A SELF-ORGANIZING CLUSTER PROCESS

By Robert M. Burton[1] and William G. Faris[2]

Oregon State University and University of Arizona

The state of the self-organizing cluster process is a finite subset of points in a bounded region. This subset represents an evolving discrete approximation to a continuous probability distribution in the region. The dynamics of the process is determined by an independent sequence of random points in the region chosen according to the distribution. At each time step the random point attracts the nearest point in the finite set. In this way the subset learns to approximate its environment. It is shown that initial states approach each other exponentially fast for all time with probability one. Thus all memory of the initial state is lost; the environment alone determines future history.

1. Introduction. The subject is a Markov chain whose state is a finite subset of a bounded region in Euclidean space. The time evolution of the chain is determined by a random environment consisting of an independent and identically distributed sequence of points in the bounded region. At each time step the nearest point in the finite subset is moved toward the random point. This chain is the simplest case of a family of models introduced by Kohonen [8] in which the state is an embedding of a graph in the bounded region. These models are Markov chains in which the state evolves to give a discrete graphical approximation to the probability distribution of the random points.

These models have several possible interpretations. Abstractly one can think of the graph as a geometrical approximation to the probability distribution. A one- or two-dimensional graph embedded in a Euclidean space of higher dimension would evolve to an approximation to a space-filling curve or surface that attempts to parameterize the regions of high probability.

Other interpretations are closer to neurobiology [12]. One can think of each point of the graph as a neuron. The dimensions of the higher dimensional Euclidean space correspond to different fibers carrying input stimuli. An embedding of the graph in this space assigns to each neuron a connection weight for each fiber. An overall stimulus is a point in this Euclidean space. The neuron that responds is that neuron whose input fiber weights best correspond to the stimulus. This neuron and its neighbors in the graph then adjust their weights according to the stimulus. In this way the neuronal weights evolve to make the neuronal responses into a kind of map of the incoming stimuli.

The particular case treated here corresponds to the situation when the graph has vertices but no edges. Even for this case there are new mathematical challenges when the dimension of the Euclidean space is greater than 1. We give conditions that guarantee that two initial states approach each other exponentially fast for all time with probability 1. Thus the initial state does not matter; the environment determines future history. There is a sense in which the system is guaranteed to learn from experience.

Processes of this type fall into a general scheme for constructing Markov chains. In this scheme there is a state space and a sequence $F_1, \ldots, F_n, \ldots$ of independent random functions from the state space to itself. Take an initial point, X_0. The random orbit defined by the iterations $X_n = F_n \cdots F_1 X_0$ is the Markov chain.

The self-organizing cluster process under consideration is constructed in this way. Consider a bounded region in $\mathbf{R}^\nu$ (We shall always take this to be a rectangular product of intervals). Fix a number l. The state space is the collection of finite subsets of the region, each with l elements. Fix a probability measure μ on the region and a shrinking parameter α with $0 < \alpha < 1$.

For mathematical investigations it is convenient to take μ to be the uniform measure on a rectangular product of intervals. We shall always make this choice. Furthermore, we shall always assume that we are working in dimension $\nu \geq 2$. (The problems in the case of dimension $\nu = 1$ are simpler, but the geometry requires special consideration.)

DEFINITION 1.   The self-organizing cluster process is defined as follows. For each n choose an independent point ω_n in the bounded region from the probability distribution μ. Then choose the point x in the finite subset X that minimizes the Euclidean distance $|x - \omega_n|$. Then $F_n X$ is the new subset where x is replaced by $\alpha \omega_n + (1 - \alpha)x$ and the other points are unchanged. For each initial subset X_0 the process is defined by $X_n = F_n X_{n-1}$.

We consider the stability question, formulated as follows. Take two initial states X and Y, and look at their orbits under the same random sequence of functions. The basic question is whether they will eventually get close and stay close for all future time.

If X and Y are close, then each point of X is paired with a point of Y. It is possible that the random ω may respect the pairing, in the sense that the x that is closest to ω may be paired with the y that is closest to ω. In that case, $x - y$ is replaced by $(1 - \alpha)(x - y)$. Thus the two points have been shrunk together. On the other hand, if the pairing is violated by ω, then the close particles split apart, and this shrinking process may be destroyed. So the question is whether an overall shrinking effect can persist, in spite of the risk of such a split.

This question was suggested by work on the Kohonen process of topological self-organization ([7], [8]) where the state is a function from a graph into a bounded region in $\mathbf{R}^\nu$. Its ergodic behavior was studied rigorously in a fundamental paper by Cottrell and Fort [3]. However they only treat the one-

dimensional case $\nu = 1$. Bouton and Pagés [1, 2] have made further progress in this case.

Topological self-organization in higher dimensions has been studied by various authors. Ritter, Martinetz and Schulten [12] present an overview. Tolat [13] and Lo and Bavarian [10] provide useful insights. Erwin, Obermayer and Schulten [6] and Li, Gasteiger and Zupan [9] give references to other work in this area. There are still aspects of the problem that have eluded rigorous analysis.

Our objective is to develop tools that may be used to study topological self-organization in higher dimensions. The model that we study here does not have topological features, but it is interesting in its own right as a model of how a finite set of points can evolve to mimic a probability distribution. The strategy of the analysis is to compare two trajectories; the general result of Theorem 1 then gives ergodicity. The simplest properties of the model are presented in Lemmas 1, 2 and 3. Lemmas 4 and 5 show that there is a nonzero probability of getting into a good starting configuration and then having the two trajectories approach each other exponentially; this situation is summarized in Lemma 6. Lemma 7 shows that if something goes wrong with this convergence, then it is likely to be rather soon. There are repeated chances for the convergence to take place. Thus Lemma 6 and Lemma 7 combine to show that the exponential convergence takes place after a random lag; this is Theorem 2, the main result of the paper.

2. Markov chains as dynamical systems.

This section formulates the notion of stability for a Markov chain with continuous state space in a more abstract and precise form. This form of stability turns out to be considerably stronger than the condition of uniqueness of the invariant probability measure.

