Controlling Search Dynamics by Manipulating Energy Landscapes

David S. Touretzky

December, 1989

CMU-CS-89-113 2

School of Computer Science
Carnegie Mellon University
Pittsburgh, PA 15213-3890

Abstract

Touretzky and Hinton's DCPS (Distributed Connectionist Production System) is a neural network with complex dynamical properties. Visualization of the energy landscapes of some of its component modules leads to a better intuitive understanding of the model. Three visualization techniques are used in this paper. Analysis of the way energy landscapes change as modules interact during an annealing search suggests ways in which the search dynamics can be controlled, thereby improving the model's performance on difficult match cases.

This work was supported by the Office of Naval Research under contract number N00014-86-K-0678, and by National Science Foundation grant EET-8716324.

The views and conclusions contained in this document are those of the author, and should not be interpreted as representing the official policies, either expressed or implied, of the Office of Naval Research, the National Science Foundation, or the U.S. Government.

Contents

1 Introduction

Competition through mutual inhibition appears in a wide variety of network designs, such as Rumelhart and McClelland's interactive activation model [McClelland & Rumelhart 1981], Feldman and Ballard's winner-take-all (WTA) networks [Feldman & Ballard 1982], and various competitive learning architectures [Rumelhart & Zipser 1986], [Grossberg 1987], [Kohonen 1984], [Hecht-Nielsen 1987]. In this paper we analyze a system with more complicated competitive dynamics than any of the preceding. The system is DCPS, the Distributed Connectionist Production System [Touretzky & Hinton 1988]. DCPS is a Boltzmann machine composed of five modules, as shown in Figure 1.

>>> Insert Figure 1 here. <<<

Visualization of the energy landscapes of some of these modules leads to a better intuitive understanding of the model. Three visualization techniques are used in this paper. Subsequent analysis of the way energy landscapes change as modules interact during an annealing search suggests ways in which the search dynamics can be controlled, thereby improving the model's performance on difficult match cases. The behavior of DCPS is complex because it contains two independent winner-take-all networks, called rule space and bind space, which interact via their effects on two attentional components called clause spaces. Clause spaces are instances of *pullout networks* [Mozer 1987], another type of competitive architecture which uses mutual inhibition but does not involve WTA dynamics. The clause spaces serve as sources of evidential input to the two WTA nets, but are also controlled by those nets. Thus, unlike most other competitive architectures, in DCPS the evidential input to a WTA net does not remain constant as the network's state evolves. Rather, the present output of the WTA net helps to determine (stochastically) which evidence will become visible to it in the future. This dynamic attentional mechanism is necessary to allow the rule and bind spaces to work together, as explained in section 3.

In the following sections I will first describe the organization of distributed winner-take-all networks and give a general characterization of their energy landscapes. This will be followed by a summary of the way such networks are used in DCPS. Then I will consider in detail the energy landscapes of DCPS' component modules during an annealing search, and use the previously mentioned visualization techniques to explain why the model has trouble solving certain difficult match cases. Two techniques for manipulating energy landscapes, "threshold growing" and "corner clipping," are proposed. Experiments confirm that manipulation of energy landscapes during settling improves the model's performance with no increase in computational demands.

2 Distributed Winner-Take-All Networks

A winner-take-all network uses relaxation to choose the strongest member of a set of competing hypotheses. Each unit in the network receives external excitation in proportion to the evidence supporting its hypothesis. Simultaneously, all-pairs inhibitory connections allow each unit to suppress its rivals in proportion to its own current activation level. The unit with the most support (highest initial activation) receives the least inhibition, and so over time its activation level will increase toward the maximum value while the outputs of rival units fall to zero.

In classical WTA networks a unit's output value is a continuous quantity that reflects its activation level. DCPS implements a stochastic version of winner-take-all dynamics using Boltzmann machines [Hinton & Sejnowski 1986]. Boltzmann units have binary outputs, so there is no direct way for competing units to communicate their activation levels. However, there are indirect ways. The amount of evidence supporting a unit determines its energy gap, ΔE, which in turn determines its probability of being active. The network's degree of confidence in a hypothesis is reflected in the amount of time the unit spends in the active state. Therefore a good approximation to strength of support can be obtained by representing each hypothesis by a clique of k independent units looking at the evidence. The number of active units in a clique reflects the strength of its hypothesis. The architecture just described is called a *distributed* winner-take-all (DWTA) network, because its representation of hypotheses is distributed across multiple units. DCPS uses a clique size $k = 40$.

One of the advantages of DWTAs is that they allow limited fan-in of individual hypothesis units, since the entire space of evidence will still be amply covered by multiple units with overlapping receptive fields. In DCPS, for example, a rule as a whole looks for evidence from 115 units in each clause space, but individual rule units are restricted to looking at randomly-chosen 40 unit subsets of these 115 units.[1] The fact that each member of the clique has a unique receptive field distinguishes this scheme from the 'ensemble' Boltzmann machines of [Derthick & Tebelskis 1988].

The maximum amount of evidence a winning rule should be able to muster in each clause space is about 28 active units. Therefore, a single rule unit should see an average of $(40/115) \times 28$ or 9.7 pieces of evidence per clause. It is important to remember that this is only an expected value; the actual values are binomially distributed about this mean, as shown in Figure 2. Therefore some rule units will see less than the average amount of evidence; this variance has a significant impact on the choice of weights and thresholds in the model. The probability that a rule unit sees exactly r units of evidence when 28 units

[1] By coincidence, rules are implemented as cliques of 40 units, and each rule unit looks at exactly 40 clause units in each clause space. These two numbers are not required to be the same.

are available is:

$$\left(\prod_{i=1}^{r}\frac{28-i}{115-i}\right)\times\left(\prod_{i=r+1}^{40}\frac{115-28-i+1}{115-r-i+1}\right)\times\binom{40}{r} \tag{1}$$

>>> *Insert Figure 2 here.* <<<

One way to smooth out the variation caused by limited-size receptive fields is to introduce excitatory connections between rule units in the same clique. This way, if a particular unit receives below average evidence but most of its siblings are active, its net input will still be high enough for the unit to turn on. This can be illustrated by an "energy tour" diagram such as Figure 3.

>>> *Insert Figure 3 here.* <<<

We use the standard definition for energy from [Hopfield 1982]:

$$E = \sum_i s_i\theta_i - \sum_{i<j} s_i s_j w_{ij} \tag{2}$$

where s_i is the state of the ith unit, θ_i is its threshold, and w_{ij} is the weight on the connection between units i and j. Evidence units are assumed to have zero thresholds, so only their states are important. The energy gap ΔE_i of an individual unit i is the amount by which its input exceeds its threshold:

$$\Delta E_i = \sum_{j\neq i} s_j w_{ji} - \theta_i \tag{3}$$

Consider a clique of 40 units in the Rule Space of DCPS. Each rule unit has a threshold of 69, and receives evidential support from active clause units with a weight of +5 each. Siblings have a weight of +2, and rivals a weight of -2. One such rule unit is shown in Figure 4. Since these are Boltzmann units, the probability p_i that unit i is in the on state is a function of its energy gap and the current temperature T:

$$p_i = \frac{1}{1+e^{-\Delta E_i/T}} \tag{4}$$

>>> *Insert Figure 4 here.* <<<

Rule units receive evidential support from the two clause spaces. If there are 8 clause units active in a rule unit's receptive field, the total evidential support it receives is 40; not enough to put the unit above

its threshold of 69. But recall that sibling rule units (those voting for the same rule) have excitatory connections with a weight of +2. If a unit's siblings are all active, the net input it receives will be $40 + (39 \times 2)$ or 118, putting it well above threshold.

To generate an energy tour diagram such as Figure 3, we start with no rule units active. As rule units are turned on one at a time, the energy rises since the units are below threshold. Boltzmann machines are reluctant to move uphill in energy space, so initially the network will be unlikely to want to turn more units on. But with each new unit forced to turn on, the excitatory effect on sibling units grows. Eventually the network reaches a point where the activation a unit receives from siblings is equal to the difference between the evidence and its threshold. This is the peak in the energy graph. The difference between evidence and threshold is 29 in our example, and the sibling weight is +2, so the peak is reached when 15 units are active. If we turn on one more unit the energy will begin to fall. Once it is over this hump the network is eager to turn more units on. Eventually the energy goes negative, and with all 40 rule units turned on the network is sitting at the bottom of an energy well from which it cannot escape if it is at low temperature. In the second half of an energy tour we reverse direction and turn the rule units off one by one, bringing the system out of its energy well, over the hump, and eventually back to the zero energy state. By ramping the number of active units from zero up to forty and then back down to zero again, we make the graph symmetric, which strengthens the visual impression of the energy wells we're trying to study.

Energy tours in a DWTA take one of four basic shapes. Examples may be seen in Figure 5. (The curves are not to scale.) Let e be the amount of external evidence available to each unit, θ be the unit's threshold, k the clique size, and w_s the excitatory weight between siblings. The four shapes are:

>>> Insert Figure 5 here. <<<

Eager vee: the evidence is above threshold ($e > \theta$). The system is eager to turn units on; energy decreases as the number of active units goes up. We have a broad, deep energy well, which the system will naturally fall into given the chance.

Reluctant vee: the evidence is below threshold, but a little sibling influence (fewer than $k/2$ siblings) is enough to make up the difference. We have $e < \theta < e + w_s(k-1)/2$. The system is initially reluctant to turn units on because that causes the energy to go up, but once siblings take it over the hump it willingly turns on more units. With all units in the clique active the system is in an energy well whose energy is below zero.

Dimpled peak: with higher thresholds the total energy of the system may remain above zero even

when all units are on. This happens when more than half of the siblings must be active to boost each unit above threshold, i.e., $e + w_s(k-1) > \theta > e + w_s(k-1)/2$. The system can still be trapped in the small energy well that remains, but only at low temperatures. It is hard to get into this well in the first place since to do so the system must travel way uphill in energy space. It can do better by simply turning all its units off. Even if it visits the well, the system may easily bounce out if the well is shallow.

Smooth peak: when $\theta > e + w_s(k-1)$, units will be below threshold even with full sibling support. In this case there is no energy well, only a peak. The system wants to turn all units off.

In the next section we discuss in detail the task and the structure of the DCPS model. We will then return to the discussion of energy landscapes and their relevance to DCPS.

3 The Task of DCPS

DCPS matches production rules against a working memory. The working memory contains triples of symbols, such as **(P A B)**, in a coarse-coded, distributed representation. We use 25 letters, so there are $25^3 = 15.625$ possible triples. The model has 2000 working memory units, each of which has a $6 \times 6 \times 6$ receptive field table, so each unit participates in the representation of 216 distinct triples. Conversely, each triple is represented by $(6/25)^3 \times 2000$ or approximately 28 units. See [Rosenfeld & Touretzky 1988] for a full discussion of the mathematics of this representation, called a *coarse-coded symbol memory*.

To store a triple in working memory we turn on the 28 units that code for it, forming a distributed pattern of activation. Several triples may be stored simultaneously by superimposing their patterns. Working memory typically contains from two to a dozen triples at a time.

Each DCPS production rule consists of a "left hand side" condition and some "right hand side" actions. Only the left hand side condition will concern us here. Conditions are pairs of clauses, where each clause specifies a triple. A single variable =x appears in the first position of each clause. This forces the two clauses to match triples that both begin with the same letter. A typical set of production rules is:

```
Rule-1:   (=x A B)   (=x C D)   -->   ...

Rule-2:   (=x E F)   (=x G H)   -->   ...

Rule-3:   (=x I J)   (=x K L)   -->   ...
```

Rule-1 can match the triples **(P A B)** and **(P C D)**, or the triples **(Q A B)** and **(Q C D)**, but it cannot match **(P A B)** and **(Q C D)** since the variable binding constraint would be violated. The same constraint applies to all rules.

The job of selecting a production rule that matches the current working memory contents is shared by the rule and bind spaces, which are both DWTA networks. The rule units only examine the last two positions of a triple to see if they match the constant part of the rule. Each clause has an evidence pool of 115 units to examine.[2] If a triple is present that matches a particular left hand side clause, there will be approximately 28 units of evidence visible in that clause space.

At the same time as the rule network is searching for a winning rule, the bind network is searching for a value for the bound variable. It does this by looking for two working memory triples that both begin with the same letter.

