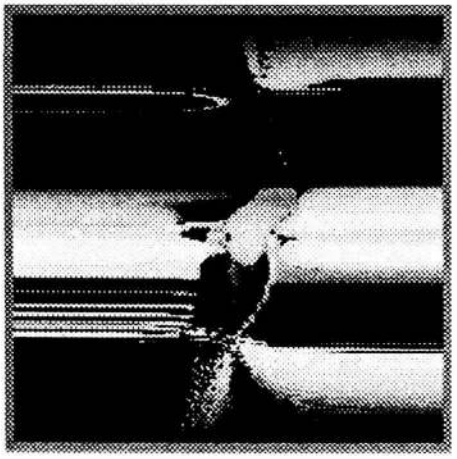

Figure 10 : (+1+8-3-6X-1+1+8Y) η=3.50 α=0.90

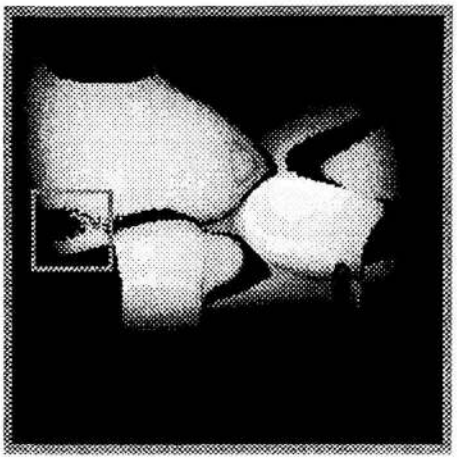

Figure 11 : (+7+4-9-9-5Y-3+9X) η=3.00 α=0.70

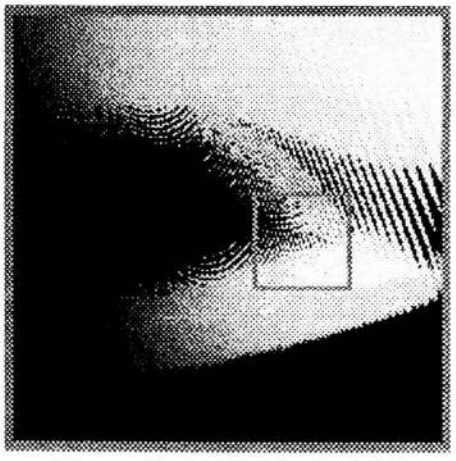

Figure 12 : (-9.0,-1.8) step 0.018

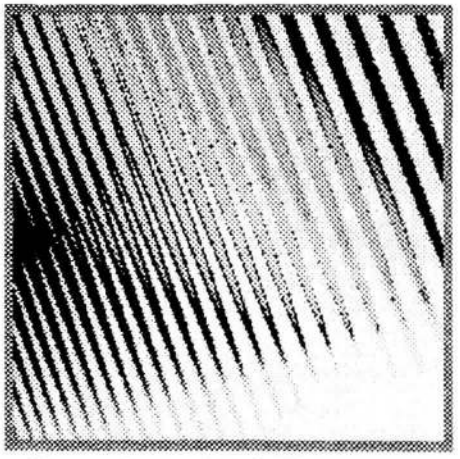

Figure 13 : (-6.966,-0.500) step 0.004

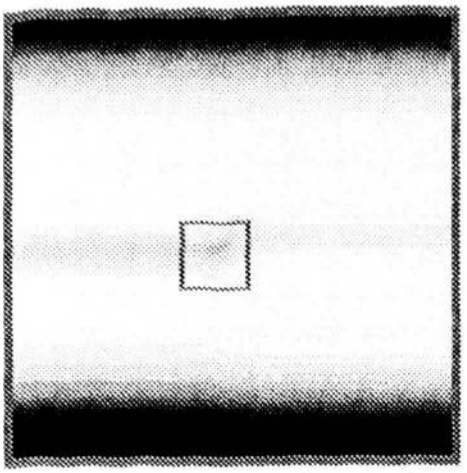

Figure 14 : (-20.00000, -20.00000) step 0.200000