Suppose that (Y, d) is a bounded separable metric space. Assume without loss of generality that the diameter of Y is 1. In this context we shall let ν denote a probability measure defined on Borel subsets of Y. One way to define Markov chains on Y is as a random dynamical system. Suppose that $F_1, F_2, \ldots$ is an i.i.d sequence of random transformations from Y to itself. A Markov chain $\{X_n\}$ may be defined by setting X_0 to have an initial distribution ν and setting $X_n = F_n \circ F_{n-1} \circ \cdots \circ F_1(X_0)$, where $\circ$ denotes composition.

All Markov chains may be expressed this way (we do not assume that the transformations are continuous). Often this is the natural way to define the chain. This is the case for the self-organizing feature maps; indeed they were defined in exactly this way.

DEFINITION 2. A Markov chain given as a dynamical system with random transformation distribution F is superstable if (i) there is a constant $\gamma > 0$ so that for every pair of initial values x and y there is a random variable $W < \infty$ with integer values $0, 1, 2, 3, \ldots$, so that

$$(1) \qquad d(F_n \circ F_{n-1} \circ \cdots \circ F_1(x), F_n \circ F_{n-1} \circ \cdots \circ F_1(y)) \le e^{-\gamma(n-W)}$$

and (ii) there is an invariant probability measure for the chain. We say the chain is superstable with exponential rate if in addition (iii) there are constants $k > 0$ and $0 < \beta < 1$ independent of x and y so that $\mathbf{P}[W = n] \le k\beta^n$.

Note that this definition is in some sense dual to the property of being a chaotic attractor in which all the randomness is in the initial conditions. Here the initial conditions are irrelevant in the long term; the environment alone gives the behavior.

Condition (i) says that two trajectories approach each other exponentially fast. Condition (ii) guarantees at least one invariant measure $\bar{\nu}$. Together this is a very strong stability condition, giving a unique invariant probability measure that attracts all other measures ν. If we also have (iii), then this convergence is exponentially fast, in a precise sense to be explained below.

The content of the following remark is that if we replace the fixed initial values x and y with initial random variables X_0 and Y_0, then we get an equivalent formulation of superstability.

REMARK 1. A Markov chain given as a dynamical system with random transformation distribution F is superstable if and only if (i) there is a constant $\gamma > 0$ so that for every pair of initial random variables X_0 and Y_0 with (X_0, Y_0) independent of the F_i there is a random variable $W < \infty$ with integer values $0, 1, 2, 3, \ldots$, so that

$$(2) \qquad d(F_n \circ F_{n-1} \circ \cdots \circ F_1(X_0), F_n \circ F_{n-1} \circ \cdots \circ F_1(Y_0)) \le e^{-\gamma(n-W)}$$

and (ii) there is an invariant probability measure for the chain. The chain is superstable with exponential rate if and only if in addition (iii) there are constants $k > 0$ and $0 < \beta < 1$ independent of X_0 and Y_0 so that $\mathbf{P}[W = n] \le k\beta^n$.

Let $\mathcal{M}$ be the set of Borel probability measures on Υ and let $\mathcal{C}$ be the set of bounded continuous real-valued functions on Υ. Define $T : \mathcal{M} \to \mathcal{M}$ by $T(\nu) = \mathrm{dist}(F(X))$ where $\nu = \mathrm{dist}(X)$ and F is distributed as $F_1, F_2, \ldots$.

We may also put a metric on $\mathcal{M}$, the Wasserstein metric, which we define by $\rho(\nu_1, \nu_2) = \inf\{\mathbf{E}[d(X_1, X_2)] \mid \mathrm{dist}(X_i) = \nu_i, \ i = 1, 2\}$. The Kantorovich–Rubinstein theorem together with the assumption that the metric d is bounded implies ρ is a metric which characterizes convergence in distribution. That is, $\rho(\nu_n, \nu) \to 0$ iff $\nu_n \to \nu$ (dist) ([4], page 329). This is a complete metric if Υ is Polish.

THEOREM 1. *Suppose that a Markov chain $\{X_n\}$ defined as a dynamical system on Υ with random transformation distributed as F is superstable with $\bar{\nu}$ an invariant probability. Then $\bar{\nu}$ is the unique invariant measure; indeed the distribution of X_n converges weakly to $\bar{\nu}$ with respect to the function class $\mathcal{C}$. If the chain is superstable with exponential rate, then this convergence is exponentially fast with respect to the Wasserstein metric.*

PROOF.   Fix $f \in \mathscr{C}$ and $\nu \in \mathscr{M}$. Let Y be a random variable with distribution $\bar{\nu}$ and let X have distribution ν so that $T^n \nu$ is the distribution of $F_n \circ \cdots \circ F_1(X)$ and $T^n \bar{\nu} = \bar{\nu}$ is the distribution of $F_n \circ \cdots \circ F_1(Y)$. Then

$$(3) \qquad \int f \, dT^n \nu - \int f \, d\bar{\nu} = \mathbf{E}[f(F_n \circ \cdots \circ F_1(X)) - f(F_n \circ \cdots \circ F_1(Y))].$$

From the remark on the definition of superstability, the distance between $F_n \circ \cdots \circ F_1(X)$ and $F_n \circ \cdots \circ F_1(Y)$ goes to zero as $n \to \infty$ almost surely. Since f is continuous, it follows that the distance between $f(F_n \circ \cdots \circ F_1(X))$ and $f(F_n \circ \cdots \circ F_1(Y))$ also goes to zero as $n \to \infty$ almost surely. Thus the bounded convergence theorem implies that the expectation goes to zero as $n \to \infty$.