Consider what happens during a match when working memory contains the following triples. (I have grouped these six triples in pairs to neatly correspond to the rules they trigger, but in general this isn't possible because a triple may match any number of rules, and a rule may partially match any number of triples.)

```
(P A B)      (Q C D)

(Q E F)      (P G H)

(R I J)      (R K L)
```

The only correct match in this case is Rule-3 with the variable **=x** bound to R. But the two individual DWTA nets do not have this global perspective. Considering rule space in isolation for a moment, it is possible to satisfy Rule-1 because there is a triple of form **(– A B)** and a triple of form **(– C D)**. (One of these triples begins with P and the other with Q, but rule units don't notice that.) It is also possible to satisfy Rule-2, since there is a triple of form **(– E F)** and a triple of form **(– G H)**. And Rule-3 can be satisfied because there is a triple of form **(– I J)** and a triple of form **(– K L)**. So, in the absence of a variable binding constraint, all the rules appear to have equal support.

[2]There are 25 possible triples that could match a pattern such as **(=x A B)**, and each triple is represented by 28 units, so one might expect to have to examine $25 \times 28 = 700$ units. But due to the use of coarse coding, similar triples have overlapping representations, so the total number of units that must be examined is only $(6/25)^2 \times 2000$ or 115.

Similarly, if we ignore the rules for a moment, we see that there are also three equally well-supported winners in bind space. It is possible to completely satisfy the P bind units because there are two triples of form (P – –); likewise the Q and R bind units. Yet we know that the only totally correct match is the one with R and Rule-3 as the winners.

The reason it's difficult to get the network to find the correct match is that rule and bind spaces are selecting triples independently from working memory. The constraints that each space imposes on the match are not being communicated to the other space.

This problem is solved in DCPS by introducing an attentional mechanism, known as clause spaces, to mediate between the DWTA nets and working memory. Rule and bind units are not allowed to look directly at the contents of working memory; they look only at the two clause spaces. Each clause space contains 2000 units with one-one excitatory connections from corresponding working memory units. Due to strong lateral inhibition, each clause space can only represent a small subset of the working memory pattern at any one time. The inhibition level within a clause space is carefully set so that by the completion of a match only about 28 clause units will remain active: just enough to represent one triple. But which triple should they represent? Clause spaces receive top-down guidance via their connections from the rule and bind spaces. They use this input to focus on subsets of the working memory pattern that both the rule and bind spaces appear to be interested in. In order to conclude a successful match, rule and bind spaces are forced to agree on which two triples from working memory they wish to see in the clause spaces. In the sample problem we've been considering, only the selection of (R I J) in the first clause space and (R K L) in the second can satisfy both a rule hypothesis and a variable binding hypothesis simultaneously.

4 Dynamics of DCPS

When DCPS begins a rule matching cycle, it starts out at a high temperature, putting its units in essentially random states. The clause units extract a random subset of the working memory pattern, and the rule and bind units, although influenced somewhat by the clause units, are also nearly random in their behavior. After two cycles at a high temperature of 300, the temperature is dropped to a moderate value of 33. This is where the real matching work takes place. The model performs a stochastic search through the space of possible rule matches. Coalitions of rule, bind, and clause units form, then break up and later reform again. If one particular rule starts to become more active, more of its units will turn on; this provides a positive influence to clause units that support that rule. If the clause units can also achieve support from bind units (because there is a triple in the other clause space that begins with the same letter), the clause

units will be more likely to remain on, which in turn helps even more rule units in the winning rule's clique to become active.

It is important to emphasize that the model does not consider individual rules, variable values, or triples discretely. During the early phases of the match the clause spaces will typically have 40 to 60 units active, representing a mixture of several different triples. At the same time, there will be several rules partially active in rule space, and several variable binding hypotheses active in bind space. As the winning rule and bind cliques grow stronger, they eventually dominate their respective spaces and suppress all rivals, but this is a complex process due to the two-way interaction with the clause spaces.

4.1 Dynamics with Full Evidence

The model's dynamics are easiest to understand when full evidence is available, in the late phase of the match, so let's start by examining the energy landscape of rule space when there is ample evidence in the clause spaces for the winning rule. Assume there are three rules, A, B, and C, with disjoint evidence populations. The state space of the rule module can be drawn as a cube 40 units on a side. The energy tour through state space shown in Figure 6 does not include the interior of the cube, but does extend along each of the major axes.

>>> Insert Figure 6 here. <<<

Suppose that B is the best-supported rule, with evidence 100 (10 connections from active units in each clause space, with weights of +5, times 2 clauses), and that A has evidence 40 and C has evidence 5. Under these conditions, the energy tour looks like Figure 7.

>>> Insert Figure 7 here. <<<

The left half of Figure 7 shows the energy curves for hypotheses A, B, and C, when a value of 69 is used for for the rule unit thresholds.[3] Since the evidence for rule A is a bit below threshold, its curve is of the "reluctant vee" type. We see some initial reluctance to turn on A units, indicated by a rise in energy, but this is eventually overcome by sibling support. The evidence for rule B is well above threshold, so there is no reluctance to turn on B units. The energy goes negative as soon as the first B unit comes on. Rule C has almost no evidential support; its tiny local energy minimum is due almost entirely to sibling support.

[3] All the weights and thresholds in this paper are actual DCPS values, taken from [Touretzky & Hinton 1988].

Two important observations can be made from these curves. First, the energy well for B is considerably deeper than for A. This means at moderate temperature the model could pop out of A's energy well, but it is more likely to remain in B's well. The well for C is so shallow, and its energy is so far above zero, that it is very unstable; the model is certain to pop out of it at any temperature above zero.

The second observation we can make is that the well for B is somewhat *broader* than the well for A. This means that it is easier for the B attractor to capture the model, since its attractor region spans a larger portion of state space.

The tours we've made for hypotheses A, B, and C correspond to traversing the three orthogonal edges extending from the origin of a $40 \times 40 \times 40$ cube, as shown in Figure 6. During the stochastic search, A, B, and C units will be flickering on and off simultaneously, so the model will also visit internal points of the cube not covered in the energy tour diagram. Therefore we will experiment with some additional graphic representations of energy landscapes. First, note that hypothesis C gets so little support that we safely can ignore it and concentrate on A and B. This allows us to focus on just the front face of the state space cube in Figure 8. In this graph the number of active A units runs from zero to forty along the vertical axis, and the number of active B units runs from zero to forty along the horizontal axis. The arrows at each point on the graph show legal state transitions at zero temperature. For example, at the point where there are are 38 active B units and 3 active A units there are two arrows, pointing down and to the right. This means there are two states the model could enter next: It could either turn off one of the active A units, or turn on one more B unit, respectively. At nonzero temperatures other state transitions are possible, corresponding to uphill moves in energy space, but these two remain the most probable.

>>> *Insert Figure 8 here.* <<<

The points in the upper left and lower right corners of Figure 8 are marked by "Y" shapes. These represent point attractors at the bottoms of energy wells; the model will not move out of these states unless the temperature is greater than zero.

A point in state space is said to be within the region of a particular fixedpoint attractor if all legal transition sequences (at $T = 0$) from that point lead eventually to the fixedpoint. By looking at the arrows, the attractor regions of A and B are easily determined. They are outlined in the figure. Note that the attractor region for B covers more area than the one for A, as predicted by its greater breadth in the energy tour diagram. Note also that there is a small ridge between the two attractor regions. From starting points on the ridge the model can end up in either final state.

>>> *Insert Figure 9 here.* <<<

While the legal state transition diagram shows us the breadth of various attractor regions, Figure 9 shows us the depth. The energy well for B is substantially deeper than the well for A. Starting at the point in the lower left corner where there are zero A units and zero B units active, the energy falls off immediately when moving in the B direction, but rises initially in the A direction before dropping into a modest energy well when most of the A units are on. Points in the interior of the diagram, representing a combination of A and B units active, have higher energy than points along the edges due to the inhibitory connections between units in rival cliques.

Although DCPS is a Boltzmann machine it does not operate by simulated annealing in the usual sense. True annealing requires a slow reduction in temperature over many update cycles so the model can thoroughly search the state space and with high probability end up in the minimum energy state. The stochastic search in DCPS takes place at a single temperature which has been determined empirically to be the model's approximate melting point. (The melting point is the temperature at which the model repeatedly enters shallow energy wells and then escapes. At lower temperatures escape is too difficult; at higher temperatures the model will not remain trapped if it enters a deep well.) The entire search takes only a few update cycles; typically less than 10. Therefore, the breadth of energy wells is quite important, as it determines how likely the model is to wander into particular attractor regions. Once inside an attractor region, by definition it can reach the bottom by moving strictly downhill.

We can see from Figures 8 and 9 that the attractor for A, although smaller and narrower than the one for B, is still sizable. This is likely to mislead the model, so that some of the time it will get trapped in the wrong energy minimum. The fact that there is an attractor for A at all is due largely to sibling support, since the raw evidence for A is less than the rule unit threshold.

We can eliminate the unwanted energy well for A by raising the rule unit thresholds to a level that exceeds maximum sibling support. DCPS uses a value of 119. The energy tours for A, B, and C with high thresholds are shown in the right half of Figure 7. With high thresholds, A is no longer an energy well. If we turn on all 40 A units and let the model run, it will turn all of them off again, even at zero temperature. B is still an energy well, but it is narrower and shallower than before. Also, there is a slight reluctance to turn B units on initially, since full evidential support by itself is no longer enough to put a unit above threshold. Once a few B units become active, though, the model quickly ends up in the attractor region, where it turns on the rest and settles into a global energy minimum.

>>> Insert Figure 10 here. <<<

Figure 10 shows the legal zero-temperature state transitions with the threshold set at 119. We see that the area of the B attractor is reduced somewhat, but it still covers a sizable portion of state space. The

attractor for A is gone. There is, however, another attractor, located at the origin where all units are off. The null hypothesis has become an attractor because the threshold is now higher than the maximum external evidence, so when very few units are active there is insufficient sibling support for any rule to win the competition. The only way the model can move downhill in energy space from this point is by turning all its units off.

>>> Insert Figure 11 here. <<<

The majority of points in state space do not belong uniquely to either of the two attractors, so in theory the model can reach either one from those points. However, the two choices are not equiprobable. Units follow the Boltzmann probability distribution, so they are more likely to move in the direction of steepest descent in energy. As Figure 11 shows, the energy gradient favors attractor B over the null attractor. In a later section I will introduce a technique for actively avoiding the null attractor.

4.2 Dynamics with Low to Moderate Evidence

There are two reasons why a rule unit may be receiving only low to moderate evidential support from clause space. The unit may simply have chosen an unlucky 40-clause unit subset for its receptive field, in which instead of the expected 9-10 active clause units it will see only, say, 6. Sibling support can make up for this deficit once enough siblings have become active.

But all rule units receive low evidential support initially, for a different reason. Early in the match the clause units are extracting an essentially random subset of the working memory pattern. There has not been time yet for the rule and bind units to form coalitions and focus the clause spaces on particular triples. Therefore, rule units will unavoidably go through a period where they are getting much less evidence from the clause spaces than they will eventually end up with after the match has concluded successfully. At the same time, early in the match, the rule units are receiving little sibling support, and quite a bit of inhibition from competing rules. Thus they are quite likely to be near or below threshold.

>>> Insert Figure 12 here. <<<

The right half of Figure 12 shows an energy tour of the landscape when hypothesis B has evidence 60 rather than 100, and rule thresholds are set at 119. The attractor for B has energy way above zero. This means it is very unlikely for the model to travel so far uphill in energy space to reach that local energy minimum. And at a temperature of 33, it is unlikely to stay in that minimum if it does reach it.

>>> Insert Figure 13 here. <<<

Figure 13 shows that the attractor for B is very narrow, indicating that it is heavily dependent on sibling support. The attractor for the null hypothesis covers most of the state space. We can see from this diagram that if we start out with high thresholds early in the search, before the clause units have time to focus their attention and bring in adequate evidence, the model will simply turn all its units off and give up. Figure 14 shows that giving up is in fact the best thing the model can do. The null hypothesis is the global energy minimum (zero); the B attractor's energy is quite a bit higher than that.

