

A Unifying Objective Function for Topographic Mappings

Geoffrey J. Goodhill

Sloan Center for Theoretical Neurobiology, Salk Institute for Biological Studies, La Jolla, CA 92037, U.S.A. and Georgetown Institute for Cognitive and Computational Sciences, Georgetown University Medical Center, Washington, DC 20007, U.S.A.

Terrence J. Sejnowski

The Howard Hughes Medical Institute, Salk Institute for Biological Studies, La Jolla, CA 92037, U.S.A. and Department of Biology, University of California San Diego, La Jolla, CA 92037, U.S.A.

Many different algorithms and objective functions for topographic mappings have been proposed. We show that several of these approaches can be seen as particular cases of a more general objective function. Consideration of a very simple mapping problem reveals large differences in the form of the map that each particular case favors. These differences have important consequences for the practical application of topographic mapping methods.

1 Introduction

The notion of a topographic mapping that takes nearby points in one space to nearby points in another appears in many domains, both biological and practical. A number of algorithms have been proposed that claim to construct such mappings (reviewed in Goodhill & Sejnowski, 1996). However, a fundamental problem with these claims, or the claim that a particular given mapping is topographic, is that the computational-level meaning of topography has not been formally defined. Given the wide variety of contexts in which topography or degrees of topography are discussed, it is unlikely that an application-independent definition would be generally useful. A more productive goal is to attempt to lay bare the assumptions behind different approaches to topographic mappings, and thus clarify the relationships among them. In this way, the hope is to make it easier to choose the appropriate approach for a particular problem.

In this article, we introduce an objective function we call the *C* measure and show that it has the capacity to unify several different approaches to topographic mapping. These approaches constitute particular instantiations of the functions from which the *C* measure is constructed. We then consider a simple mapping problem and explicitly optimize different versions of the *C* measure. It becomes apparent that the “optimally topographic” map can radically change depending on the particular measure employed.

2 The C measure

For the purposes of this article, we consider only bijective (1-1) mappings and a finite number of points in each space. Some mapping problems intrinsically have this form, and many others can be reduced to this by separating out a clustering or vector quantization step from the mapping step.

Consider an input space V_{in} and an output space V_{out} , each of which contains N points (see Figure 1). Let M be the mapping from points in V_{in} to points in V_{out} . We use the word *space* in a general sense: either or both of V_{in} and V_{out} may not have a geometric interpretation. Assume that for each space there is a symmetric similarity function that, for any given pair of points in the space, specifies by a nonnegative scalar value how similar (or dissimilar) they are. Call these functions F for V_{in} and G for V_{out} . Then we define a cost function C as follows,

$$C = \sum_{i=1}^N \sum_{j<i} F(i, j) G(M(i), M(j)), \quad (2.1)$$

where i and j label points in V_{in} , and $M(i)$ and $M(j)$ are their respective images in V_{out} . The sum is over all possible pairs of points in V_{in} . Since we have assumed that M is a bijection, it is therefore invertible, and C can equivalently be written as

$$C = \sum_{i=1}^N \sum_{j<i} F(M^{-1}(i), M^{-1}(j)) G(i, j), \quad (2.2)$$

where i and j label points in V_{out} , and M^{-1} is the inverse map. A good mapping is one with a high value of C . However, if one of F or G is given

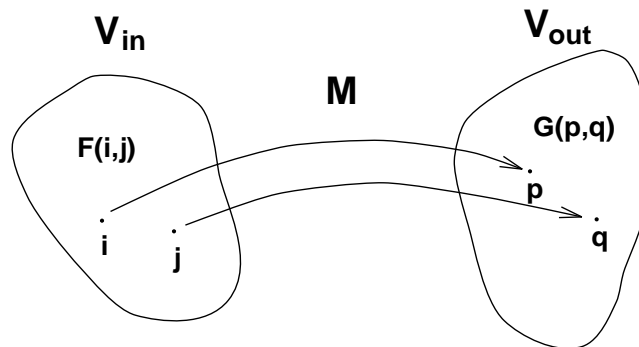


Figure 1: The mapping framework.