To see convergence in the Wasserstein metric with exponential rate, write

$$
\begin{aligned}
\rho(T^n(\nu), \bar{\nu}) &= \rho(T^n \nu, T^n(\bar{\nu})) \\
&\le \mathbf{E}[d(F_n \circ \cdots \circ F_1(X), F_n \circ \cdots \circ F_1(Y))] \\
&\le \mathbf{E}[\min(1, e^{-(n-W)\gamma})] \\
(4) \qquad &= \mathbf{P}[W > n] + \mathbf{E}[e^{-(n-W)\gamma}\mathbf{1}_{\{W \le n\}}] \\
&\le k\beta^{n+1}/(1-\beta) + ke^{-n\gamma} \sum_{w=1}^{n} e^{w\gamma}\beta^w \\
&\le k_1[\max(\beta, e^{-\gamma})]^n,
\end{aligned}
$$

where k_1 is another constant. $\square$

3. Hitting and splitting.   The self-organizing cluster process fits into the general scheme of the previous section. The metric space Υ consists of all l element subsets of the bounded region in $\mathbf{R}^\nu$. (Henceforth we revert to the notation where ν denotes the dimension of the Euclidean space.) The metric δ should be defined such that two subsets are close if there is some matching of points in the two subsets that makes the points in each of the matched pairs close.

Let X and Y be l-element finite subsets of the bounded region in $\mathbf{R}^\nu$. (We shall often think of X as representing a collection of l particles and of Y as representing another collection of l shadow particles.) Consider functions π making a correspondence from X onto Y. The distance between X and Y is defined by

$$(5) \qquad\qquad \delta(X, Y) = \min_{\pi} \max_{x \in X} |x - \pi x|,$$

where the norm is the Euclidean distance in $\mathbf{R}^\nu$.

Let X be an l-element finite subset of the region. We shall be concerned about the minimum distance between particles in this one subset. This is the *separation* defined by

$$(6) \qquad\qquad d(X) = \min_{x \ne x'} |x - x'|.$$

In the starting lemma (Lemma 4) and gluing lemma (Lemma 5), we shall need to bound the initial separation from below by a suitable $d_0 > 0$.

Recall that the evolution of the chain with state X at some particular time is determined by moving the particle closest to the random point ω part way toward ω. We say that a particle x is *hit* at this time if x is the element of X closest to ω_k.

Assume that π provides the correspondence between X and Y that defines $\delta(X, Y)$. We say that X and Y *split* if the random ω is closest to x in X and to y in Y with $y \neq \pi x$. We wish to show that the states evolving from X and Y eventually cease to split. Notice that if X and Y do not split, then the uniform distance from X to Y cannot increase.

The following simple shrinking lemma is fundamental.

LEMMA 1. *If initially $\delta(X, Y) \leq \varepsilon$, and after a certain number of steps, all l points are hit and there is no split, then the final X and Y satisfy $\delta(X, Y) \leq \varepsilon(1 - \alpha)$.*

The fundamental problem is to control the probability of a split. In dimension greater than 1 there can be a serious danger of a split when the separation is small; the following no-split lemma gives a bound on the probability of a split in terms of the separation.

LEMMA 2. *For each $\varepsilon > 0$ and $d > 0$, if the initial X and Y satisfy $\delta(X, Y) \leq \varepsilon$ and if the separation $d(X) \geq d > 0$, then the probability that X and Y split in the next step is bounded by $C_0 \varepsilon / d$.*

PROOF. Let $y = \pi x$ and $y' = \pi x'$ be the shadow particles corresponding to x and x'. We have $|x - y| \leq \varepsilon$ and $|x' - y'| \leq \varepsilon$, while $|x - x'| \geq d$. Consider the event that x and y' are hit. For this event to happen, the random point ω must satisfy $|\omega - x| < |\omega - x'|$ and $|\omega - y| > |\omega - y'|$. This says that $(x - x') \cdot (\omega - (x + x')/2) > 0$ and $(y - y') \cdot (\omega - (y + y')/2) < 0$. So ω must lie on one side of one half-plane and on the other side of the other half-plane. In order for this condition to define a small region, the half-planes must be nearly the same. The two planes contain the two points $(x + x')/2$ and $(y + y')/2$, which are both in the region and which are within distance ε of each other. So if the planes were parallel, the region would have volume proportional to ε. However, in general the size of the region will also have a contribution from a multiple of the sine of the angle between the two planes. The sine of the angle is bounded by the distance between the unit vectors $(x - x')/|x - x'|$ and $(y - y')/|y - y'|$ that are normal to the planes. This distance can itself be bounded in two steps. The distance between $(x - x')/|x - x'|$ and $(y - y')/|x - x'|$ is clearly bounded by $2\varepsilon/d$. On the other hand, the distance between $(y - y')/|x - x'|$ and $(y - y')/|y - y'|$ is bounded by $(|y - y'| - |x - x'|)/|x - x'|$, which is also bounded by $2\varepsilon/d$. So the sine is bounded by $4\varepsilon/d$. We conclude that the volume of the region of ω that satisfies both inequalities is bounded by a multiple of ε/d. □

The problem in dimension greater than 1 when the separation is small requires an estimate that shows that this is unlikely. This is the following repelling lemma.

LEMMA 3. *There is a constant k_0 such that if the separation between X particles is $d > 0$, then the probability that the two particles achieve a separation smaller than d on the next step is bounded by $k_0 d$.*

PROOF.   If two particles are separated by a distance, then the closest that they can come after one step is half that distance. So we only need to worry about particles that are within $2d$ of each other.

For x to approach x' we need that $|\alpha\omega + (1 - \alpha)x - x'| < |x - x'|$. This says that ω must fall within a ball of radius $|x - x'|/\alpha$. The probability of this happening is proportional to $|x - x'|^\nu$ which in turn is bounded by a multiple of d^ν. Since d itself is bounded above by the size of the region, this is bounded by a multiple of d. □

4. The starting lemma.   The starting lemma of this section states that there is a nonzero (although perhaps very small) probability that a desirable configuration is reached, whatever the initial configuration. The reason for the validity of the lemma is that practically anything can happen with some nonzero probability. However we must make sure that a bad initial state does not make this probability arbitrarily small.

The method of proof is to describe a sequence of events that can bring the state into the desired configuration. The probability of this sequence must be bounded below independent of the initial configuration. For simplicity we describe only the case when the dimension $\nu = 2$ and the region is the unit square $[0, 1]^2$.