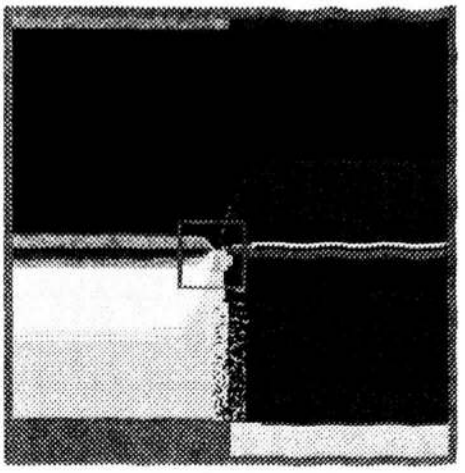

Figure 15 : Solution Networks

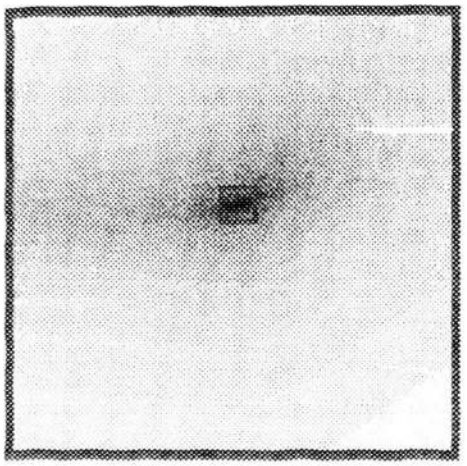

Figure 16 : (-4.500000, -4.500000) step 0.030000

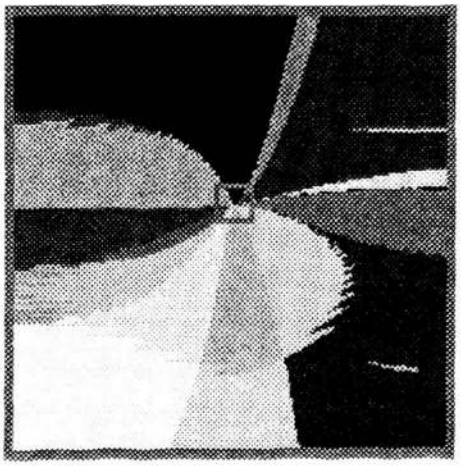

Figure 17 : Solution Networks

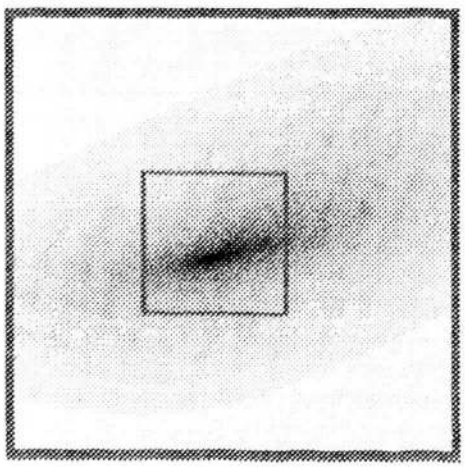

Figure 18 : (-1.680000, -1.350000) step 0.002400

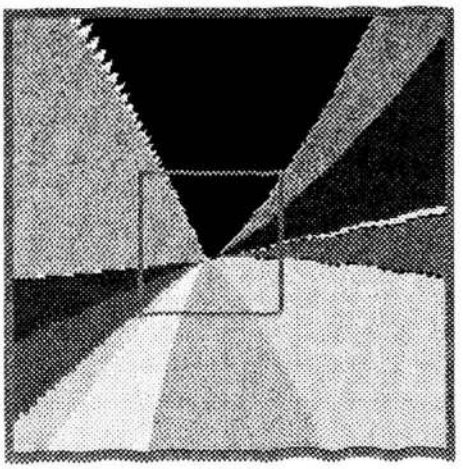

Figure 19 : Solution Networks