as a dissimilarity function (i.e., increasing with decreasing similarity), then a good mapping has a low value of C . How F and G are defined is problem specific. They could be Euclidean distances in a geometric space, some (possibly nonmonotonic) function of those distances, or they could just be given, in which case it may not be possible to interpret the points as lying in some geometric space. C measures the correlation between the F 's and the G 's. It is straightforward to show that if a mapping that preserves orderings exists, then maximizing C will find it. This is equivalent to saying that for two vectors of real numbers, their inner product is maximized over all permutations within the two vectors if the elements of the vectors are identically ordered, a proof of which can be found in Hardy, Littlewood, and Pólya (1934, p. 261).

3 Measures Related to C

Several objective functions for topographic mapping can be seen as versions of the C measure, given particular instantiations of the F and G functions.

3.1 Metric Multidimensional Scaling. Metric multidimensional scaling (metric MDS) is a technique originally developed in the context of psychology for representing a set of N entities (e.g., subjects in an experiment) by N points in a low- (usually two-) dimensional space. For these entities one has a matrix that gives the numerical dissimilarity between each pair of entities. The aim of metric MDS is to position points representing entities in the low-dimensional space so that the set of distances between each pair of points matches as closely as possible the given set of dissimilarities. The particular objective function optimized is the summed squared deviations of distances from dissimilarities. The original method was presented in Torgerson (1952); for reviews, see Shepard (1980) and Young (1987).

In terms of the framework presented earlier, the MDS dissimilarity matrix is F . Note that there may not be a geometric space of any dimensionality for which these dissimilarities can be represented by distances (for instance, if the dissimilarities do not satisfy the triangle inequality), in which case V_{in} does not have a geometric interpretation. V_{out} is the low-dimensional, continuous space in which the points representing entities are positioned, and G is Euclidean distance in V_{out} . Metric MDS selects the mapping M by adjusting the positions of points in V_{out} , which minimizes

$$\sum_{i=1}^N \sum_{j<i} (F(i, j) - G(M(i), M(j)))^2. \quad (3.1)$$

Under our assumptions, this objective function is identical to the C mea-

sure. Expanding out the square in equation 3.1 gives

$$\sum_{i=1}^N \sum_{j<i} (F(i, j)^2 + G(M(i), M(j))^2 - 2F(i, j)G(M(i), M(j))). \quad (3.2)$$

The last term is twice the C measure. The entries in F are fixed, so the first term is independent of the mapping. In metric MDS, the sum over the G 's varies as the entities are moved in the output space. If one instead considers the case where the positions of the points in V_{out} are fixed, and the problem is to find the assignment of entities to positions that minimize equation 3.1, then the sum over the G 's is also independent of the map. In this case the metric MDS measure becomes exactly equivalent to C .

3.2 Minimal Wiring. In minimal wiring (Mitchison & Durbin, 1986; Durbin & Mitchison, 1990), a good mapping is defined as one that maps points that are nearest neighbors in V_{in} as close as possible in V_{out} , where closeness in V_{out} is measured by, for instance, Euclidean distance raised to some power. The motivation here is the idea that it is often useful in sensory processing to perform computations that are local in some space of input features V_{in} . To do this in V_{out} (e.g., the cortex) the images of neighboring points in V_{in} need to be connected; the dissimilarity function in V_{out} is intended to capture the cost of the wire (e.g., axons) required to do this. Minimal wiring is equivalent to the C measure for

$$F(i, j) = \begin{cases} 1 & : i, j \text{ neighboring} \\ 0 & : \text{otherwise} \end{cases}$$

$$G(M(i), M(j)) = \|M(i) - M(j)\|^p.$$

For the cases of one- or two-dimensional square arrays investigated in Mitchison and Durbin (1986) and Durbin and Mitchison (1990), neighbors are taken to be just the two or four adjacent points in the array, respectively.