>>> *Insert Figure 14 here.* <<<

Clearly we want thresholds to be low so rule units can turn on early, and thereby invite more evidence to appear in clause space. Let's go back to using thresholds of 69 and see what effect that has on energy space. The energy tours with a threshold of 69 are shown in the left half of Figure 12; the legal state transitions at zero temperature are shown in Figure 15, and the energy space is mapped in Figure 16.

>>> *Insert Figure 15 here.* <<<

>>> *Insert Figure 16 here.* <<<

With low thresholds the attractor for the null hypothesis is negligible, while the attractor for B is of a healthy size. But we saw earlier that we wanted thresholds to be high to rule out spurious minima, such as the one for A which has reappeared. This demonstrates a fundamental tension between different needs of the model with respect to thresholds in DWTA modules:

- We want thresholds to be low so units can become active with just a little evidence visible.
- We want thresholds to be high so that only well-supported hypotheses can be winners.
- We want thresholds to be low so that siblings can help turn a unit on if it receives below-average evidential support. This is important for counteracting the variance in evidence distribution illustrated in Figure 2, and also for supporting tentative hypotheses while they assemble more evidence in the clause spaces.
- We want thresholds to be high to prevent the "mass delusion" effect, in which the model is captured in a hypothesis' energy well just because of high sibling support, not external evidence.

In section 6 I suggest two techniques for resolving these conflicts.

5 Coarse-Coded DWTA Networks

The bind space in DCPS is similar to rule space: it is organized as a DWTA, it gets its input from the two clause spaces, and it in turn affects the clause spaces. It contains 25 cliques, one for each of the letters A through Y, and each clique has 40 members, the same as for rule space. But bind space uses a coarse coded representation in which each bind unit codes for three values instead of one. So instead of $25 \times 40 = 1000$ bind units, there are only 333.

The coarse-coded WTA has two advantages. The first is a savings in units without any reduction in the number of units per clique. The second advantage is that, with a coarse-coded WTA, every hypothesis is likely to have a few units active, even when one hypothesis dominates the competition. To take the most extreme example, suppose hypothesis P has won the competition outright, and so there are exactly 40 units active, all of which code for P. Since each unit actually codes for three letters, in addition to there being 40 P votes affecting the clause spaces, there are also an average of $(40 \times 2)/24 = 3.3$ votes for each of the other 24 letters. Earlier in the search, before a clear winner has emerged, the bind units are even more likely to spread the votes around. This seems to help promising hypotheses over the initial hump of a "reluctant vee" energy landscape.

In a coarse-coded WTA, two units will have an excitatory connection if their receptive fields have at least one letter in common. If their receptive fields are disjoint they have an inhibitory connection. Given that each unit codes for three letters, and each letter has a clique of 40 units, each bind unit ends up with slightly less than 3×39 siblings, which is about one third of the entire bind space. But this does not cause the network to choose three winners. If unit x is in the same clique as y, and unit y is in the same clique as z, x and z may still be rivals if they have no letters in common. So there is plenty of inhibition available, and the attractor states of the network are still those in which only the units in a single clique are active.

Since bind units also have limited-size receptive fields (they look at 240 clause units out of a relevant population of 1440), they are subject to the same variance in the distribution of evidence as rule units. But bind units also have a second source of variance, because the three-letter receptive fields of these units are all different. Units within the clique for the letter P, for example, will experience varying levels of inhibition from non-P units, since some will view these units as rivals while others view them as siblings because they have some other letter in common.

6 Threshold Growing and Corner Clipping

Due to the interaction between rule and clause spaces during the stochastic search, the model goes through a phase transition from a low evidence to a high evidence phase. We need low thresholds in the low evidence phase. They should be high enough to weed out totally random hypotheses, yet not so high as to pose a barrier to hypotheses with the potential to develop real support. But we do need high thresholds once the model has begun to assemble evidence and form coalitions, because we only want strong hypotheses to have below-zero energy wells.

One way to solve this problem is to let the thresholds grow during the simulated annealing process. By growing the thresholds we are dynamically re-shaping the energy landscape, switching from the left half of Figure 12 to the right half of Figure 7, or from Figure 15 to Figure 10. The energy well for B remains because evidence for B is increasing at the same time as the thresholds increase. The energy well for A disappears because the total evidence available for A is insufficient to maintain it as the match criteria become more stringent. In other words, by raising the thresholds we are "pulling the rug out from under" the A attractor while leaving the model time to head toward the B attractor.

Another technique, *corner clipping*, might be used to prevent the model from selecting the null hypothesis and turning all its units off. At the conclusion of any successful match there will be 40 rule units active, all belonging to the same clique. Therefore rule space can be restricted to always have at least R rule units active, where R is some parameter close to 40. Whenever there are fewer than R rule units active, the model could be forced to turn on some more units at random. In the final step of the annealing this restriction would be removed. If there were a valid match, there would be 40 units active. If the model failed to find a valid match, it would be allowed to turn all its units off; this would be a signal that the match had been unsuccessful and should be restarted from the beginning.

>>> *Insert Figure 17 here.* <<<

This technique is called "corner clipping" because establishing a requirement for a minimum number of active rule units amounts to clipping the origin corner from the state space hypercube, as shown in Figure 17. The model is never permitted to enter this corner; therefore it can't turn all its units off.

>>> *Insert Figure 18 here.* <<<

A possible improvement on this idea is to make the boundary reactive, like the bumper on a pinball machine:[4] Whenever the model has turned off one too many units, we kick it away from the boundary by

[4]This analogy was suggested by Lokendra Shastri.

turning on several extra units at random. Over time the model should eventually find the attractor region for B, even if that region is small. Figure 18 shows that the gradient of the energy surface favors the B attractor even outside of what we defined as the strict attractor region, because the barrier prevents the model from choosing the null hypothesis and the high threshold removes A as an energy well.

7 Experimental Results

The original version of DCPS used the parameters shown in Table 1 and the annealing schedule of Table 2. The model ran the match at a temperature of 33, with low thresholds, until the energy dropped below a certain value. It then dropped the temperature to 0.1 to settle the two DWTA modules and remove noise from the two clause spaces. In a final verification step called "rebiasing," rule and clause unit thresholds were increased by 50, and bind unit thresholds by 30.

>>> Insert Table 1 here. <<<

>>> Insert Table 2 here. <<<

The post-match threshold-raising was used to test whether the model was trapped in a deep energy well representing a valid match, or lying in a local minimum representing only a partial match. Rebiasing changed the shape of the energy landscape as shown in Figure 19. If the model was trapped in the shallow well of a local minimum, raising the thresholds left it on a slope; it then "rolled downhill" by turning all its units off. If the model was in a global minimum, rebiasing left it trapped in a well that had positive energy but very steep sides, so it remained trapped in that state.

>>> *Insert Figure 19 here.* <<<

This scheme worked well for simple match problems, but it did not do well on complicated problems in which there were many partial matches, hence too many confusing local minima. The example given at the beginning of this paper is one of those hard match problems. With six triples in working memory, there are nine possible pairs of triples the clause spaces might reasonably extract.[5] One pair yields the correct match. But four other pairs have moderately deep energy minima corresponding to partial matches:

1. **(P A B)** and **(Q C D)**, choosing Rule-1 as the rule and a combination of P and Q as the variable value. This violates the unique variable binding constraint.

[5]The clause spaces don't actually extract triples as whole objects, though. Each of the 2000 units operates independently, so clause space usually contains a mixture of partial representations for several triples at once.

2. **(P A B)** and **(P G H)**, choosing a combination of Rule-1 and Rule-2 as the winning rule, and a variable value of P. This violates the constraint that there must be a unique winner in Rule space. Each rule gets support from only one of the two clause spaces.

3. **(Q E F)** and **(P G H)**, choosing Rule-2 and a combination of P and Q, violating the unique variable binding constraint.

4. **(Q E F)** and **(Q C D)**, choosing a combination of Rule-2 and Rule-1, with Q as the variable value, violating the unique rule constraint.

When the model failed to match correctly, frequently this was because it ended up trapped in one of the partial minima, and after rebiasing, turned all its units off. One reason for this failure may be that the partial minima interacted synergistically, e.g., Rule-1 and Rule-2 both support the same two bindings (P and Q), and conversely each of those bindings partially supports both Rule-1 and Rule-2. The correct winners, Rule-3 and R, have nothing in common with any other match candidate, so there is no synergy for them to capitalize on.

The model did solve the match problem correctly some of the time, but its success rate was rather low, averaging 32% over five experiments. In each experiment the model was run overnight for at least 400 annealings. The five experiments differed only in the wiring pattern between the rule and bind spaces and the two clause spaces. Units in rule and bind space are wired to randomly-chosen subsets of the relevant clause space populations each time the model is initialized. Different wiring patterns result in slightly different success rates for any particular match problem.

A modified version of DCPS was constructed to see how threshold growing and corner clipping could improve the search dynamics. Initially, the corner clipping parameter was set to keep a minimum of 40 units active in each WTA space, and threshold growing was used to gradually raise the rule, bind, and clause unit thresholds from their initial values to somewhat higher values. The other model parameters (i.e., connection strengths and annealing temperatures) were unchanged. A run of five match experiments gave an average performance level of 35%, which is not a statistically significant improvement.

Observing the model in light of the analysis of section 4 suggested that changes to some of its other parameters might make it possible to take advantage of threshold growing. Consequently, the model was re-tuned, resulting in different choices for rule unit thresholds, lateral excitation and inhibition levels, and annealing temperatures, as shown in Table 3. The revised annealing schedule is shown in Table 4. The result was an average performance in five experiments of 45%, a clear improvement over the 32% the original model achieved.

Corner clipping was inapplicable in the new model because, with the lower thresholds, there were always more than 40 rule and bind units active. Another series of experiments was performed with high initial thresholds, relying on corner clipping to keep the model from turning all its units off before adequate evidence could be assembled in the clause spaces to support some hypothesis. The model did poorly under these conditions because it was forced to turn on lots of units for unsupported hypotheses (in order to have at least 40 units on). Inhibitory connections between units in rival cliques therefore caused all the WTA units to receive substantial inhibition, so the model never settled into an attractor for any hypothesis. In another experiment, instead of turning on units at random to keep the total number of active WTA units above 40, the model was modified to turn on the inactive unit with the highest energy gap. This turned out to be worse than the random strategy, because once a WTA net got itself into a partial minimum, it could not easily get out again by turning units off. Every time the number of active units went below 40, the unit with the highest energy gap would be one of the members of the clique the model was trying to abandon, so it would be forced back into the weak attractor it was trying to leave.

Apparently, for corner clipping to be useful in high dimensional spaces one must push away from the boundary in a random direction, and omit the inhibitory connections between units in rival WTA cliques. Removing lateral inhibition in WTA space should not eliminate its competitive dynamics, because lateral inhibition in the clause spaces would prevent multiple cliques from simultaneously assembling enough evidence for full support. This hypothesis has not been pursued.

8 Discussion

Tuning DCPS is a painstaking task, but one that would have been impossible without visualization tools. The first visualization technique used was the basic state display. The model contains more than 6500 units, all of whose states can be observed simultaneously throughout the stochastic search. Monitoring the units' collective states and the model's energy level over time is an easy way to recognize when it has undergone a phase transition or bounced out of a local minimum. But just watching these parameters is not enough to fine tune the weights and thresholds that govern the model's complex dynamic behavior. More sophisticated visualization tools, such as energy tour diagrams and legal state transition diagrams, are necessary.

The three types of diagram introduced in this paper complement each other. Energy tour diagrams help us understand the nature of particular fixedpoints: are their attractor regions easy to enter, or is there some energy barrier? How high is the barrier? How deep is the energy well on the other side?

Legal state transition diagrams serve as a map of state space, showing the size of attractor regions and

the places in state space where competing attractors interact. Energy surface diagrams show the depth as well as the breadth of attractor regions, and the gradient allows us to judge the probabilities of particular state transitions. Studying these diagrams inspired the threshold growing and corner clipping techniques described in this paper.

It should be stressed that the stochastic search DCPS performs is not truly annealing, since all the matching work is done at a single temperature and the model settles extremely quickly. The matching portion of the "annealing" schedule requires just ten cycles. Surely the model's performance could be improved by adopting a much longer annealing period with a gradual temperature reduction, but since rapid settling is also important in DCPS, optimal performance on the trickiest match cases was sacrificed for efficiency. The retuning and introduction of threshold growing resulted in improved performance without lengthening the annealing period or in any way increasing the model's computational workload.