OBSERVATION 1.   Suppose a curve C has endpoints u and v. Fix $\bar{\varepsilon} > 0$. Suppose C is sufficiently direct that there are disjoint open balls $B_1, \ldots, B_n$ of radius $\bar{\varepsilon}$ with the property that $u \in B_1$, $v \in B_n$, $C \subset \bigcup_{i=1}^{n} \bar{B}_i$, and for $1 \le i \le n - 1$, $\bar{B}_i \cap \bar{B}_{i+1}$ is a single point. Suppose there is an X particle in B_1. Then there exists $\bar{N} > 0$, $\bar{\delta} > 0$ depending only on $\bar{\varepsilon}$ and the shrinking parameter α so that, after $\bar{N}$ additional moves of the Markov chain, with probability greater than $\bar{\delta}$, there is an X particle in B_n that was originally within $4\bar{\varepsilon}$ of C. (Thus if the X particle in B_1 is the only X particle within $4\bar{\varepsilon}$ of C, it will have moved to within $2\bar{\varepsilon}$ of v after $\bar{N}$ moves, with probability greater than $\bar{\delta}$.)

PROOF.   Let $D_1, \ldots, D_{n-1}$ be open balls of radius $\bar{\varepsilon}$ with the center of D_i in $\bar{B}_i \cap \bar{B}_{i+1}$. Let B_i' and D_i' be open balls of radius $\bar{\varepsilon}/2$ and with the same centers as the corresponding B_i and D_i. If in the evolution of the Markov chain the noise process ω occurs sufficiently often in D_1', an X particle that was originally within $4\bar{\varepsilon}$ of C will come to be in D_1. Repeat this with B_2', D_2', B_3', D_3' and so on until there is an X particle in B_n. We have to pull the

particle into each of $2n$ balls in order and each time takes k moves, where k is chosen so $(1-\alpha)^k 2\sqrt{2} < \bar{\varepsilon}$. The number of hits needed is therefore $2nk$. Since the balls are disjoint, the number of balls n times the area $\pi\bar{\varepsilon}^2/4$ of a ball is bounded by 1. So n is less than the reciprocal of this number. We may thus take the upper bound on the number of hits to be $\bar{N} = 8k/(\pi\bar{\varepsilon}^2)$. Then the corresponding probability is bounded below by $\bar{\delta} = (\pi\bar{\varepsilon}^2/4)^{\bar{N}}$. □

OBSERVATION 2. Let A and B be disjoint subsets of the unit square $[0,1]^2$. Let $g_1 = \sup\{|u - v| \mid u \in A, \ v \in B\}$ and $g_2 = \inf\{|u - v| \mid u \in A, \ v \in B\}$. Suppose

(7) $$\alpha \operatorname{diam}(A) + (1-\alpha)g_1 < g_2.$$

Suppose there are X particles in B and no other X particles within g_1 of A. If the noise ω occurs repeatedly in A, then exactly one X particle from B will be affected. Note that the condition $\operatorname{diam}(A) + \operatorname{diam}(B) < \alpha g_2$ is sufficient to ensure condition (7) above.

PROOF. We need to show that a noise occurrence in A will bring an X particle from B to a location closer to any point in A than any other X particle in B. That is, we must show

(8) $$\sup\{|\alpha z + (1-\alpha)x - z_1| \mid z, z_1 \in A, \ x \in B\} < g_2.$$

But using the inequality (7) above we have $|\alpha z + (1-\alpha)x - z_1| = |\alpha(z - z_1) + (1-\alpha)(x - z_1)| \le \alpha|z - z_1| + (1-\alpha)|x - z_1| \le \alpha \operatorname{diam}(A) + (1-\alpha)g_1 < g_2$.

Now suppose $\operatorname{diam}(A) + \operatorname{diam}(B) < \alpha g_2$. The triangle inequality gives $g_1 \le \operatorname{diam}(A) + \operatorname{diam}(B) + g_2$, so $\alpha \operatorname{diam}(A) + (1-\alpha)g_1 \le \alpha \operatorname{diam}(A) + (1-\alpha)\operatorname{diam}(A) + (1-\alpha)\operatorname{diam}(B) + (1-\alpha)g_2 \le \operatorname{diam}(A) + \operatorname{diam}(B) + (1-\alpha)g_2 < \alpha g_2 + (1-\alpha)g_2 = g_2$. □

These observations are essential for the proof of the starting lemma.

LEMMA 4. *There exists an initial separation $d_0 > 0$ such that for all $\varepsilon_0 > 0$ there exist N and $\delta > 0$ such that for all initial states X_0 and Y_0 the probability that the separation $d(X_n) > d_0$ and the distance $\delta(X_n, Y_n) \le \varepsilon_0$ exceeds δ.*

PROOF. The proof proceeds in several steps.

Step 1. There is a event that gets all the X particles close to the origin at the lower left corner of the square.

Define the triangular region $F(\varepsilon) = \{(x, y) \mid 0 \le x, \ 0 \le y, \ x + y \le \varepsilon\}$. Then $\operatorname{diam}(F(\varepsilon)) = \sqrt{2}\varepsilon$. Pick a small $\varepsilon_1 > 0$. Let the noise occur often enough in $F(\varepsilon_1/2)$ that there is at least one X particle in $F(\varepsilon_1)$. If all the X particles are in $F(\varepsilon_1)$, then stop. Otherwise we want to get an additional X particle in $F(2\varepsilon_1)$. If there is already such a particle, stop. Otherwise draw a straight line segment between an X particle outside $F(\varepsilon_1)$ to the point $(\frac{3}{2}\varepsilon_1, \frac{3}{2}\varepsilon_1)$. Using Observation 1 with $\bar{\varepsilon} = \varepsilon_1/6$ and this segment for a curve, move an X particle

into $F(2\varepsilon_1) \setminus F(\varepsilon_1)$. At this point we have at least two X particles in $F(2\varepsilon_1)$. Continue this process with successive bands until all the X particles are in $F(l\varepsilon_1)$.

Step 2. There is an event that gets all X and Y particles close to the origin.