3.3 Minimal Path Length. In this scheme, a good map is one such that, in moving between nearest neighbors in V_{out} , one moves the least possible distance in V_{in} . This is, for instance, the mapping required to solve the traveling salesman problem (TSP), where V_{in} is the distribution of cities and V_{out} is the one-dimensional tour. This goal is implemented by the elastic net algorithm (Durbin & Willshaw, 1987; Durbin & Mitchison, 1990; Goodhill & Willshaw, 1990), which measures dissimilarity in V_{in} by squared distances:

$$F(i, j) = \|\mathbf{v}_i - \mathbf{v}_j\|^2$$

$$G(p, q) = \begin{cases} 1 & : p, q \text{ neighboring} \\ 0 & : \text{otherwise} \end{cases}$$

where \mathbf{v}_k is the position of point k in V_{in} (we have only considered here the regularization term in the elastic net energy function, which also includes a term matching input points to output points). Thus, minimal wiring and minimal path length are symmetrical cases under equation 2.1, with respect to exchange of V_{in} and V_{out} . Their relationship is discussed further in Durbin & Mitchison (1990), where the abilities of minimal wiring and minimal path length are compared with regard to reproducing the structure of the map of orientation selectivity in the primary visual cortex (see also Mitchison, 1995).

3.4 The Approach of Jones, Van Sluyters, and Murphy (1991). Jones, Van Sluyters, and Murphy (1991) investigated the effect of the shape of the cortex (V_{out}) relative to the lateral geniculate nuclei (V_{in}) on the overall pattern of ocular dominance columns in the cat and monkey, using an optimization approach. They desired to keep both neighboring cells in each LGN (as defined by a hexagonal array), and anatomically corresponding cells between the two LGNs, nearby in the cortex (also a hexagonal array). Their formulation of this problem can be expressed as a maximization of C when

$$F(i, j) = \begin{cases} 1 & : i, j \text{ neighboring, corresponding} \\ 0 & : \text{otherwise} \end{cases}$$

and

$$G(p, q) = \begin{cases} 1 & : p, q \text{ first or second nearest neighbors} \\ 0 & : \text{otherwise.} \end{cases}$$

For two-dimensional V_{in} and V_{out} they found a solution such that if $F(i, j) = 1$, then $G(M(i), M(j)) = 1, \forall i, j$. Alternatively this problem could be expressed as a minimization of C when $G(p, q)$ is the stepping distance (see below) between positions in the V_{out} array. They found this gave appropriate behavior for the problem addressed.

3.5 Minimal Distortion. Luttrell and Mitchison have introduced mapping functionals for the continuous case. Under appropriate assumptions to reduce them to the discrete case, these are equivalent to C . Luttrell (1990, 1994) defined a minimal distortion principle that can be interpreted as a measure of mapping quality. He defined "distortion" D as

$$D = \int d\mathbf{x} P(\mathbf{x}) d\{\mathbf{x}, \mathbf{x}'[\mathbf{y}(\mathbf{x})]\}.$$

\mathbf{x} and \mathbf{y} are vectors in the input and output spaces, respectively. \mathbf{y} is the map in the input to output direction, \mathbf{x}' is the (in general different) map back again, and $P(\mathbf{x})$ is the probability of occurrence of \mathbf{x} . \mathbf{x}' and \mathbf{y} are suitably adjusted to minimize D . An augmented version of D includes additive noise

in the output space,

$$D = \int d\mathbf{x} P(\mathbf{x}) \int d\mathbf{n} \pi(\mathbf{n}) d\{\mathbf{x}, \mathbf{x}'[\mathbf{y}(\mathbf{x}) + \mathbf{n}]\},$$

where $\pi(\mathbf{n})$ is the probability density of the noise vector \mathbf{n} . Intuitively, the aim is now to find the forward and backward maps so that the reconstructed value of the input vector is as close as possible to its original value after being corrupted by noise in the output space. In Luttrell (1990), the d function was taken to be $\{\mathbf{x} - \mathbf{x}'[\mathbf{y}(\mathbf{x}) + \mathbf{n}]\}^2$. In this case, Luttrell showed that the minimal distortion measure can be differentiated to produce a learning rule that is almost the self-organizing map rule of Kohonen (1982).

For the discrete version of this, it is necessary to assume that appropriate quantization has occurred so that $\mathbf{y}(\mathbf{x})$ defines a 1-1 map, and $\mathbf{x}'(\mathbf{y})$ defines the same map in the opposite direction. In this case the minimal distortion measure becomes equivalent to the C measure, with $F(i, j)$ the squared Euclidean distance between the positions of vectors i and j (assuming these to lie in a geometric space) and $G(p, q)$ the noise process in the output space (generally assumed gaussian).

The minimal distortion principle was generalized in Mitchison (1995) by allowing the noise process and reconstruction error to take arbitrary forms. For instance, they can be reversed, so that F is a gaussian and G is Euclidean distance. In this case, the measure can be interpreted as a version of the minimal wiring principle, establishing a connection between minimal distortion (and hence the SOM algorithm) and minimal wiring. This identification also yields a self-organizing algorithm similar to the SOM for solving minimal wiring problems, the properties of which are briefly explored in Mitchison (1995).

Luttrell (1994) generalized minimal distortion to allow a probabilistic match between points in the two spaces, which in the discrete case can be expressed as

$$D = \sum_i \sum_j F(i, j) \sum_k P(k|i)P(j|k),$$

where D is the distortion to be minimized, F is Euclidean distance in the input space, $P(k|i)$ is the probability that output state k occurs given that input state i occurs, and $P(j|k)$ is the corresponding Bayes inverse probability that input state j occurs given that output state k occurs. This reduces to C in the special case that $P(k|i) = P(i|k)$, which is true for bijections.

3.6 Dynamic Link Matching. Bienenstock and von der Malsburg (1987a, 1987b) considered position-invariant pattern recognition as a problem of graph matching. Their formalism consists of two sets of nodes, one with links expressing neighborhood relationships given by $F(i, j)$ and one with analogous links given by $G(p, q)$. The aim is to find the bijective mapping between the two sets of nodes that optimally preserves the link structure. By

approximating the Hamiltonian for a particular type of spin neural network, they derived an objective function H that in our notation (taking $p = M(i)$, $q = M(j)$) can be written as

$$H = \sum_{i,j=1}^N \sum_{p,q=1}^N F(i, j) G(p, q) w_{ip} w_{jq} + \gamma \left(\sum_{i=1}^N \left(\sum_{p=1}^N w_{ip} - a \right)^2 + \sum_{p=1}^N \left(\sum_{i=1}^N w_{ip} - a \right)^2 \right). \quad (3.3)$$

Here γ and a are constants, and w_{ip} is the strength of connection between node i and node p . The first term is a generalized version of the C measure, which allows a weighted match between all pairs of nodes i and p . The second term enforces the regularization constraint that the sum of the weights from or to each node is equal to a . When $a = 1$ and γ is large, the minima of H are clearly the same as the maxima of C . More recent work has investigated its potential as a general objective function for topographic mappings (Wiskott & Sejnowski, 1997).

4 Application to a Simple Mapping Problem

To show the range of optimal maps that the C measure implies for different F and G , we consider the very simple case of mapping between a regular 10×10 square array of points (V_{in}) and a regular 1×100 linear array of points (V_{out}). This is a paradigmatic example of dimension reduction (also used in Durbin & Mitchison, 1990 and Mitchison, 1995). It has the virtue that solutions are easily represented (see Figures 2 and 3) and easily compared with intuition. Calculation of the global optimum for this problem by exhaustive search is clearly impractical, since there are of the order of $100!$ possible mappings. Instead we employ simulated annealing (Kirkpatrick, Gelatt, & Vecchi, 1983) to find a solution close to optimal. The parameters used are given in the legend to Figure 2.

Figure 2 shows the best maps obtained by this method for the minimal path, minimal wiring, and metric MDS versions of the C measure. Also shown for comparison is a solution to this problem produced by the elastic net algorithm, taken from Durbin & Mitchison (1990). An optimal map for the minimal path version (A) is one with no diagonal segments. Our simulated annealing algorithm found one with three diagonal segments and a cost of 100.243, 1.3 percent larger than the optimal of 99.0. The optimal map for the minimal wiring measure has been explicitly calculated (Mitchison & Durbin, 1986), and has a cost of 914.0. Our algorithm found the map shown in (B), which is similar in shape and has a cost of 917.0 (0.3 percent larger than the optimal value). These results confirm that the simulated annealing algorithm employed can find close to optimal solutions. The cost for the

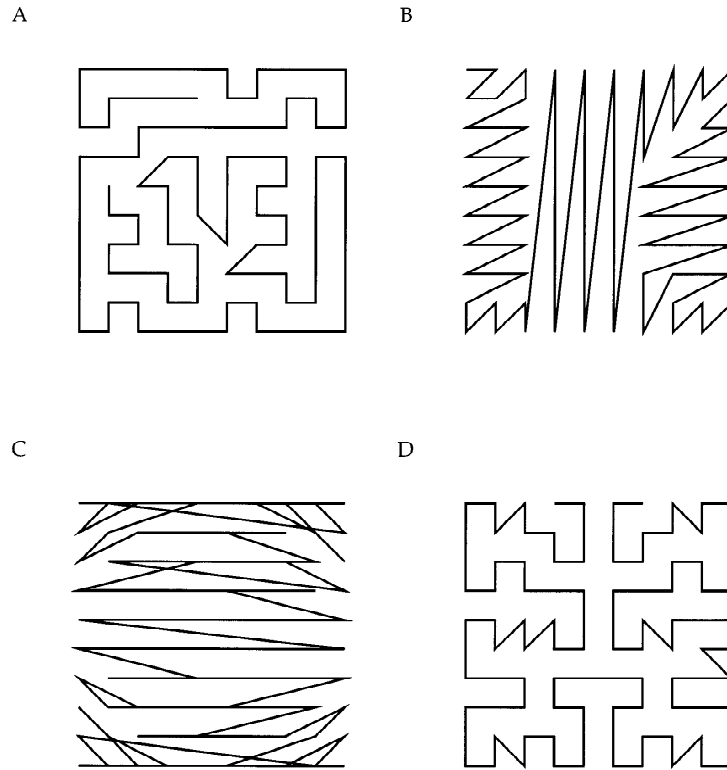


Figure 2: Maps found by optimizing using simulated annealing. (A) Minimal path length solution. (B) Minimal wiring solution. (C) Metric MDS solution. (D) Map found by the elastic net algorithm (see Durbin & Mitchison, 1990). Parameters used for simulated annealing were as follows (for further explanation, see van Larhoven & Aarts, 1987). Initial map: random. Move set: pairwise interchanges. Initial temperature: $3 \times$ the mean energy difference over 10,000 moves. Cooling schedule: exponential. Cooling rate: 0.998. Acceptance criterion: 1000 moves at each temperature. Upper bound: 10,000 moves at each temperature. Stopping criterion: zero acceptances at upper bound.

metric MDS map (C) is 6,352,324. The illusion of multiple ends to the line is due to the map frequently doubling back on itself. For instance, in the fourth row up the square, the horizontal component for neighboring points in the line progresses in the order 5, 6, 4, 7, 3, 8, 2, 9, 10. Local discontinuity arises because this measure takes into account neighborhood preservation at all scales: local continuity is not privileged over global continuity. The costs of the map produced by the elastic net (D) are 102.3 for the minimal

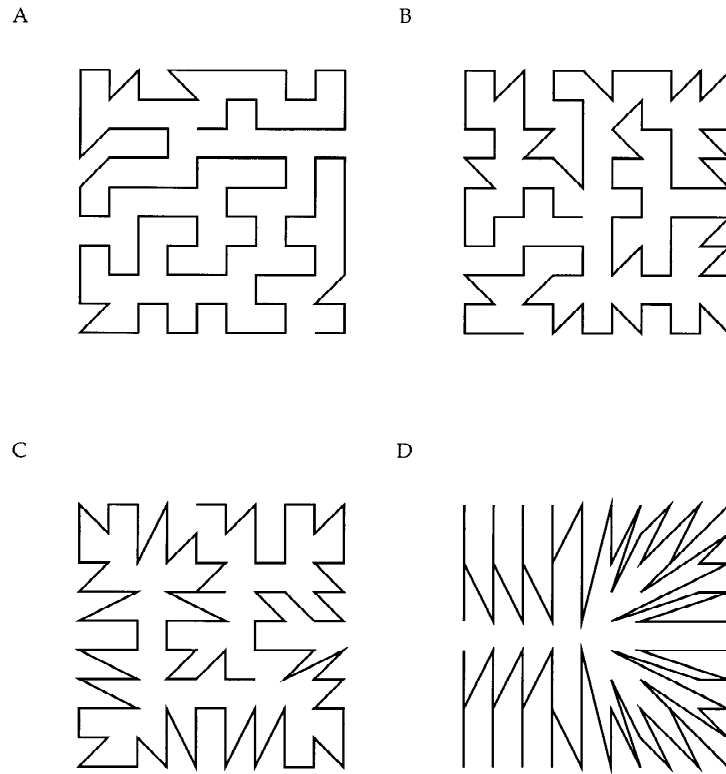


Figure 3: Minimal distortion solutions. (A) $\sigma = 1.0$, cost = 43.3. (B) $\sigma = 2.0$, cost = 214.7. (C) $\sigma = 4.0$, cost = 833.2. (D) $\sigma = 20.0$, cost = 18467.1. Simulated annealing parameters as in Figure 2.

path function, 1140 for the minimal wiring function, and 6,455,352 for the metric MDS function. In each case this is a higher cost than for the maps shown in (A), (B), and (C), respectively.

Figure 3 shows close-to-optimal maps for the minimal distortion version of the C measure (F is squared Euclidean distance in the square, G is given by e^{-d^2/σ^2} where d is Euclidean distance along the line), for varying σ . For small σ , the solution resembles the minimal path optimum of Figure 2A, since the contribution from more distant neighbors on the line compared to nearest neighbors is negligible. However, as σ increases, the map changes form. Local continuity becomes less important compared to continuity at the scale of σ , the map becomes spikier, and the number of large-scale folds in the map gradually decreases until at $\sigma = 20$ there is just one. This

last map also shows some of the frequent doubling back behavior seen in Figure 2C.¹

5 Discussion

5.1 Types of Mappings. Several qualitatively different types of map are apparent in Figures 2 and 3. For the metric MDS measure, Figure 2C, the line progresses through the square by a series of locally discontinuous jumps along one axis of the square and a comparatively smooth progression in the orthogonal direction. One consequence is that nearby points in the square never map too far apart on the line. For the minimal distortion measure with $\sigma = 20$ (see Figure 3D), the strategy is similar except that there is now one fold to the map. This decreases the size of local jumps in following the line through the square, at the cost of introducing a seam across which nearby points in the square map a very long way apart on the line. For the minimal distortion measure with small σ (see Figures 3A–C), the strategy is to introduce several folds. This gives strong local continuity, at the cost that now many points that are nearby in the square map to points that are far apart on the line.

What do these maps suggest about which measures are most appropriate for different problems? If it is desired that generally nearby points should always map to generally nearby points as much as possible in both directions, and one is not concerned about very local continuity, then something like the MDS measure is useful. This may be appropriate for some data visualization applications where the overall structure of the map is more important than the fine detail. If, on the other hand, one desires a smooth progression through the output space to imply a smooth progression through the input space, one should choose something like minimal path or minimal distortion for small σ . This may be important for data visualization where it is believed the data actually lie on a lower-dimensional manifold in the high-dimensional space. However, an important weakness of this representation is that some neighborhood relationships between points in the input space may be completely lost in the resulting representation. For understanding the structure of cortical mappings, self-organizing algorithms that optimize objectives of this type have proved useful (Durbin & Mitchison, 1990; Obermayer, Blasdel, & Schulten, 1992). However, very few other objectives have been applied to this problem, so it is still an open question which are most appropriate. The type of map seen in Figure 3D has not been much investigated. There may be some applications for which such maps are worthwhile, perhaps in a neurobiological context for understand-

¹ We have also explicitly optimized the objective functions proposed by Sammon (1969) and Bezdek and Pal (1995). The former produces a map similar to that of Figure 2C, while the latter resembles Figure 3D (for more details, see Goodhill & Sejnowski, 1996).

ing why different input variables are sometimes mapped into different areas rather than interdigitated in the same area.

5.2 Relation of C to Quadratic Assignment Problems. Formulating neighborhood preservation in terms of the C measure sets it within the well-studied class of quadratic assignment problems (QAPs). These occur in many different practical contexts and take the form of finding the minimal or maximal value of an equation similar to C (see Burkard, 1984, for a general review and Lawler, 1963, and Finke, Burkard, & Rendl, 1987, for more technical discussions). An illustrative example is the optimal design of typewriter keyboards (Burkard, 1984). If $F(i, j)$ is the average time it takes a typist to sequentially press locations i and j on the keyboard, while $G(p, q)$ is the average frequency with which letters p and q appear sequentially in the text of a given language (note that in this example F and G are not necessarily symmetrical), then the keyboard that minimizes average typing time will be the one that minimizes the product

$$\sum_{i=1}^N \sum_{j=1}^N F(i, j) G(M(i), M(j))$$

(see equation 2.1), where $M(i)$ is the letter that maps to location i . The substantial amount of theory developed for QAPs is directly applicable to the C measure. As a concrete example, QAP theory provides several different ways of calculating bounds on the minimum and maximum values of C for each problem. This could be very useful for the problem of assessing the quality of a map relative to the unknown best possible. One particular case is the eigenvalue bound (Finke et al., 1987). If the eigenvalues of symmetric matrices F and G are λ_i and μ_i , respectively, such that $\lambda_1 \leq \lambda_2 \leq \dots \leq \lambda_n$ and $\mu_1 \geq \mu_2 \geq \dots \geq \mu_n$, then it can be shown that $\sum_i \lambda_i \mu_i$ gives a lower bound on the value of C .

QAPs are in general known to be of complexity NP-hard. A large number of algorithms for both exact and heuristic solution have been studied (see, e.g., Burkard, 1984; the references in Finke et al., 1987; and Simić, 1991). However, particular instantiations of F and G may allow efficient algorithms for finding good local optima, or alternatively may beset C with many bad local optima. Such considerations provide an additional practical constraint on what choices of F and G are most appropriate.

5.3 Many-to-One Mappings. In many practical contexts, there are many more points in V_{in} than V_{out} , and it is necessary also to specify a many-to-one mapping from points in V_{in} to N exemplar points in V_{in} , where $N =$ number of points in V_{out} . It may be desirable to do this adaptively while simultaneously optimizing the form of the map from V_{in} to V_{out} . For instance, shifting a point from one cluster to another may increase the clustering cost, but by moving the positions of the cluster centers decrease the sum of this

and the continuity cost. The elastic net, for instance, trades off these two contributions explicitly with a ratio that changes during the minimization, so that finally each cluster contains only one point and the continuity cost dominates (Durbin & Willshaw, 1987; for discussions, see Simić, 1990, and Yuille, 1990). The self-organizing map of Kohonen (1982) trades off these two contributions implicitly. Luttrell (1994) discusses allowing each point in the input space to map to many in the output space in a probabilistic manner, and vice-versa. The H function (Bienenstock & von der Malsburg, 1987a, 1987b) offers an alternative way of introducing a weighted match between input and output points.

Acknowledgments

We thank Steve Finch for very useful discussions at an earlier stage of this work. We are grateful to Graeme Mitchison, Steve Luttrell, Hans-Ulrich Bauer, and Laurenz Wiskott for helpful discussions and comments on the manuscript. Research was supported by the Sloan Center for Theoretical Neurobiology at the Salk Institute and the Howard Hughes Medical Institute.

References

- Bezdek, J. C., & Pal, N. R. (1995). An index of topological preservation for feature extraction. *Pattern Recognition* 28:381–391.
- Bienenstock, E., & von der Malsburg, C. (1987a). A neural network for the retrieval of superimposed connection patterns. *Europhys. Lett.* 3:1243–1249.
- Bienenstock, E., & von der Malsburg, C. (1987b). A neural network for invariant pattern recognition. *Europhys. Lett.* 4:121–126.
- Burkard, R. E. (1984). Quadratic assignment problems. *Europ. J. Oper. Res.* 15:283–289.
- Durbin, R., & Mitchison, G. (1990). A dimension reduction framework for understanding cortical maps. *Nature* 343:644–647.
- Durbin, R., & Willshaw, D. J. (1987). An analogue approach to the traveling salesman problem using an elastic net method. *Nature* 326:689–691.
- Finke, G., Burkard, R. E., & Rendl, F. (1987). Quadratic assignment problems. *Annals of Discrete Mathematics* 31:61–82.
- Goodhill, G. J., & Sejnowski, T. J. (1996). Quantifying neighbourhood preservation in topographic mappings. In *Proceedings of the 3rd Joint Symposium on Neural Computation* (pp. 61–82). Pasadena: California Institute of Technology. Available from <http://www.cnl.salk.edu/~geoff/>.
- Goodhill, G. J., & Willshaw, D. J. (1990). Application of the elastic net algorithm to the formation of ocular dominance stripes. *Network* 1:41–59.
- Hardy, G. H., Littlewood, J. E., & Pólya, G. (1934). *Inequalities*. Cambridge: Cambridge University Press.
- Jones, D. G., Van Sluyters, R. C., & Murphy, K. M. (1991). A computational model for the overall pattern of ocular dominance. *J. Neurosci.* 11:3794–3808.

- Kirkpatrick, S., Gelatt, C. D., & Vecchi, M. P. (1983). Optimization by simulated annealing. *Science* 220:671-680.
- Kohonen, T. (1982). Self-organized formation of topologically correct feature maps. *Biol. Cybern.* 43:59-69.
- Lawler, E. L. (1963). The quadratic assignment problem. *Management Science* 9:586-599.
- Luttrell, S. P. (1990). Derivation of a class of training algorithms. *IEEE Trans. Neural Networks* 1:229-232.
- Luttrell, S. P. (1994). A Bayesian analysis of self-organizing maps. *Neural Computation* 6:767-794.
- Mitchison, G. (1995). A type of duality between self-organizing maps and minimal wiring. *Neural Computation* 7:25-35.
- Mitchison, G., & Durbin, R. (1986). Optimal numberings of an $N \times N$ array. *SIAM J. Alg. Disc. Meth.* 7:571-581.
- Obermayer, K., Blasdel, G. G., & Schulten, K. (1992). Statistical-mechanical analysis of self-organization and pattern formation during the development of visual maps. *Phys. Rev. A* 45:7568-7589.
- Sammon, J. W. (1969). A nonlinear mapping for data structure analysis. *IEEE Trans. Comput.* 18:401-409.
- Shepard, R. N. (1980). Multidimensional scaling, tree-fitting and clustering. *Science* 210:390-398.
- Simić, P. D. (1990). Statistical mechanics as the underlying theory of "elastic" and "neural" optimizations. *Network* 1:89-103.
- Simić, P. D. (1991). Constrained nets for graph matching and other quadratic assignment problems. *Neural Computation* 3:268-281.
- Torgerson, W. S. (1952). Multidimensional scaling, I: Theory and method. *Psychometrika* 17:401-419.
- van Laarhoven, P. J. M., & Aarts, E. H. L. (1987). *Simulated annealing: Theory and applications*. Dordrecht: Reidel.
- Wiskott, L., & Sejnowski, T. J. (1997). Objective functions for neural map formation (Tech. Rep. INC-9701). Institute for Neural Computation.
- Young, F. W. (1987). *Multidimensional scaling: History, theory, and applications*. Hillsdale, NJ: Erlbaum.
- Yuille, A. L. (1990). Generalized deformable models, statistical physics, and matching problems. *Neural Computation* 2:1-24.