Another important observation about this work, due to Ralph Linsker (personal communication), is that it shows that an annealing process can involve more parameters than just temperature. Threshold growing dynamically changes the shape of the energy landscape. In other words, the model shifts from one annealing problem to another as its thresholds change. Modifying the energy landscape in this way should give better control of the search than would varying temperature alone.

Acknowledgments

This research was supported by the Office of naval research under contract N00014-86-K-0678, and by National Science Foundation Grant EET-8716324. I thank Dean Pomerleau, Roni Rosenfeld, Paul Gleichauf, and Lokendra Shastri for helpful comments, and Geoffrey Hinton for his collaboration in the development of DCPS.

References

[Derthick & Tebelskis 1988] Derthick, M. A., and Tebelskis, J. (1988) 'Ensemble' Boltzmann machines have collective computational properties like those of Hopfield and Tank Neurons. In D. Z. Anderson (ed.), *Neural Information Processing Systems*. New York: American Institute of Physics.

[Feldman & Ballard 1982] Feldman, J. A., and Ballard, D. H. (1982) Connectionist models and their properties. *Cognitive Science* **6**:205-254.

[Grossberg 1987] Grossberg, S. (1987) Competitive learning: from interactive activation to adaptive resonance. *Cognitive Science* **11**(1):23-63.

[Hecht-Nielsen 1987] Hecht-Nielsen, R. (1987) Counter-propagation networks. *Applied Optics* **26**(23):4979-4984.

[Hinton & Sejnowski 1986] Hinton, G. E., and Sejnowski, T. J. (1986) Learning and relearning in Boltzmann machines. In D. E. Rumelhart and J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. Cambridge, MA: Bradford Books/The MIT Press.

[Hopfield 1982] Hopfield, J. J. (1982) Neural networks and physical systems with emergent collective computational abilities. *Proceedings of the National Academy of Sciences USA*, **79**:2554-2558.

[Kohonen 1984] Kohonen, T. (1984) *Self-Organization and Associative Memory*. Berlin: Springer-Verlag.

[McClelland & Rumelhart 1981] McClelland, J. L., and Rumelhart, D. E. (1981) An interactive activation model of context effects in letter perception: part 1. An account of basic findings. *Psychological Review* **88**:375-407.

[Mozer 1987] Mozer, M. C. *The perception of multiple objects: a parallel, distributed processing approach*. Doctoral dissertation, University of California at San Diego.

[Rosenfeld & Touretzky 1988] Rosenfeld, R., and Touretzky, D. S. (1988) Coarse-coded symbol memories and their properties. *Complex Systems*, **2**(4):463-484.

[Rumelhart & Zipser 1986] Rumelhart, D. E., and Zipser, D. (1986) Feature discovery by competitive learning. In D. E. Rumelhart and J. L. McClelland (eds.), *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume 1. Cambridge, MA: Bradford Books/The MIT Press.

[Touretzky 1989] Touretzky, D. S. (in press) Connectionism and compositional semantics. In J. A. Barnden and J. B. Pollack (eds.) *Advances in Connectionist and Neural Computational Theory.* Norwood, NJ: Ablex. Also available as technical report CMU-CS-89-147, School of Computer Science, Carnegie Mellon University.

[Touretzky & Hinton 1988] Touretzky, D. S., and Hinton, G. E. (1988) A distributed connectionist production system. *Cognitive Science* **12**(3):423-466.

Figure Captions

Fig. 1. The five modules comprising DCPS, the Distributed Connectionist Production System.

Fig. 2. Percent of DCPS rule units with a given number of active clause units in their receptive fields (out of 40 clause units total), at the completion of a successful match. About 18% of rule units will have 10 active clause units; 17% will have 9 units; 13.5% will have 8 units; 15% will have 7 or fewer units. This curve was generated from Equation 1.

Fig. 3. A tour through the energy landscape of a DWTA with one clique of 40 units. Moving along the x axis, the number of active units goes from zero up to 40 and then back down to zero. The y axis is energy. The cusp in the middle of the curve is due to the abrupt switch from turning units on to turning them off.

Fig. 4. One rule unit in a clique of 40.

Fig. 5. The four basic shapes of an energy tour diagram.

Fig. 6. An energy tour that follows the axes of the state space cube. This tour considers only one rule unit clique as active at a time, so the internal points of the cube are not visited.

Fig. 7. Energy tours for three rule cliques with varying amounts of evidence. Left side: low rule unit thresholds; right side: high thresholds.

Fig. 8. Legal state transitions at zero temperature, with low thresholds.

Fig. 9. Energy of the model with various combinations of A and B units active. Attractor regions are outlined.

Fig. 10. Legal transitions at zero temperature when rule unit thresholds are set at 119, and the evidence for A is 40 and for B 100.

Fig. 11. The energy surface for rule units with thresholds set at 119, where evidence for A is 40 and B 100.

Fig. 12. Energy tours when the best-supported rule clique has evidence 60 instead of 100. Left: low

thresholds; right: high thresholds.

Fig. 13. Legal state transitions when the best-supported clique has evidence 60, and thresholds are high.

Fig. 14. Energy surface when the best-supported clique has evidence 60, and thresholds are high.

Fig. 15. Legal state transitions when the best-supported clique has evidence 60, and thresholds are low.

Fig. 16. Energy surface when the best-supported clique has evidence 60, and thresholds are low.

Fig. 17. Clipping the origin corner from the state space cube forces the model to keep at least R units active at all times. This prevents it from turning all its units off.

Fig. 18. Energy surface with corner clipping. The pinball bumper effect causes the model to rebound off the barrier by turning on extra units at random.

Fig. 19. "Rebiasing" (raising thresholds) changes the shape of the energy landscape.

Table Captions

Table 1. Weight and threshold values used in the original DCPS model.

Table 2. Annealing schedule from the original DCPS model.

Table 3. Weight and threshold values for the modified version of DCPS incorporating threshold growing.

Table 4. Modified annealing schedule incorporating threshold growing.

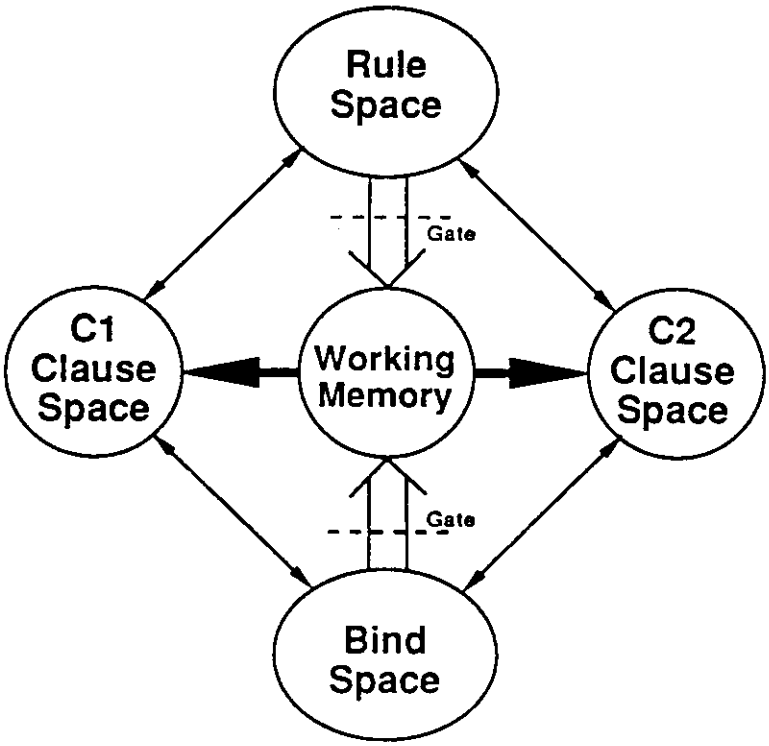

Rule Space
Gate
C1 Clause Space
Working Memory
C2 Clause Space
Gate
Bind Space

Figure 1:

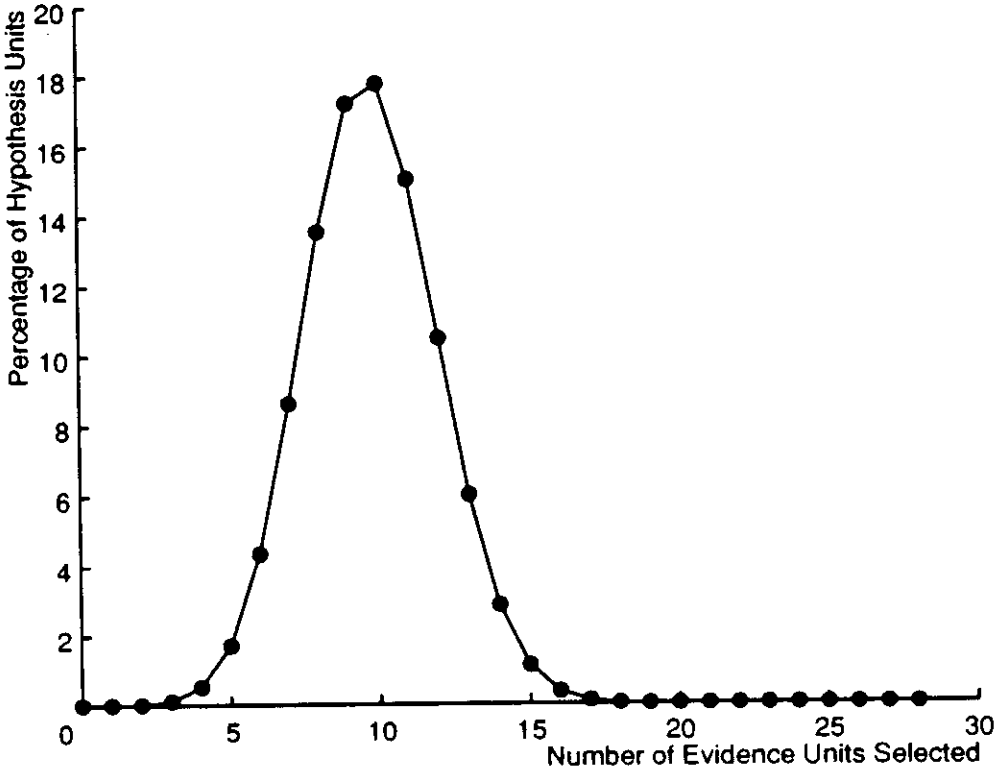

Percentage of Hypothesis Units
20
18
16
14
12
10
8
6
4
2
0
5
10
15
20
25
30
Number of Evidence Units Selected

Figure 2:

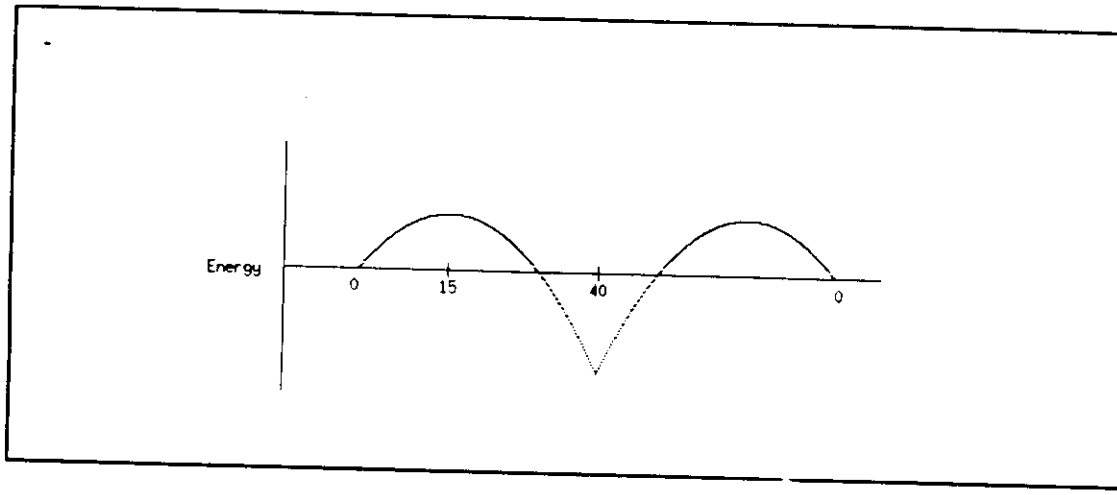


Figure 3:

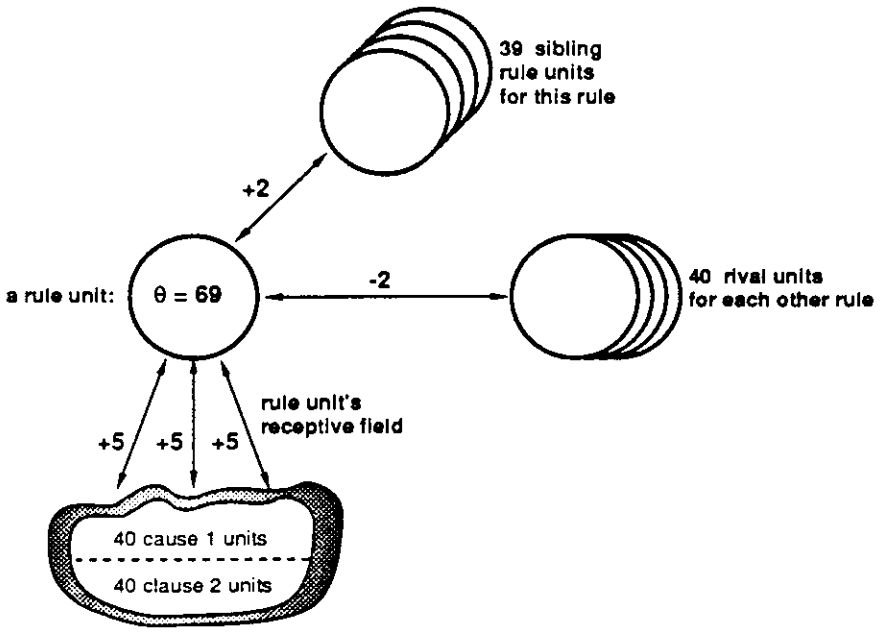

39 sibling
rule units
for this rule
+2
a rule unit:
θ = 69
-2
40 rival units
for each other rule
rule unit's
receptive field
+5
+5
+5
40 cause 1 units
40 clause 2 units

Figure 4:

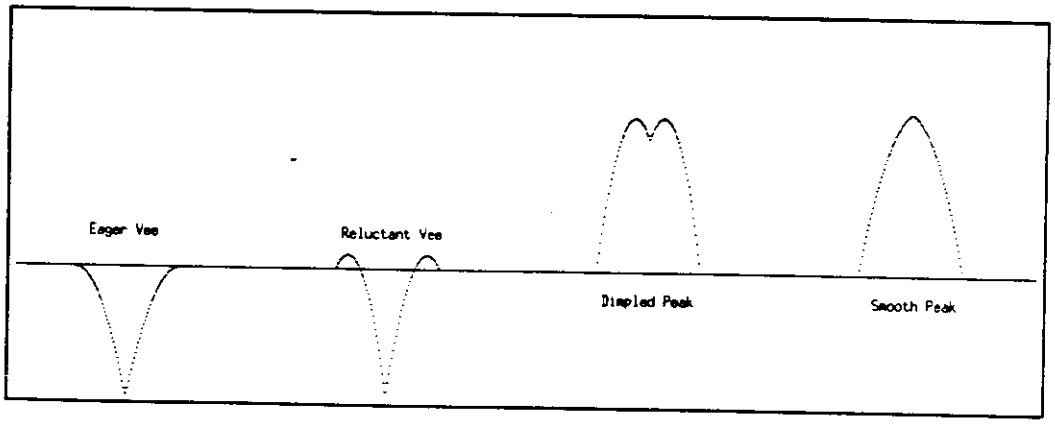


Figure 5:

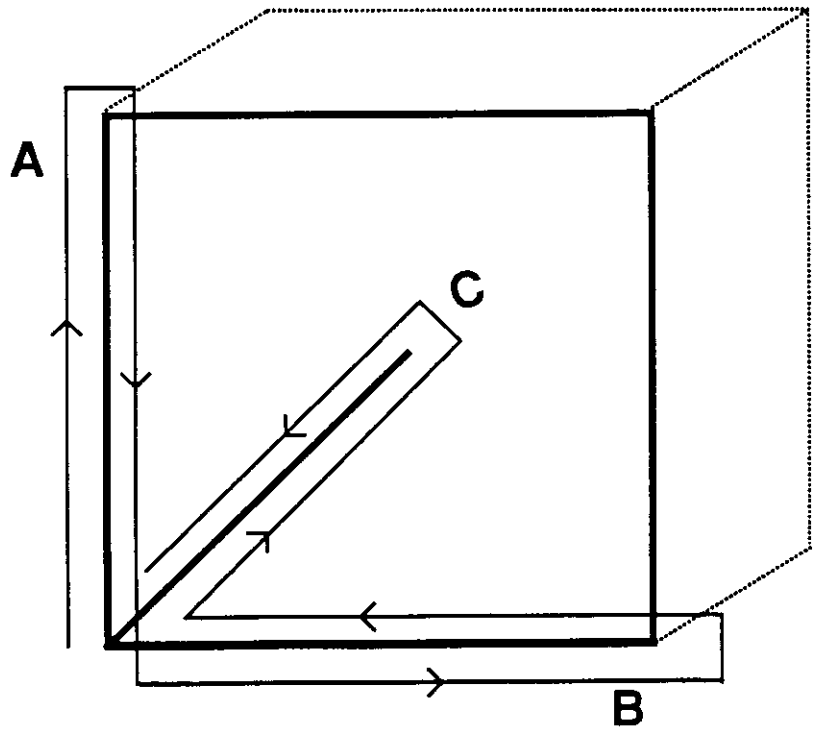


Figure 6:

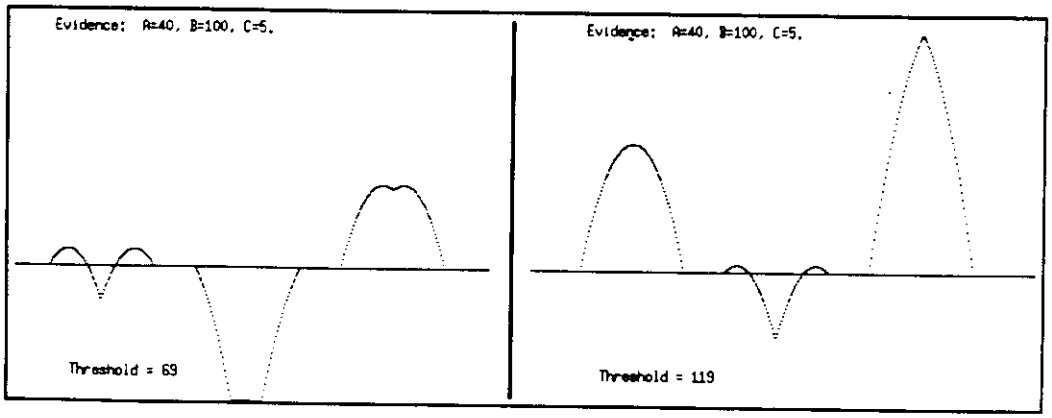

Evidence: A=40, B=100, C=5.
Threshold = 69
Evidence: A=40, B=100, C=5.
Threshold = 119

Figure 7:

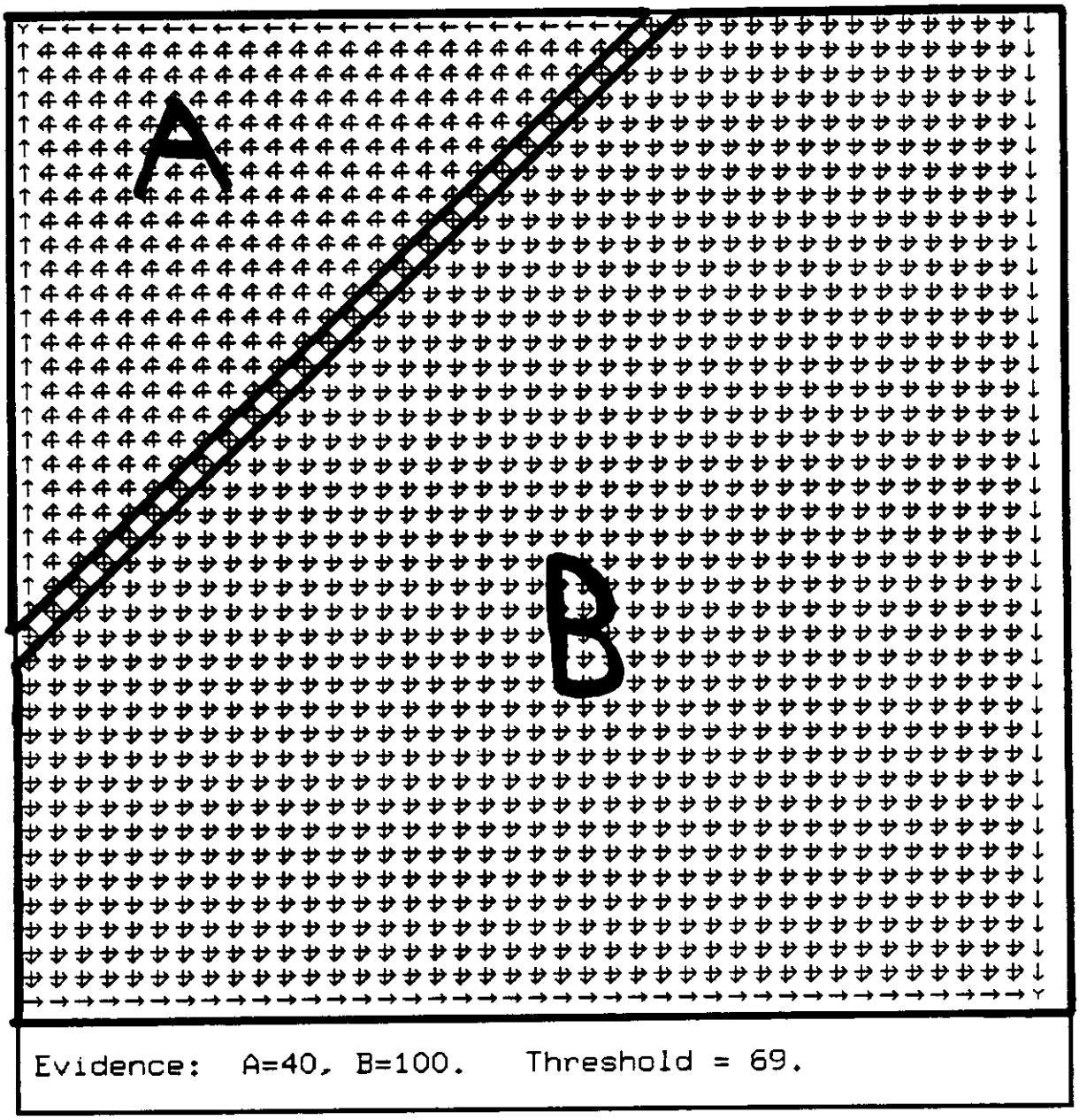
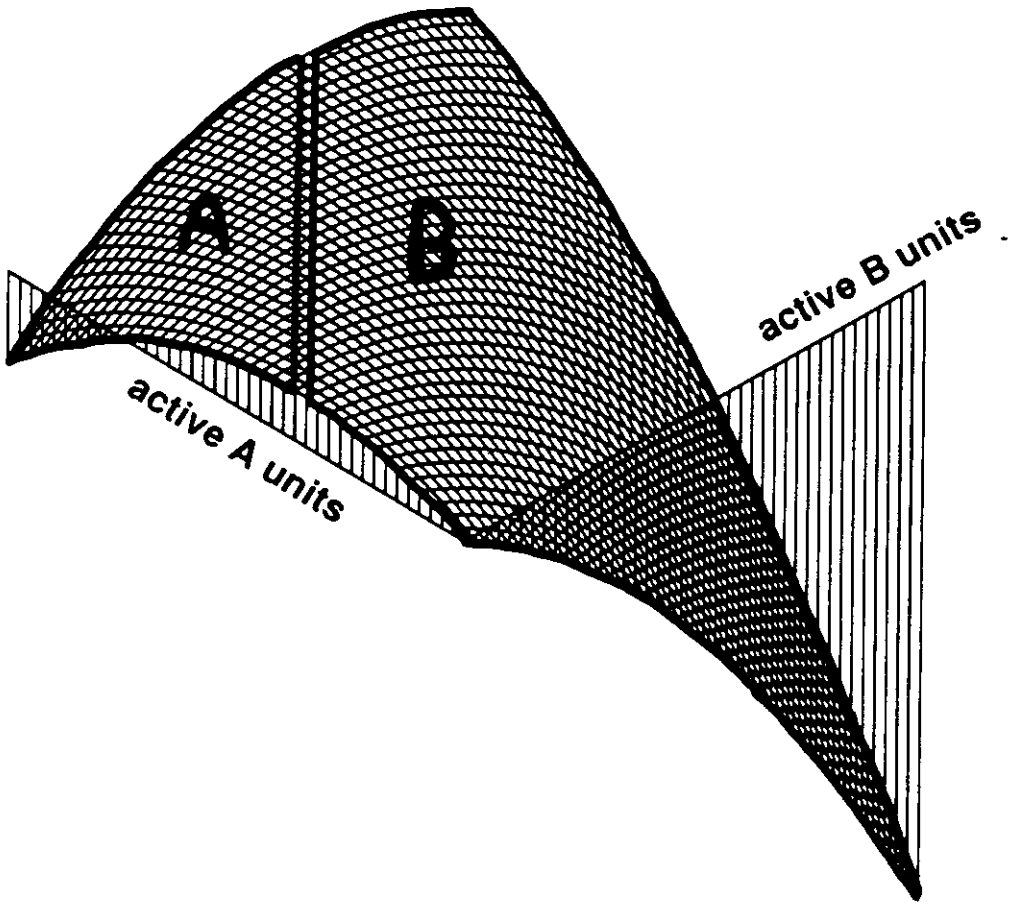