We already have all the X particles in $F(\varepsilon_2)$ where $\varepsilon_2 = l\varepsilon_1$. If all the Y particles are in $F((4/\alpha)\varepsilon_2)$, then we are done. Otherwise we want to place at least one Y particle in $F((4/\alpha)\varepsilon_2)$ while keeping all the X particles in $F((4/\alpha)\varepsilon_2)$. If there is already a Y particle in $F((4/\alpha)\varepsilon_2)$, stop. Otherwise for $\bar\varepsilon = \varepsilon_2/12$ let the noise occur in an $\bar\varepsilon/2$ neighborhood of one of the Y particles. By Observation 2 a unique X particle will be brought from $F(\varepsilon_2)$ to an $\bar\varepsilon$ neighborhood of this Y particle. Now draw a straight line segment from this Y particle to the point $((3/\alpha)\varepsilon_2, (3/\alpha)\varepsilon_2)$ and use Observation 1 with this $\bar\varepsilon$ to bring the X particle back into $F((4/\alpha)\varepsilon_2)$. Also at least one Y particle is in $F((4/\alpha)\varepsilon_2)$. At this point we have all X particles in $F((4/\alpha)\varepsilon_2)$ and at least one Y particle in $F((4/\alpha)\varepsilon_2)$. Proceeding in this fashion, we will eventually have all X and Y particles in $F(\varepsilon_3)$ where $\varepsilon_3 = (4/\alpha)^l\varepsilon_2$.

Step 3. There is an event that puts pairs of X, Y particles together.

Assume ε_1 was chosen small enough so that $\varepsilon_3 < \frac{1}{10} \min(1/l, 1/8)$ and $8\varepsilon_3[(1/\alpha) + 1] < 1$. Let A be an open ball of radius $\varepsilon_3/2$ and center $(\frac{1}{4}, \frac{1}{4})$ and $B = F(\varepsilon_3)$. By Observation 2, if the noise occurs repeatedly in A it will attract a unique (X, Y) pair to within ε_3 of $(\frac{1}{4}, \frac{1}{4})$. Now use Observation 1 to place this pair in an ε_3 neighborhood of $((1/2l), 7/8)$. Continue this until we have (X, Y) pairs in ε_3 neighborhoods of $((2i - 1/2l), 7/8)$ for $i = 1, 2, \ldots, l$.

Step 4. There is an event that gets the paired particles close.

Take $d_0 = 2/l$. Assume $\varepsilon_0 < 4/l$. By having the noise occur repeatedly in neighborhoods of radius $\varepsilon_0/4$ of the points $((2i - 1/2l), 7/8)$, $i = 1, 2, \ldots, l$, we may get with positive probability corresponding X, Y particles within ε_0 of each other. Thus $\delta(X_N, Y_N) \leq \varepsilon_0$ where N is the large but bounded number of steps used in this algorithm.

This completes the proof. $\square$

The starting lemma refers to particles and their shadow particles. The following corollary of the starting lemma refers only to the particles. We may refer to this as the hitting lemma.

COROLLARY 1. *There exists N and a strictly positive $\gamma > 0$ such that for all initial states X_0 the probability that all X particles are hit during N steps exceeds γ.*

5. Possible exponential attraction. In the following proofs we shall need to construct certain dominating random variables with specified probabilities. A typical situation is that we have a given Bernoulli random variable K with $\mathbf{P}[K = 1] = p$ and we are given p' with $p \leq p'$. We want to construct a Bernoulli random variable K' with $\mathbf{P}[K' = 1] = p'$ and such that $K = 1$ implies $K' = 1$.

The solution is to enlarge the probability space by introducing a new independent random variable U with uniform distribution in the unit interval. Then we define $K' = 1$ if $K = 1$. We also define $K' = 1$ if $K = 0$ and $U \leq (p' - p)/(1 - p)$. Otherwise we define $K' = 0$.

The main lemma in this section is the following gluing lemma.

LEMMA 5. *There exists $\varepsilon_0 > 0$, $d_0 > 0$, $\beta > 0$ and $\rho > 0$ such that for all X_0 and Y_0 with $\delta(X_0, Y_0) < \varepsilon_0$ and $d(X_0) \geq d_0$, the probability that for all n we have the estimate $\delta(X_n, Y_n) \leq \varepsilon_0 e^{-\beta n}$ exceeds ρ.*

PROOF. Since we are concerned about close approaches, we say that there is a *record value* at n if $d(X_n) < \min\{d(X_m) \mid m < n\}$. Now we want to divide the time index set into blocks of size N. The ith block goes from $(i-1)N + 1$ to iN. We say that the block i has a *record event* if there is at least one record value for the X in block i.

We say that the block i has an *all-hit event* if each X particle is moved sometime in block i.

Let S_i be the event of a split somewhere in the first i blocks. We take S_0 to be impossible. The S_i are increasing with i, and our ultimate objective is to bound the probability that S_i occurs for some i.

The main problem is to show that there are sufficiently few record events and sufficiently many all-hit events. For this purpose we define Bernoulli random variables J_i^0 and K_i^0 that score a record event in block i and an all-hit event in block i. Let $j^0(i) = \sum_{i'=1}^{i} J_{i'}^0$ and $k^0(i) = \sum_{i'=1}^{i} K_{i'}^0$ be the accumulated number of record events and all-hit events.

We need several probability estimates. Let $\mathscr{G}_i$ be the σ-algebra defined by the history of the process in the first i blocks.

(i) There is always some probability of an all-hit event:

$$(9) \qquad \mathbf{P}[K_i^0 = 1 \mid \mathscr{G}_{i-1}] \geq \gamma > 0.$$

(ii) The probability of a record value is less than $k_0 d(X)$. Furthermore, each record event shortens $d(X)$ by at most 2^{-N}. Thus the conditional probability of a record event given the past history is bounded above by $N k_0 d_0 2^{-Nj}$, where j is the number of previous record events. That is:

$$(10) \qquad \mathbf{P}[J_i^0 = 1 \mid \mathscr{G}_{i-1}] \leq N k_0 d_0 2^{-Nj^0(i-1)}.$$

(iii) The conditional probability of a first split in i given the past is bounded by $N C_0 \varepsilon_0 (1 - \alpha)^k / (d_0 2^{-N(j+1)})$. Thus:

$$(11) \qquad \mathbf{P}[S_i \mid \mathscr{G}_{i-1}] \leq N C_0 \frac{\varepsilon_0}{d_0} 2^N (1 - \alpha)^{k^0(i-1)} 2^{Nj^0(i-1)}$$

on S_{i-1}^c.