Figure 20 : (-1.536000, -1.197000) step 0.000780

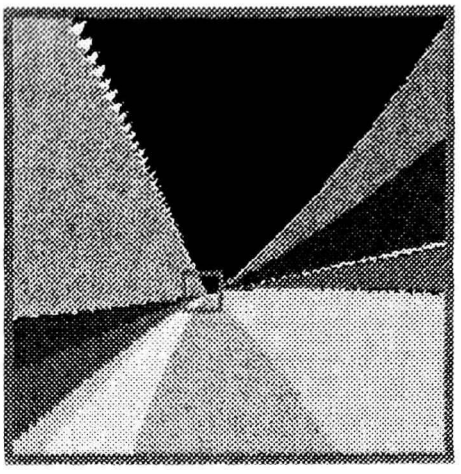

Figure 21 : Solution Networks

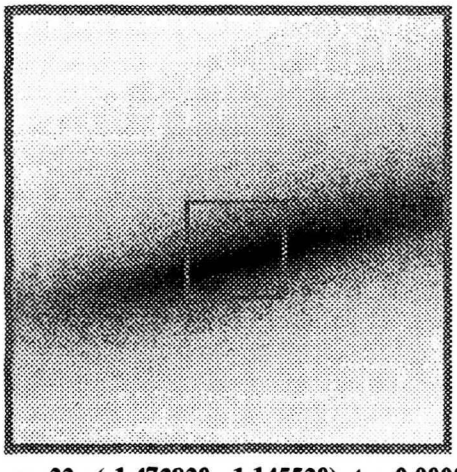

Figure 22 : (-1.472820, -1.145520) step 0.000070

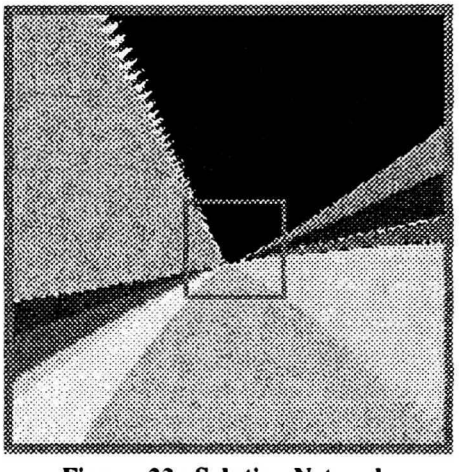

Figure 23 : Solution Networks

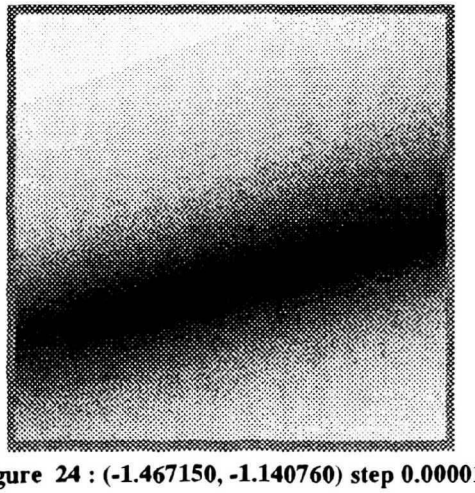

Figure 24 : (-1.467150, -1.140760) step 0.000016

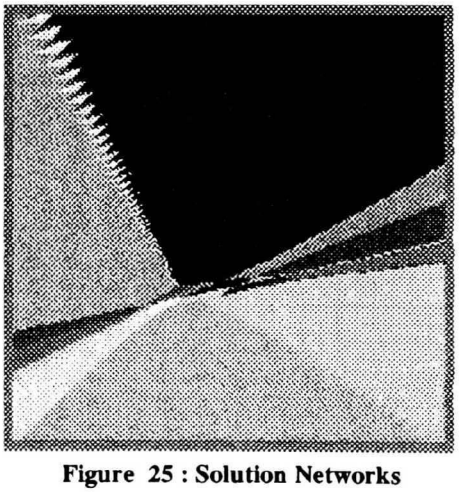

Figure 25 : Solution Networks