Figure 8:



Evidence: A=40, B=100. Threshold = 69.

Figure 9:

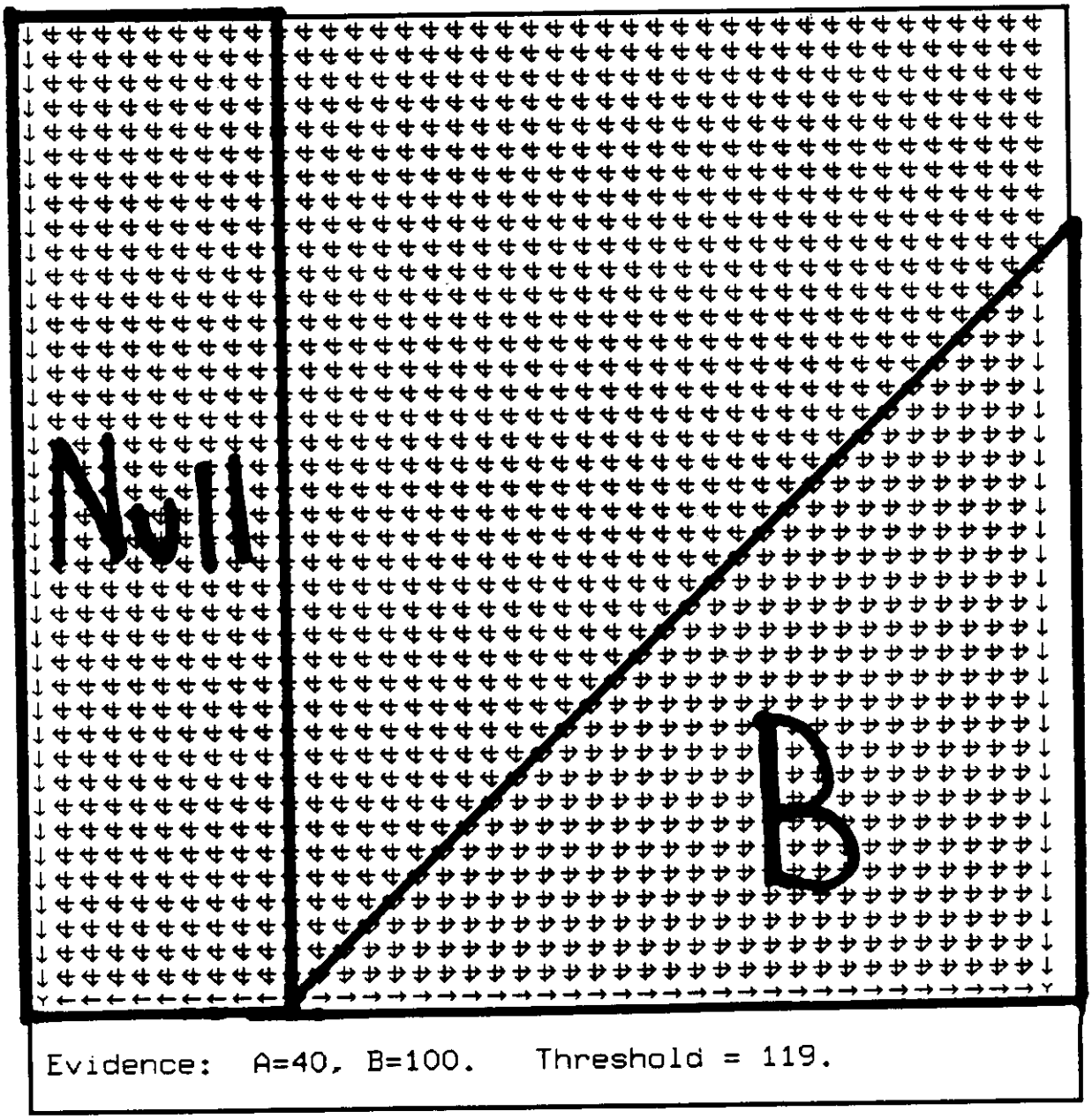
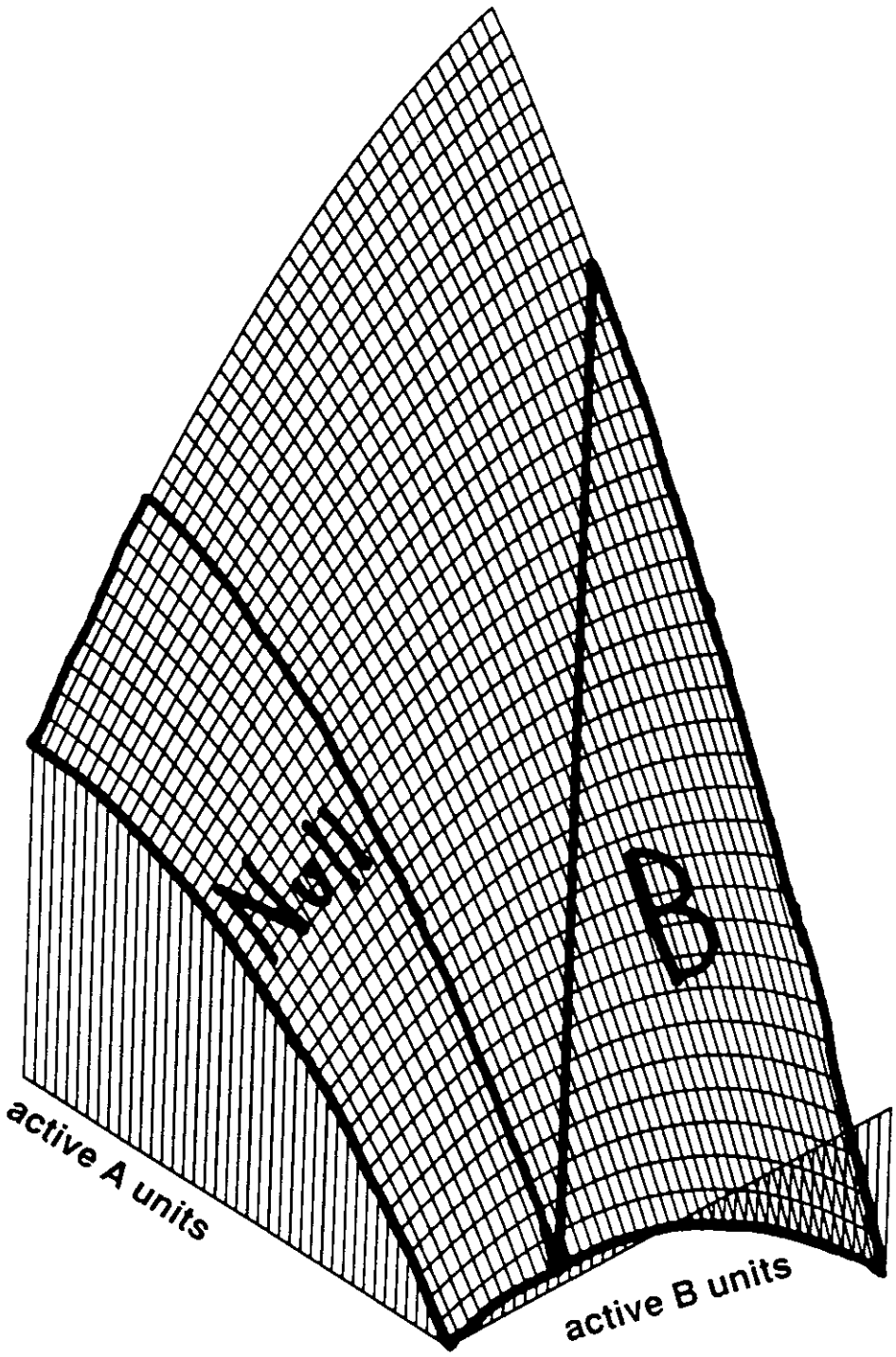


Figure 10:



Evidence: A=40, B=100. Threshold = 119.

Figure 11:

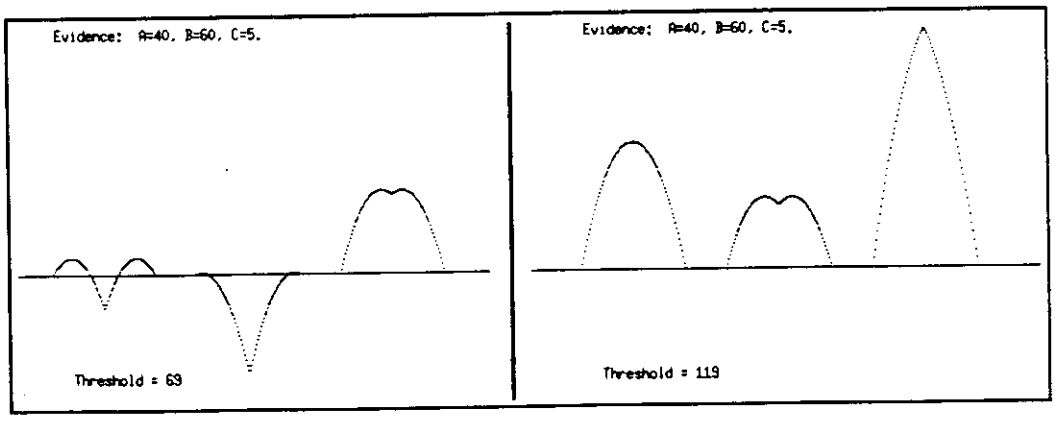

Evidence: A=40, B=60, C=5.
Threshold = 69
Evidence: A=40, B=60, C=5.
Threshold = 119

Figure 12:

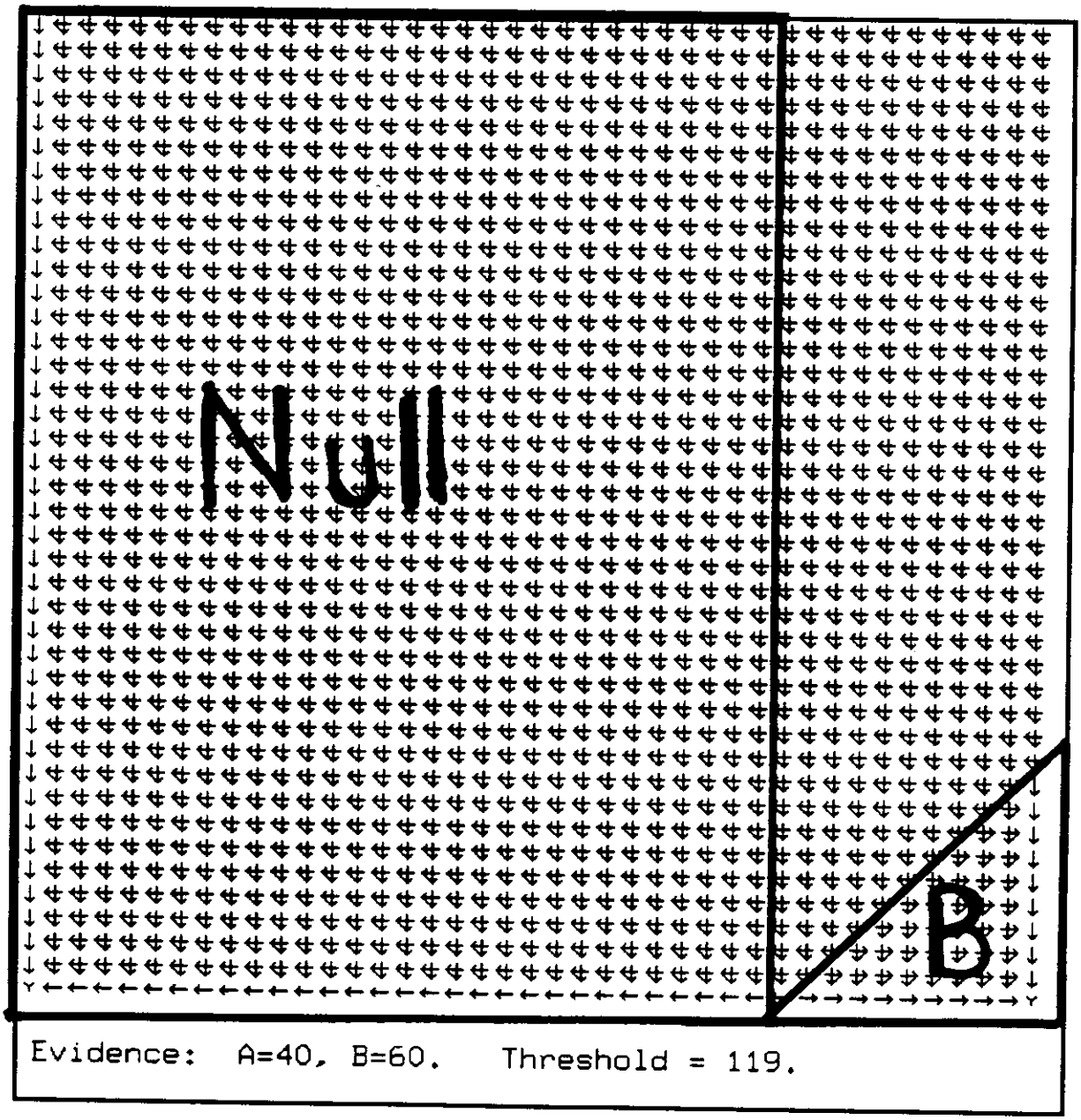
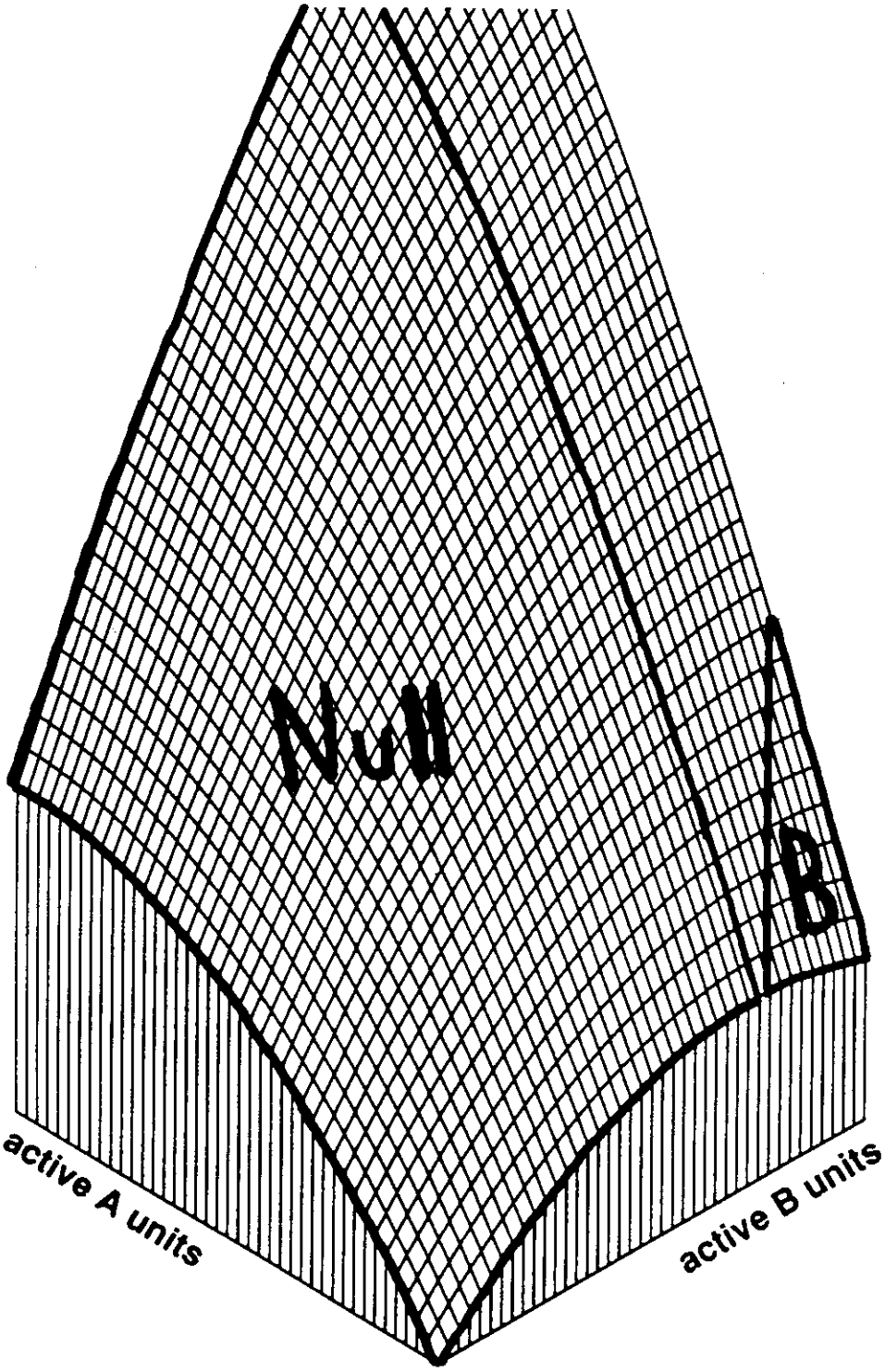


Figure 13:



Evidence: A=40, B=60. Threshold = 119.

Figure 14:

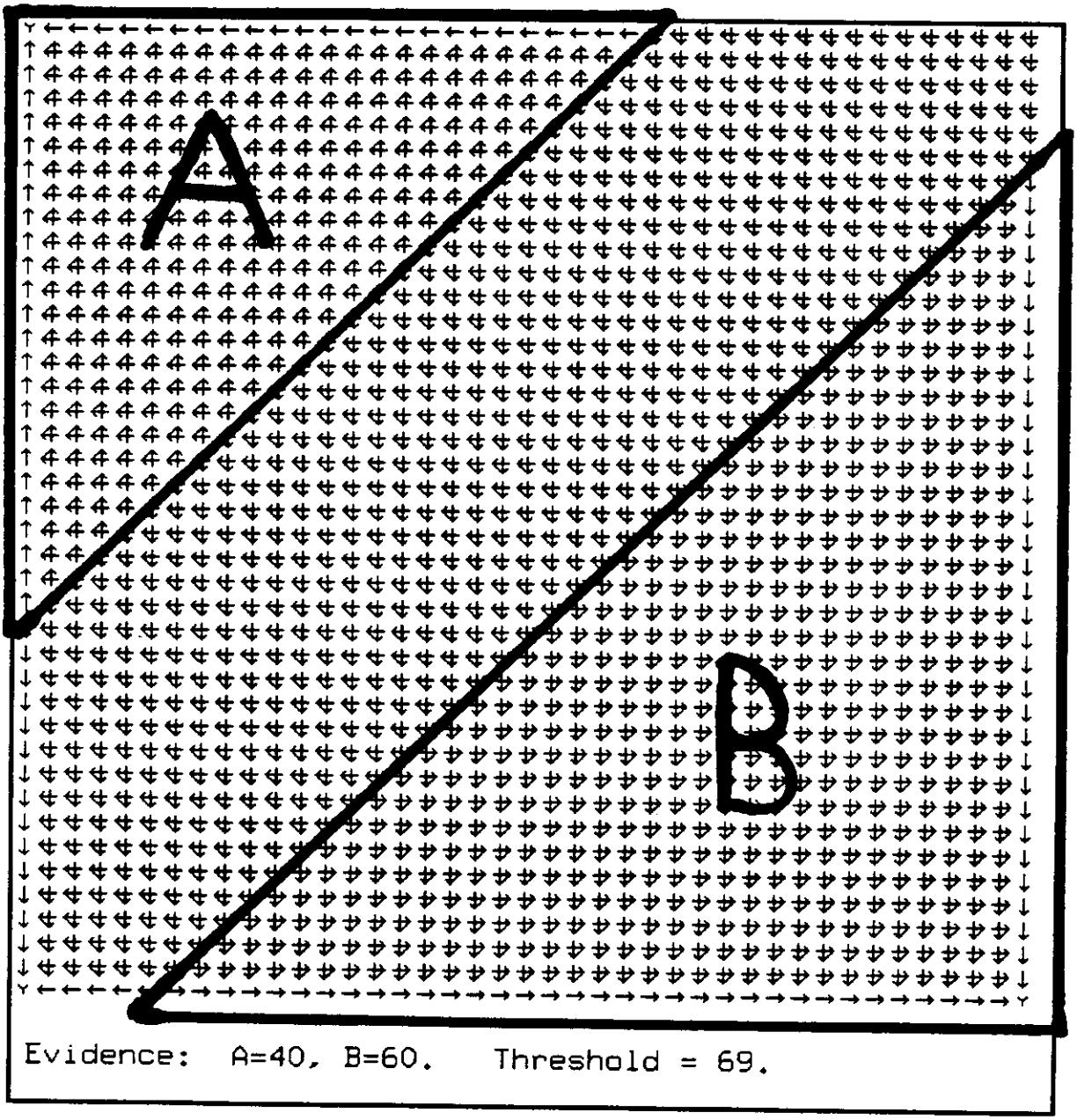
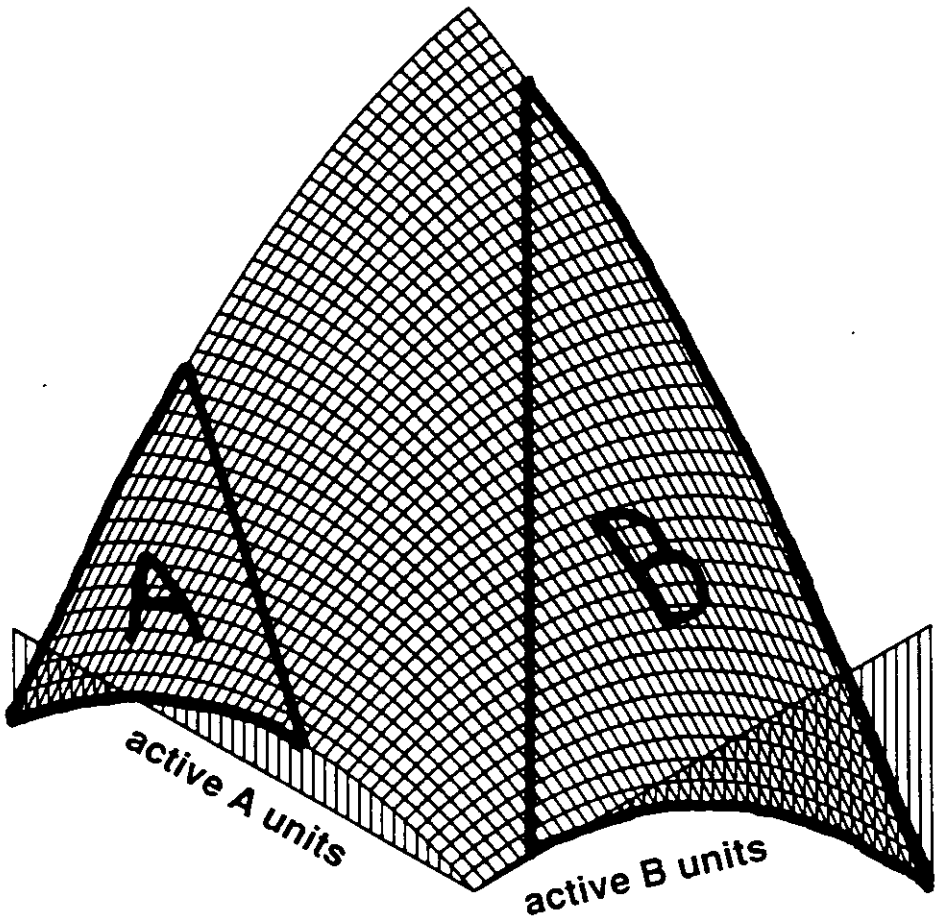


Figure 15:



Evidence: A=40, B=60. Threshold = 69.