These estimates are awkward to use because there is no reason to believe that the J_i^0 and K_i^0 variables are independent. The idea of the proof is to construct independent Bernoulli random variables J_i and K_i for $i \geq 1$ that

under certain conditions give upper bounds and lower bounds, respectively, on the occurrence of record events and all-hit events. Let $j(i) = \sum_{i'=1}^{i} J_{i'}$ and $k(i) = \sum_{i'=1}^{i} K_{i'}$ be counters corresponding to accumulated record event bounds and accumulated all-hit event bounds.

The construction of the random variables is by induction. We let $\mathscr{F}_{i-1}$ be the σ-algebra generated by $S_{i-1}, J_1, \ldots, J_{i-1}, K_1, \ldots, K_{i-1}$. Thus $\mathscr{F}_0$ is the trivial σ-algebra. Similarly, we let $\mathscr{F}_{i-1}^{+}$ be the σ-algebra generated by $\mathscr{F}_{i-1}$ together with S_i and J_i.

We will construct J_i to have a conditional probability on each atom of $\mathscr{F}_{i-1}$ that only depends on $j(i-1)$. Similarly, we will construct K_i to have a constant conditional probability on each atom of $\mathscr{F}_{i-1}^{+}$. This construction automatically makes the J family and K families of random variables independent and independent of each other.

Recall that the shrinking parameter α satisfies $0 < \alpha < 1$. Fix $c_0 > 0$ such that

$$(12) \qquad \rho = \left(\frac{2^N}{1-\alpha} \right)^{c_0} (1-\alpha) < 1.$$

Let E_i be the event that $j(i') \leq c_0 k(i')$ for all $i' \leq i$. Thus E_i is the event that the number of record event bounds is not excessive compared to the number of all-hit event bounds. We take E_0 as the sure event.

The sets corresponding to given values of $j(i')$ and $k(i')$ for $i' < i$ belong to the σ-algebra $\mathscr{F}_{i-1}$. In particular E_{i-1} is in the σ-algebra $\mathscr{F}_{i-1}$.

The actual requirements on the bounding random variables are the following.

(i) If there has been no split, so S_{i-1} does not occur, and if there has been no crossing, so E_{i-1} occurs, then the occurrence of a record event $J_i^0 = 1$ implies that $J_i = 1$. Thus

$$(13) \qquad J_i^0 \leq J_i \quad \text{on } S_{i-1}^c \cap E_{i-1}.$$

(ii) If there is currently still no split, so S_i does not occur, and if there has been no crossing, so E_{i-1} occurs, and if $J_i = 0$, then $K_i = 1$ implies the occurrence of a all-hit event $K_i^0 = 1$. Thus

$$(14) \qquad K_i \leq K_i^0 \quad \text{on } S_i^c \cap E_{i-1} \cap \{J_i = 0\}.$$

We want the conditional probabilities to be prescribed; we shall show that we can take them to be

$$(15) \qquad \mathbf{P}[J_i = 1 \mid \mathscr{F}_{i-1}] = N k_0 d_0 2^{-Nj(i-1)}$$

and

$$(16) \qquad \mathbf{P}[K_i = 1 \mid \mathscr{F}_{i-1}^{+}] = \frac{\gamma}{2}.$$

We then also obtain an estimate on the conditional probability of a first split when there has been no close approach:

$$(17) \qquad \mathbf{P}[S_i \mid \mathscr{F}_{i-1}] \le NC_0 \frac{\varepsilon_0}{d_0} 2^N \rho^{k(i-1)}$$

on $S_{i-1} \cap E_{i-1}$, where $\rho < 1$ is as given above. This is a consequence of the fact that $j^0(i-1) \le j(i-1)$ and $k(i-1) - j(i-1) \le k^0(i-1)$ on $S_i \cap E_{i-1}$ and that we have the estimate of $j(i-1)$ in terms of $k(i-1)$ on E_{i-1}.

The construction follows. On each atom of $\mathscr{F}_{i-1}$ contained in $S_{i-1}^c \cap E_{i-1}$, the conditional probability of a record event $J_i^0 = 1$ in block i is bounded by $k_0 N d_0 2^{-Nj(i-1)}$. We define J_i in such a way that if there is a record event $J_i^0 = 1$ in i, then $J_i = 1$ and such that the conditional probability that $J_i = 1$ is exactly $k_0 N d_0 2^{-Nj(i-1)}$.

On the atoms of $\mathscr{F}_{i-1}$ contained in $S_{i-1} \cup E_{i-1}^c$, we define J_i in some arbitrary way to have the same conditional probability that $J_i = 1$.

Now we turn to the definition of K_i. On each atom of $\mathscr{F}_{i-1}$ contained in $S_{i-1}^c \cap E_{i-1}$, the conditional probability of a split S_i is bounded by $NC_0 2^N \varepsilon_0 / d_0$. Furthermore, the conditional probability that $J_i = 1$ is bounded by $k_0 N d_0$. We choose d_0 sufficiently small and then ε_0 sufficiently small that

$$(18) \qquad \gamma - NC_0 2^N \frac{\varepsilon_0}{d_0} - N k_0 d_0 > \frac{\gamma}{2}.$$

Thus the conditional probability of an all-hit event $K_i^0 = 1$ and no split and $J_i = 0$ exceeds $\gamma/2$. In particular, the conditional probability of an all-hit event $K_i^0 = 1$ given $S_i^c \cap E_{i-1} \cap \{J_i = 0\}$ exceeds $\gamma/2$.

We conclude that the conditional probability of an all-hit event $K_i^0 = 1$ exceeds $\gamma/2$ on each atom of $\mathscr{F}_{i-1}^+$ that is contained in $S_i^c \cap E_{i-1} \cap \{J_i = 0\}$. Thus we define K_i on these atoms so that the conditional probability is exactly equal to $\gamma/2$ and such that $K_i = 1$ implies an all-hit event $K_i^0 = 1$.

On the atoms of $\mathscr{F}_{i-1}^+$ contained in $S_i \cup E_{i-1}^c \cup \{J_i = 1\}$, we define K_i arbitrarily with the same conditional probability.

Now that we have completed the construction, we may complete the proof. Let $E = \bigcap_{i=0}^{\infty} E_i$ be the event that $j(i) \le c_0 k(i)$ for all i. We now show that $\mathbf{P}[E] > 0$.

By the strong law of large numbers $k(i)/i \to \gamma/2$ as $i \to \infty$ almost surely. By a similar argument for the Markov chain $j(i)$ we see that $j(i)/i \to 0$ as $i \to \infty$ almost surely. These two processes are independent. Therefore $j(i)/k(i) \to 0$ as $i \to \infty$ almost surely.

This implies that there is a random variable $I < \infty$ such that for $i \ge I$ we have $j(i)/k(i) \le c_0$ and $k(i)/i \le 1/2$. It is clear that there exists i_0 such that $\mathbf{P}[I = i_0] > 0$. On this set where $I = i_0$ there is a subset of strictly positive probability on which all the $i \le i_0$ with $K_i = 1$ precede the $i \le i_0$ with $J_i = 1$. This subset is contained in E, so $\mathbf{P}[E] > 0$.

Let $S = \bigcup_{i=0}^{\infty} S_i$ be the event that there is a split. We estimate $\mathbf{P}[S \cap E]$ as follows. The probability that E occurs and we have a split for the first time in block i is bounded by the probability that E_{i-1} occurs and we have a split for

the first time in block i. This is estimated by

$$(19) \qquad NC_0 2^N \frac{\varepsilon_0}{d_0} \mathbf{E}[\rho^{k(i-1)}] = NC_0 2^N \frac{\varepsilon_0}{d_0} \left(1 - \frac{\gamma}{2}(1-\rho)\right)^{i-1}.$$

If we sum from $i = 1$ to infinity we obtain a finite bound for $\mathbf{P}[S \cap E]$ proportional to ε_0.

We conclude by noting that $\mathbf{P}[S^c \cap E] = \mathbf{P}[E] - \mathbf{P}[S \cap E]$. Since $\mathbf{P}[E] > 0$ and is independent of ε_0 and $\mathbf{P}[S \cap E]$ is proportional to ε_0, there is an ε_0 that makes this strictly positive.

It follows that with strictly positive probability there are no splits and infinitely many all-hit events occurring at a linear rate. Each all-hit event with no split shrinks the distance between X_n and Y_n by a factor of $1 - \alpha$. This produces the exponential gluing. $\square$

We continue with the notation of the proof of Lemma 5. Begin the process in an arbitrary state. We want to define the occurrence of a catastrophe at some multiple of N steps. First we wait a time N to see if we get in a good starting configuration. If we do not, then we already have a catastrophe at the first unit. Otherwise we begin counting in multiples of N until the first i when we have an event $E_i^c \cup S_i$ of a close approach or a split. In this case we have a catastrophe at i units after the first unit.

The event of never having a catastrophe is the event of getting into a good starting configuration followed by the event $E \cup S^c$ of never having a close approach or a split.

LEMMA 6.   *There is a $\rho' > 0$ so that for arbitrary initial states the probability of never having a catastrophe exceeds ρ'. There are also constants α and $\beta < 1$ so that for arbitrary initial states*

$$(20) \qquad \mathbf{P}[\forall\, n, \delta(X_n, Y_n) \leq \alpha e^{-\beta n}] \geq \rho'.$$

PROOF.   According to Lemma 4, the probability of getting into a good starting configuration in N units is $\delta > 0$. By Lemma 5, the conditional probability of then getting glued is at least ρ. So the probability of both is at least $\rho' = \delta\rho$. $\square$

LEMMA 7.   *There are constants a and $b < 1$ such that for all initial states the probability of a catastrophe occurring at time index i is bounded by ab^i.*

PROOF.   We are only interested in large i, so we start the counting after the first unit. The event of having a catastrophe at i is

$$(21) \qquad (E_{i-1} \cap S_{i-1}^c) \cap (E_i^c \cup S_i) \subset (E_{i-1} \cap E_i^c) \cup (E_{i-1} \cap S_{i-1}^c \cap S_i).$$

In a previous argument we found a bound of the required form for the probability of the event of a split when there has been no close approach, which is the event $E_{i-1} \cap S_{i-1}^c \cap S_i$.

It remains to get a bound on the probability of a first close encounter, that is, of $E_{i-1} \cap E_i^c$. This event implies that $c_0 k(i) < j(i)$. We are interested in large i.

We use a large deviation bound for Bernoulli random variables [11]. The bound says that for i.i.d. random variables $Y_1, \ldots, Y_n$ that take the values one and zero and have mean p, we have the inequality

$$(22) \qquad \mathbf{P}\left[\frac{Y_1 + \cdots + Y_n}{n} \geq p + \eta\right] \leq \exp(-2\eta^2 n).$$

The same result holds for the probability of a deviation in the other direction, that is, of a proportion greater than or equal to $p - \eta$.

We can apply this to $k(i) = K_1 + \cdots + K_i$ with $p = \gamma/2$ and $\eta = \gamma/4$. We conclude that $\mathbf{P}[k(i)/i \leq \gamma/4] \leq \exp(-\gamma^2 i/8)$.

The application to $j(i) = J_1 + \cdots + J_i$ is a bit more complicated. In this case the conditional probability for $J_{i+1} = 1$ is $N k_0 d_0 / 2^{N j(i)}$. We want to estimate the probability that $j(i)/i \geq \xi$. If this is so, then there is an r such that for $i > r$ we have $N k_0 d_0 / 2^{N j(i)} \leq \xi/4$. Also there is an s such that for $i > s$ we have $j(r)/i \leq \xi/2$. So for i larger than s we must have $(j(i) - j(r))/i \geq \xi/2$. We can dominate the J_i for $i > r$ by independent random variables with probability $\xi/4$ of giving a 1. We conclude that for sufficiently large i the probability of $j(i)/i \geq \xi$ is bounded by $\exp(-\xi^2 i/8)$.

We can put the two estimates together by taking $\xi = c_0 \gamma/4$. If $c_0 k(i) < j(i)$ then either $k(i)/i \leq \gamma/4$ or $j(i)/i \geq \xi$. The probabilities of these events become exponentially small as i tends to infinity. $\square$

6. Eventual exponential attraction. The main result of the paper is the following theorem. Some of its consequences will be discussed in a later section.

THEOREM 2. *The self-organizing cluster process for a bounded region in two or more dimensions is superstable with exponential rate. Thus there is a constant $\beta > 0$ such that for every two initial states X_0 and Y_0 there is a random variable W with exponential tail such that*

$$(23) \qquad \delta(X_n, Y_n) \leq e^{-\beta(n-W)}$$

as $n \to \infty$.

PROOF. Break the time axis into intervals of length N, where N is the number in the starting lemma. Let W be the waiting time in units of N for a catastrophe. A catastrophe is defined as either not getting to a good starting configuration in the first interval of length N or of subsequently having an event $E_i^c \cup S_i$ of a close approach or a split.

According to Lemma 6 the probability that $W = \infty$ is strictly positive, greater than $\rho' > 0$. According to Lemma 7 the probability that $i \leq W < \infty$ is bounded by $a b^i$ where $b < 1$.

Now consider the independent waiting times $W_1, W_2, W_3, \ldots$ for subsequent catastrophes. Since these are independent, eventually there will be a first waiting time W_i with $W_i = \infty$. Thus we have waiting times $W_1, W_2, \ldots, W_L < \infty$ with $W_{L+1} = \infty$. The random variable in the statement of the theorem is $W = W_1 + \cdots + W_L$.

Conditioned on $L = i$ the waiting time random variables $W_1, \ldots, W_i$ are stochastically dominated by independent random variables with lagged geometric distribution. The random variable L is stochastically dominated by a random variable with a geometric distribution; in fact, the probability that $L \geq k$ is bounded below by $(1 - \rho')^k$. A geometric number of i.i.d. summands of geometric random variables is geometric. Therefore W is stochastically dominated by a sum of a geometric random variable with a multiple of L. Even though these need not be independent, their sum must have a distribution with exponential tail. $\square$

7. Ergodic properties.

THEOREM 3. *The self-organizing cluster process for a bounded region in dimension two or more has an invariant probability measure.*

PROOF. The construction in the starting lemma shows that there is a standard configuration of particles (in small open sets) that the process has a nonzero probability of hitting. This implies that the process is a Harris chain [5]. A recurrent Harris chain always has an invariant measure, and if the expected recurrence time is finite, it is a probability measure. The proof is completed by noting that the mechanism in the proof of the gluing lemma produces the configurations in the starting lemma at a steady rate, so the expected recurrence time is indeed finite. $\square$

COROLLARY 2. *The invariant probability measure for the self-organizing cluster process for a bounded region in dimension two or more is unique. The convergence of the probability to the invariant measure is exponentially fast in the Wasserstein metric.*

PROOF. Theorems 2 and 3 show that the process is superstable with exponential rate. The result follows from the general theory of superstable processes summarized in Theorem 1. $\square$

8. Acknowledgment. We thank David Plaehn for comments on the manuscript.

REFERENCES

[1] BOUTON, C. and PAGÉS, G. (1993). Self-organization and a.s. convergence of the one-dimensional Kohonen algorithm with nonuniformly distributed stimuli. *Stochastic Process. Appl.* **47** 249–274.

[2] BOUTON, C. and PAGÉS, G. (1994). Convergence in distribution of the one-dimensional Kohonen algorithms when the stimuli are not uniform. *Adv. in Appl. Probab.* **26** 80–103.

[3] COTTRELL, M. and FORT, J.-C. (1987). Étude d'un processus d'auto-organization. *Ann. Inst. H. Poincaré Probab. Statist.* **23** 1–20.

[4] DUDLEY, R. M. (1989). *Real Analysis and Probability*. Wadsworth and Brooks/Cole, Pacific Grove, CA.

[5] DURRETT, R. (1991). *Probability: Theory and Examples*. Wadsworth and Brooks/Cole, Pacific Grove, CA.

[6] ERWIN, E., OBERMAYER, K. and SCHULTEN, K. (1992). Self-organizing maps: ordering, convergence properties, and energy functions. *Biol. Cybernet.* **67** 47–55.

[7] KANGAS, J. A., KOHONEN, T. K. and LAAKSONEN, J. T. (1990). Variants of self-organizing maps. *IEEE Transactions on Neural Networks* **1** 93–99.

[8] KOHONEN, T. (1989). *Self-Organization and Associative Memory*, 3rd ed. Springer, Berlin.

[9] LI, X., GASTEIGER, J. and ZUPAN, J. (1993). On the topology distortion in self-organizing feature maps. *Biol. Cybernet.* **70** 189–198.

[10] LO, Z.-P. and BAVARIAN, B. (1991). On the rate of convergence in topology preserving neural networks. *Biol. Cybernet.* **65** 55–63.

[11] POLLARD, D. (1984). *Convergence of Stochastic Processes*. Springer, New York.

[12] RITTER, H., MARTINETZ, T. and SCHULTEN, K. (1992). *Neural Computation and Self-Organizing Maps*. Addison-Wesley, Reading, MA.

[13] TOLAT, V. V. (1990). An analysis of Kohonen's self-organizing maps using a system of energy functionals. *Biol. Cybernet.* **64** 155–164.

DEPARTMENT OF MATHEMATICS
OREGON STATE UNIVERSITY
CORVALLIS, OREGON 97331
E-MAIL: burton@math.oregon.edu

DEPARTMENT OF MATHEMATICS
UNIVERSITY OF ARIZONA
TUCSON, ARIZONA 85721
E-MAIL: faris@math.arizona.edu