Figure 16:

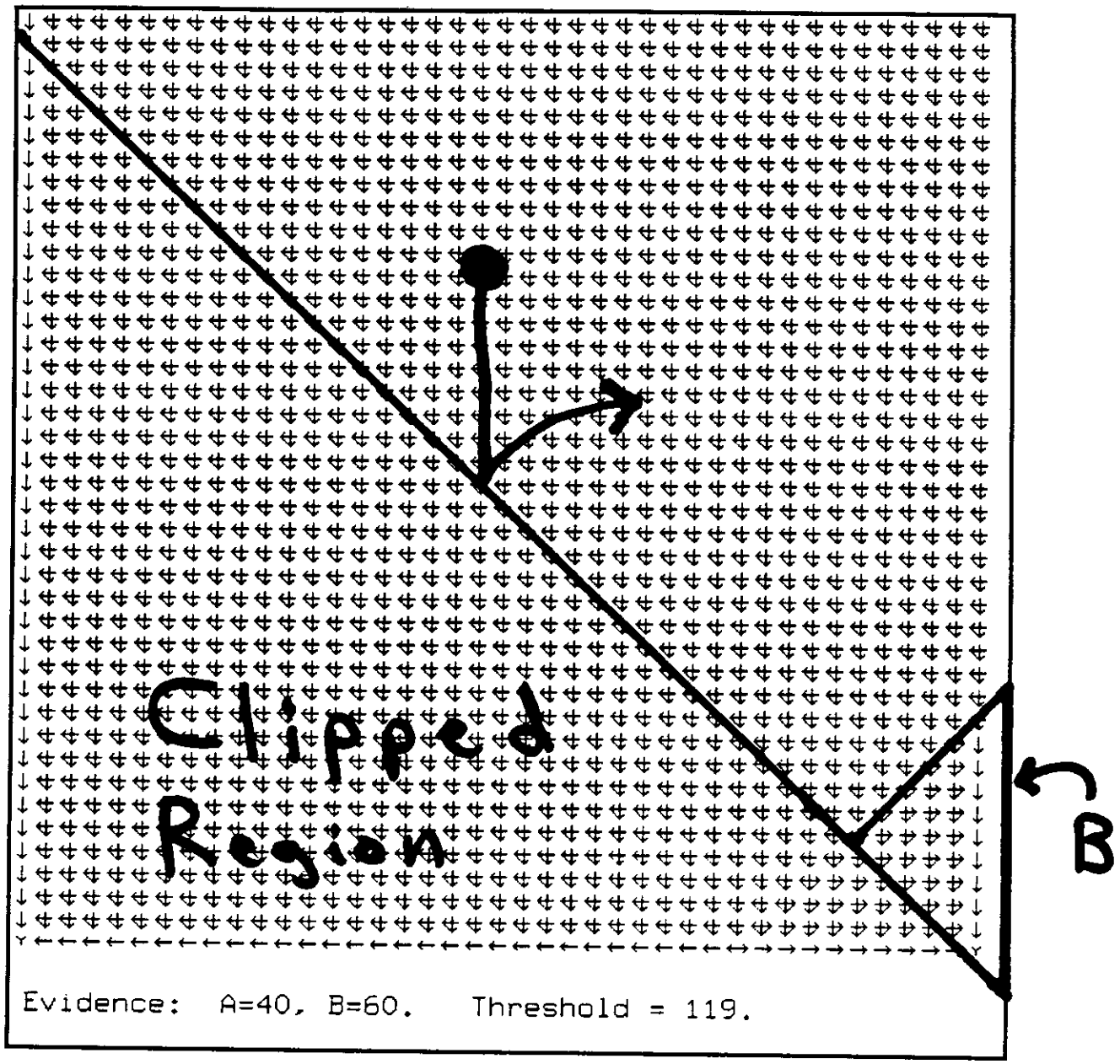
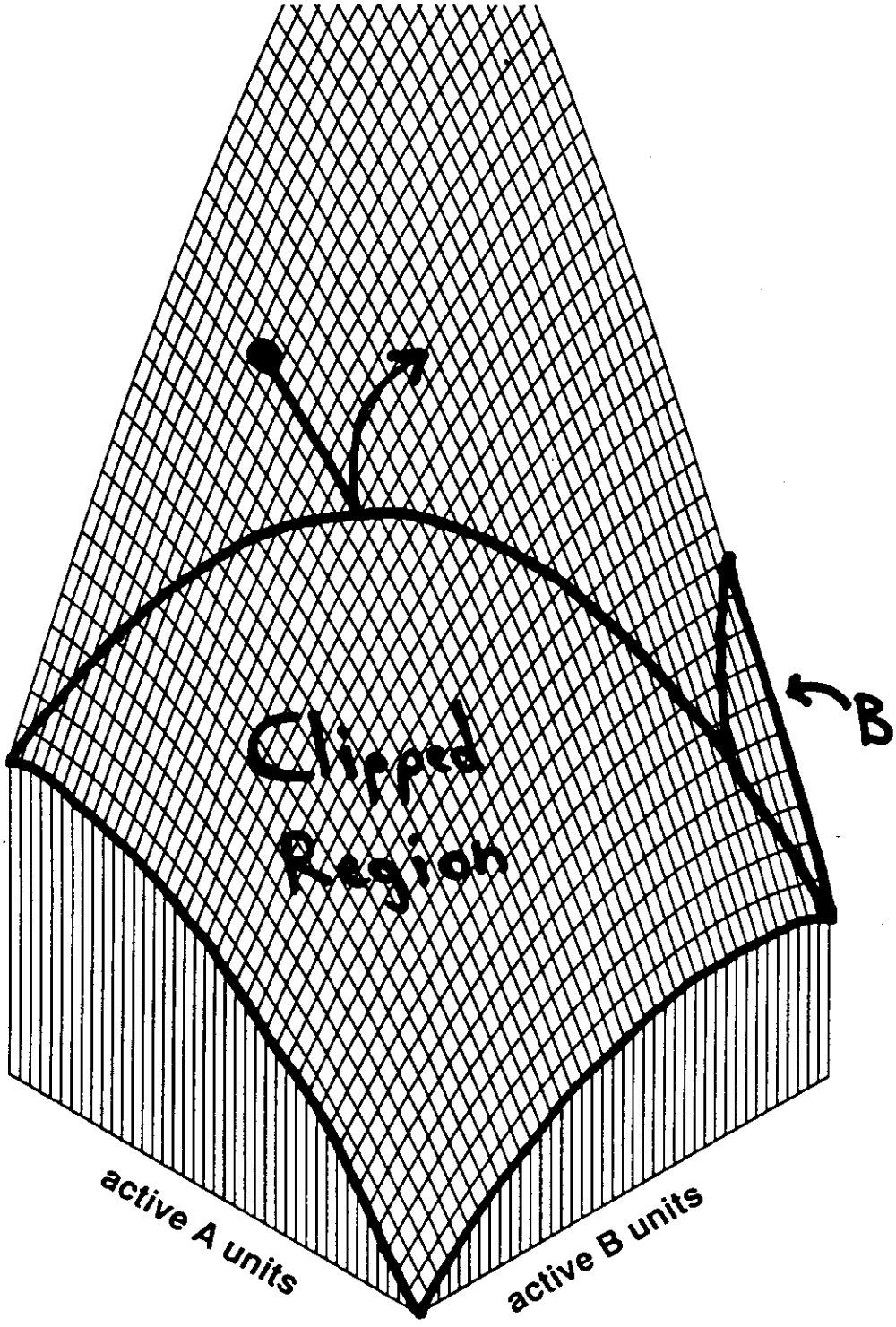


Figure 17:



Evidence: A=40, B=60. Threshold = 119.

Figure 18:

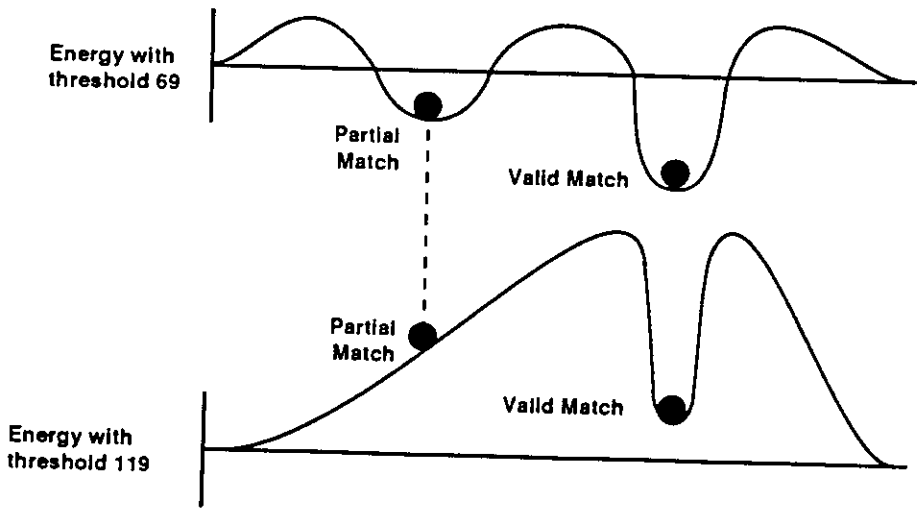

Energy with
threshold 69
Partial
Match
Valid Match
Partial
Match
Valid Match
Energy with
threshold 119

Figure 19:

Clause unit threshold	901
Clause unit WM input	900
Clause unit lateral inhibition	-2
Clause unit rule weight	5
Clause unit bind wieght	7
Rule unit threshold	69
Rule unit sibling weight	2
Rule unit rival weight	-2
Rule unit clause weight	5
Bind unit threshold	63
Bind unit sibling weight	1
Bind unit rival weight	-2
Bind unit clause weight	7

Table 1:

Step	# Cycles	Temperature	Rule Threshold Increment	Bind Threshold Increment	Clause Threshold Increment
Randomize	2	300			
Match	10	33	0	0	0
Settle	3	0.1	0	0	0
Rebias	2	0.1	+50	+30	+50

Table 2:

Clause unit threshold	901
Clause unit WM input	900
Clause unit lateral inhibition	-2
Clause unit rule weight	5
Clause unit bind wieght	7
Rule unit threshold	29
Rule unit sibling weight	1
Rule unit rival weight	-1
Rule unit clause weight	5
Bind unit threshold	63
Bind unit sibling weight	1
Bind unit rival weight	-2
Bind unit clause weight	7

Table 3:

Step	# Cycles	Temperature	Rule Threshold Increment	Bind Threshold Increment	Clause Threshold Increment
Randomize	2	300			
Match	6	30	0	0	0
Match	2	30	+5	+5	+15
Match	2	30	+10	+10	+30
Settle	2	0.1	+10	+10	+30
Rebias	2	0.1	+50	+30	+50

Table 4: