(4)

REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER	2. GOVT ACCESSION NO	3. RECIPIENT'S CATALOG NUMBER
AIM 1167		

4. TITLE (and Subtitle)

Extensions of a Theory of Networks for Approximation and Learning: Dimensionality Reduction and Clustering

5. TYPE OF REPORT & PERIOD COVERED

memorandum

6. PERFORMING ORG. REPORT NUMBER

7. AUTHOR(s)

Tomaso Poggio & Federico Girosi

8. CONTRACT OR GRANT NUMBER(s)

S1-801534-2
DACA76-85-C0010
N00014-85-K-0124

9. PERFORMING ORGANIZATION NAME AND ADDRESS

Artificial Intelligence Laboratory
545 Technology Square
Cambridge, MA 02139

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS

11. CONTROLLING OFFICE NAME AND ADDRESS

Advanced Research Projects Agency
1400 Wilson Blvd.
Arlington, VA 22209

12. REPORT DATE

13. NUMBER OF PAGES

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)

Office of Naval Research
Information Systems
Arlington, VA 22217

15. SECURITY CLASS. (of this report)

UNCLASSIFIED

15a. DECLASSIFICATION/DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Distribution is unlimited

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

learning networks
regularization

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Learning an input-output mapping from a set of examples, of the type that many neural networks have been constructed to perform, can be regarded as synthesizing an approximation of a multi-dimensional function. From this point of view, this form of learning is closely related to regularization theory. The theory developed in Poggio and Girosi (1989) shows the equivalence between regularization and a class of three-

(continued on back)

DD FORM 1 JAN 73 1473 EDITION OF 1 NOV 65 IS OBSOLETE S/N 0102-014-6601

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

Block 20 continued:

layer networks that we call regularization networks or Hyper Basis Functions. These networks are not only equivalent to generalized splines, but are also closely related to the classical Radial Basis Functions used for interpolation tasks and to several pattern recognition and neural network algorithms. In this note, we extend the theory by defining a general form of these networks with two sets of modifiable parameters in addition to the coefficients c_α: ***moving centers*** and ***adjustable norm-weights.*** Moving the centers is equivalent to task-dependent clustering and changing the norm weights is equivalent to task-dependent dimensionality reduction.

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
ARTIFICIAL INTELLIGENCE LABORATORY
and
CENTER FOR BIOLOGICAL INFORMATION PROCESSING
WHITAKER COLLEGE

A.I. Memo No.1167
C.B.I.P. Paper No. 44

April 1990

Extensions of a Theory of Networks for Approximation and Learning: dimensionality reduction and clustering

Tomaso Poggio and Federico Girosi

Abstract

Learning an input-output mapping from a set of examples, of the type that many neural networks have been constructed to perform, can be regarded as synthesizing an approximation of a multi-dimensional function. From this point of view, this form of learning is closely related to regularization theory. The theory developed in Poggio and Girosi (1989) shows the equivalence between regularization and a class of three-layer networks that we call regularization networks or Hyper Basis Functions. These networks are not only equivalent to generalized splines, but are also closely related to the classical Radial Basis Functions used for interpolation tasks and to several pattern recognition and neural network algorithms. In this note, we extend the theory by defining a general form of these networks with two sets of modifiable parameters in addition to the coefficients c_α: *moving centers* and *adjustable norm-weights*. Moving the centers is equivalent to task-dependent clustering and changing the norm weights is equivalent to task-dependent dimensionality reduction.

This paper describes research done within the Center for Biological Information Processing, in the Department of Brain and Cognitive Sciences, and at the Artificial Intelligence Laboratory. This research is sponsored by a grant from the Office of Naval Research (ONR), Cognitive and Neural Sciences Division; by the Artificial Intelligence Center of Hughes Aircraft Corporation (S1-801534-2). Support for the A. I. Laboratory's artificial intelligence research is provided by the Advanced Research Projects Agency of the Department of Defense under Army contract DACA76–85–C–0010, and in part by ONR contract N00014-85-K-0124.

1 Introduction

In previous papers (Poggio and Girosi, 1989, 1990) we have shown the equivalence between regularization and a class of three-layer networks that we called regularization networks and that are related to the classical interpolation technique of Radial Basis Functions.

Let $S = \{(\mathbf{x}_i, y_i) \in R^n \times R | i = 1, ...N\}$ be a set of data that we want to approximate by means of a function f. The regularization approach (Tikhonov, 1963; Tikhonov and Arsenin, 1977; Morozov, 1984; Bertero, 1986) selects the function f that solves the variational problem of minimizing the functional

$$H[f] = \sum_{i=1}^{N} (y_i - f(\mathbf{x}_i))^2 + \lambda \|Pf\|^2 \tag{1}$$

where P is a constraint operator (usually a differential operator), $\|\cdot\|$ is a norm on the function space to whom Pf belongs (usually the L^2 norm) and λ is a positive real number, the so called *regularization parameter*. The structure of the operator P, that is called "stabilizer", embodies the a priori knowledge about the solution, and therefore depends on the nature of the particular problem that has to be solved (for instance, it is not needed in the case of P corresponding to a Gaussian or bell-shaped Green's function). We have shown (Poggio and Girosi, 1989) that the solution of the variational problem (1) has the following simple form:

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i G(\mathbf{x}; \mathbf{x}_i) + p(\mathbf{x})$$

where $G(\mathbf{x})$ is the Green's function (Stakgold, 1979) of the self-adjoint differential operator $\hat{P}P$, $\hat{P}$ being the adjoint operator of P, $p(\mathbf{x})$ is a linear combination of functions that span the null space of P, and the coefficients c_i satisfy a linear system of equations that depend on the N "examples", i.e. the data to be approximated. The form of the term $p(\mathbf{x})$ depends on the stabilizer that has been chosen and on the boundary conditions, and therefore on the particular problem that has to be solved. For this reason, and since its inclusion does not modify the main conclusions, we will disregard it in the following. If P is an operator with radial symmetry, the Green's function G is radial and therefore the approximating function becomes:

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i G(\|\mathbf{x} - \mathbf{x}_i\|^2), \tag{2}$$

which is a sum of radial functions, each with its *center* $\mathbf{x}_i$ on a distinct data point. Thus the number of radial functions, and corresponding centers, is the same as the number of examples.

In this note we indicate how to extend the technique into three natural directions:

1. The computation of a solution of the form (2) has a complexity (number of radial functions) that is independent of the dimensionality of the input space but is on the order of the dimensionality of the training set (number of examples), which is usually high. We show how to justify in terms of the regularizationan framework an approximation of equation (2) in which the number of centers is much smaller than the number of examples and the positions of the centers are modified during learning (Poggio and Girosi, 1989). The key idea is to consider a specific form of an approximation to the solution of the standard regularization problem. Moving centers are equivalent to the free knots of nonlinear splines. In the context of networks they were first suggested as a potentially useful heuristics by Broomhead and Lowe (1988) and used by Moody and Darken (1989).

2. It is natural to try to extend the form of the solution (2) by considering the superposition of different types of Green's functions (Poggio and Girosi, 1989, 1990a) (for example basis functions of different scales). This extension is natural within the framework of regularization (and has a direct Bayesian interpretation) by considering a more general functional than equation (1) containing several stabilizers. We will show how the well-defined but underconstrained variational problem associated with the new functional can be transformed into an over-constrained problem.

3. In equation (2) the norm $\|\mathbf{x} - \mathbf{x}_i\|$ may be considered as a *weighted norm*

$$\|\mathbf{x} - \mathbf{x}_i\|_{\mathbf{W}}^2 = (\mathbf{x} - \mathbf{x}_i)^T \mathbf{W}^T \mathbf{W} (\mathbf{x} - \mathbf{x}_i)$$

where $\mathbf{W}$ is a square matrix and the superscript T indicates the transpose. In the simple case of diagonal $\mathbf{W}$ the diagonal elements w_{ii} assign a specific weight to each input coordinate, and the standard Euclidean norm is obtained when $\mathbf{W}$ is set to the identity matrix. They play a critical role whenever different types of inputs are present. We will show how the weighted norm idea can be derived from a slightly more general functional than equation (1). The associated variational problem is well-defined but underconstrained; it can be transformed into an overconstrained problem by using a certain approximation technique.

We call *Hyper Basis Functions*, in short *HyperBFs*, the most general form of regularization networks based on these three extensions.

2 Moving Centers

The solution given by standard regularization theory to the approximation problem can be very expensive in computational terms when the number of examples is very high. The computation of the coefficients of the expansion can become then a very time consuming operation: its complexity grows polynomially with N, (roughly as N^3) since an $N \times N$ matrix has to be inverted. In addition, the probability of ill-conditioning is higher for larger and larger matrices (it grows like N^3 for a $N \times N$ uniformly distributed random matrix) (Demmel, 1987). We now show a way to reduce the complexity of the problem, introducing an approximation to the regularized solution. While the exact regularization solution is equivalent to generalized splines with *fixed* knots, the approximated solution is equivalent to generalized splines with *free* knots.

2.1 An approximation to the regularization solution

A standard technique, sometimes known as Galerkin's method, that has been used to find approximate solutions of variational problems, is to expand the solution on a finite basis. The approximated solution $f^*(\mathbf{x})$ has then the following form:

$$f^*(\mathbf{x}) = \sum_{i=1}^{n} c_i \phi_i(\mathbf{x}) \tag{3}$$

where $\{\phi_i\}_{i=1}^n$ is a set of linearly independent functions (Mikhlin, 1965). The coefficients c_i are usually found according to some rule that guarantees a minimum deviation from the true solution. In the case of standard regularization, when the functional to minimize is given by equation (1), this method gives the *exact* solution if n is equal to the numer of data points N, and $\{\phi_i\}_{i=1}^n = \{G(\mathbf{x};\mathbf{x}_i)\}_{i=1}^N$, where G is the Green's function of the operator $\hat{P}P$. In this case the unknown coefficients of the expansion (3) can be obtained in a simple way by substituting expansion (3) in the regularization functional (1), that becomes a *function* $H[f^*] = H^*(c_1,\ldots,c_N)$, and then by minimizing $H[f^*]$ with respect to the coefficients, that is by setting:

$$\frac{\partial H[f^*]}{\partial c_i} = 0 \quad i = 1,\ldots,N. \tag{4}$$

It can be easily shown (Poggio and Girosi, 1989) that, if the Green's function vanishes on the boundary of the region that is considered, the set of equations (4) is a linear system whose solution gives the standard regularization coefficients. In more general cases the basis $\{\phi_i\}_{i=1}^n$ should be enlarged, to include terms that generate the null space of P, in order to obtain the correct solution. For simplicity, we disregard these terms in the following, since they do not change the main conclusions. A natural approximation to the exact solution will be then of the form:

$$f^*(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\mathbf{x};\mathbf{t}_\alpha) \tag{5}$$

where the parameters $\mathbf{t}_\alpha$, that we call "centers", and the coefficients c_α are unknown, and are in general fewer than the data points ($n \leq N$). This form of solution has the desirable property of being an universal approximator for continuous functions (Girosi and Poggio, 1989) and to be the only choice that guarantees that in the case of $n = N$ and $\{\mathbf{t}_\alpha\}_{\alpha=1}^n = \{\mathbf{x}_i\}_{i=1}^n$ the correct solution (of equation 1) is consistently recovered. We will see later in section (5) how to find the unknown parameters of this expansion.

3 Different types of Basis Functions.

This scheme can be further extended by considering in equation (5) the superposition of different types of functions G, such as Gaussians at different

scales.

The function f to be approximated is regarded as the sum of p components f^m, $m = 1, \ldots, p$, each component having a different prior probability. This assumption is clearly meaningful only if $p << N$. Therefore the functional $H[f]$ to minimize will contain p stabilizers P^m, p regularization parameters λ_m and will be written as

$$H[f] = \sum_{i=1}^{N}(y_i - \sum_{m=1}^{p} f^m(\mathbf{x}_i))^2 + \sum_{m=1}^{p} \lambda_m \|P^m f^m\|^2 \quad . \tag{6}$$

The Euler-Lagrange equations associated with equation (6) have the form:

$$\hat{P}^m P^m \ f^m(\mathbf{x}) = \frac{1}{\lambda_m} \sum_{i=1}^{N}(y_i - \sum_{k=1}^{p} f^k(\mathbf{x}_i))\delta(\mathbf{x} - \mathbf{x}_i) \quad m = 1, \ldots, p. \tag{7}$$

As in the case of standard regularization, the solution of equation (7) is a linear superposition of Green's functions:

$$f^m(\mathbf{x}) = \sum_{i=1}^{N} c_i^m G^m(\mathbf{x}; \mathbf{x}_i) \ . \tag{8}$$

The function $F(\mathbf{x})$ that minimizes the functional $H[f]$ is then a *linear superposition of linear superpositions* of the Green's functions G^m corresponding to the stabilizers P^m, that is

$$F(\mathbf{x}) = \sum_{m=1}^{p} \sum_{i=1}^{N} c_i^m G^m(\mathbf{x}; \mathbf{x}_i) \ + p(\mathbf{x}), \tag{9}$$

where $p(\mathbf{x})$ is a linear combination of functions that span the null spaces of the stabilizers. For instance, when $G^m(\mathbf{x})$ are Gaussian a polynomial is not needed, though it can always be added. For other Green's functions the theory requires an appropriate $p(\mathbf{x})$.

Substitution of equation (8) in equation (7) yields a linear system for the coefficients c_i^m. There is a simple relation between the coefficients associated to two different stabilizers, that is

$$c_i^m \lambda_m = c_i^n \lambda_n, \quad i = 1, \ldots, N; \quad n, m = 1, \ldots, p.$$

This means that if a component $f^m(\mathbf{x})$ of the solution is given, the other $p-1$ ones can be recovered by a simple scaling operation. This is expected, since the underlying variational problem is *underconstrained*: we are trying to obtain Np coefficients from a set of N data points. The form of the solution (9) is appealing: *if all the coefficients c_i^m were independent and free to vary*, the system could "choose" among different stabilizers, depending on the site. In order to retain the form (9) of the solution, while making the problem *overconstrained* instead of *underconstrained*, we choose a solution of the approximation problem of the following form (instead of equation 9):

$$\tilde{F}(\mathbf{x}) = \sum_{m=1}^{p} \tilde{f}^m(\mathbf{x}) + p(\mathbf{x}), \tag{10}$$

$$\tilde{f}^m(\mathbf{x}) = \sum_{\alpha=1}^{K_m} c_\alpha^m G^m(\mathbf{x}; \mathbf{t}_\alpha^m) \tag{11}$$

where $(1+d)\sum_{m=1}^{p} K_m < N$ and the coefficients c_α^m and the *centers* $\mathbf{t}_\alpha^m$ are unknowns. They can be found with a technique similar to the one described in section (5). Notice that equations (10) and (11) are of the same form as equation (5) and share its approximation properties.

3.1 Multiple Scales.

This method leads in particular to radial basis functions of multiple scales for the reconstruction of the function f. Suppose we know *a priori* that the function to be approximated has components on a number p of scales $\sigma_1, \ldots, \sigma_p$: we can use this information to choose a set of p stabilizers whose Green's functions are, for example, Gaussians of variance $\sigma_1, \ldots, \sigma_p$. We have (Poggio and Girosi, 1989, 1990a) :

$$\|P^m f^m\|^2 = \sum_{k=0}^{\infty} a_k^m \int_{R^n} d\mathbf{x} (D^k f^k(\mathbf{x}))^2$$

where $D^{2k} = \vec{\nabla}^{2k}$, $D^{2k+1} = \vec{\nabla}\vec{\nabla}^{2k}$ and $a_k^m = \frac{\sigma_m^{2k}}{k! 2^k}$, $\vec{\nabla}$ being the gradient operator. As a result, the solution will be a *superposition of superpositions* of Gaussians of different variances. Of course, the Gaussians with large σ should be preset, depending on the nature of the problem, to be fewer and therefore on a sparser grid, than the Gaussians with a small σ.

The HyperBF method also yields non-radial Green's functions – by using appropriate stabilizers – and also Green's functions with a lower dimensionality – by using the associated f^m and P^m in a suitable lower-dimensional subspace. Again this reflects *a priori* information that may be available about the nature of the mapping to be learned. In the latter case the information is that the mapping is of lower dimensionality or has lower dimensional components.

4 Weighted norm

The norm in equation (5) is usually intended as an Euclidean norm. If the components of $\mathbf{x}$ are of different types, it is natural to consider a *weighted norm* defined as

$$\|\mathbf{x}\|_{\mathbf{W}}^2 = \mathbf{x}^T \mathbf{W}^T \mathbf{W} \mathbf{x},$$

since the relative scale of the components is otherwise arbitrary. The case in which the matrix $\mathbf{W}$ is known (from prior information) does not present any difficulty. It is interesting, however, to see what it means in terms of the underlying regularization principle.

4.1 Weighted norm and regularization

The regularization principle consists in finding the f that minimizes the functional:

$$H_{\mathbf{W}}[f] = \sum_{i=1}^{N} (y_i - f(\mathbf{x}))^2 + \lambda \|Pf\|_{\mathbf{W}}^2 \tag{12}$$

where we assume that P is radially symmetric in the variable $\mathbf{y}$ and that $\mathbf{y} = \mathbf{W}\mathbf{x}$ (i.e. $\mathbf{y}$ is a known linear transformation of $\mathbf{x}$ that depends on the parameters $\mathbf{W}$). This means that the smoothness constraint is given in a space that is an affine transformation of the original $\mathbf{x}$ space. The Green's function associated with equation (12) is

$$G(\|\mathbf{y}\|^2) = G(\|\mathbf{x}\|_{\mathbf{W}}^2) \tag{13}$$

with $\|\mathbf{x}\|_{\mathbf{W}}^2 = \mathbf{x}^T \mathbf{W}^T \mathbf{W} \mathbf{x}$.

Suppose now that the parameters $\mathbf{W}$ are unknown. We can formulate the problem of finding f and $\mathbf{W}$ that minimize the functional $H_{\mathbf{W}}(f)$. Notice that the relevant quantity is $\mathbf{M} = \mathbf{W}^T\mathbf{W}$, since $\mathbf{W}$ only appears in this form. The matrix $\mathbf{M}$ is symmetric and positive definite; it has therefore a unique, symmetric "square root" $\mathbf{R}$, such that $\mathbf{M} = \mathbf{R}^T\mathbf{R} = \mathbf{R}^2$. One could chose to identify $\mathbf{W}$ with $\mathbf{R}$. $\mathbf{W}$ would be therefore symmetric, with $\frac{(d^2+d)}{2}$ independent parameters.

Thus finding the optimal $\mathbf{W}$ corresponds to finding the best stabilizer among those that are expressed in a coordinate system which is a linear transformation of the original one. The parameters $\mathbf{W}$ of the linear transformation become parameters of H with respect to which the functional is minimized.

The simplest case is the case of $\mathbf{W}$ diagonal and $G(x) = e^{-x^2}$. In this case

$$G(\|\mathbf{x}\|^2_{\mathbf{W}}) = e^{-x_1^2 w_1^2} e^{-x_2^2 w_2^2} \dots e^{-x_n^2 w_n^2}$$

and thus the components w_i of $\mathbf{W}$ are equivalent to the inverse of the variance σ of each component of the multidimensional Gaussian.

In the probabilistic interpretation of standard regularization (see Poggio and Girosi, 1989) the term $\lambda\|Pf\|^2$ in the regularization functional corresponds to the following prior probability in a Bayesian formulation in which the MAP (Maximum A Posteriori) estimate is sought:

$$Prob(f) = e^{-\lambda\|Pf\|^2}.$$

Our extension corresponds to choosing the stabilizer $P_{\mathbf{W}} = \|Pf(\mathbf{y})\|^2$, with $\mathbf{y} = \mathbf{W}\mathbf{x}$. The stabilizer $P_{\mathbf{W}}$ is parametrized by the matrix $\mathbf{W}$ and defines a prior $Prob_{\mathbf{W}}(f)$ which is also parametrized by $\mathbf{W}$.

The solution of the variational problem (12) has the form

$$f(\mathbf{x}) = \sum_{i=1}^{N} c_i G(\|\mathbf{x} - \mathbf{x}_i\|^2_W), \tag{14}$$

where the coefficients c_i and the elements of the matrix $\mathbf{W}$ must be estimated. Here again we are facing an underconstrained variational problem, since we trying to determine $N + \frac{(d^2+d)}{2}$ parameters from N data points. The same considerations of section (3) apply: in order to transform the problem into an overconstrained problem, we look for a solution of the form

$$f^*(\mathbf{x}) = \sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x} - \mathbf{t}_\alpha\|^2_{\mathbf{W}}) \tag{15}$$

5 How to learn centers' positions and norm weights

Suppose that we look for an approximated solution of the regularization problem of the form (15). We now have the problem of finding the n coefficients c_α, the $d \times n$ coordinates of the centers $\mathbf{t}_\alpha$ and the $\frac{(d^2+d)}{2}$ elements of the matrix $\mathbf{M}$ so that the expansion (12) is optimal. To avoid too many indeces, we will only consider here the case $p = 1$ in eq. 10. The extension is obvious. In this case we can use the natural definition of optimality given by the functional H. We then impose the condition that the set $\{c_\alpha, \mathbf{t}_\alpha | \alpha = 1, ..., n\}$ and the matrix $\mathbf{M}$ must be such that they minimizes $H[f^*]$, and the following equations must be satisfied:

$$\frac{\partial H[f^*]}{\partial c_\alpha} = 0 \ , \qquad \frac{\partial H[f^*]}{\partial \mathbf{t}_\alpha} = 0, \quad \frac{\partial H[f^*]}{\partial \mathbf{M}} = 0, \quad \alpha = 1, ..., n \ .$$

Gradient-descent is probably the simplest approach for attempting to find the solution to this problem, though, of course, it is not guaranteed to converge. Several other iterative methods, such as versions of conjugate gradient and simulated annealing (Kirkpatrick et al., 1983) may be more efficient than gradient descent and should be used in practice. Since the function $H[f^*]$ to minimize is in general non-convex, a stochastic term in the gradient descent equations may be advisable to avoid local minima. In the stochastic gradient descent method the values of c_α, $\mathbf{t}_\alpha$ and $\mathbf{M}$ that minimize $H[f^*]$ are regarded as the coordinates of the stable fixed point of the following stochastic dynamical system:

$$\dot{c}_\alpha = -\omega \frac{\partial H[f^*]}{\partial c_\alpha} + \eta_\alpha(t), \quad \alpha = 1, \ldots, n$$

$$\dot{\mathbf{t}}_\alpha = -\omega \frac{\partial H[f^*]}{\partial \mathbf{t}_\alpha} + \mu_\alpha(t), \quad \alpha = 1, \ldots, n$$

$$\dot{\mathbf{M}} = -\omega \frac{\partial H[f^*]}{\partial \mathbf{M}} + \Omega(t)$$

where $\eta_\alpha(t)$, $\mu_\alpha(t)$ and $\Omega(t)$ are white noise of zero mean and ω is a parameter determining the microscopic timescale of the problem and is related to the rate of convergence to the fixed point. Defining

$$\Delta_i \equiv y_i - f^*(\mathbf{x}) = y_i - \sum_{\alpha=1}^{n} c_\alpha G(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2_{\mathbf{W}})$$

and setting $\lambda = 0$ for simplicity (the more general case can be approached in a similar way) in equation (1) we obtain

$$H[f^*] = H_{\mathbf{c},\mathbf{t},\mathbf{M}} = \sum_{i=1}^{N} (\Delta_i)^2.$$

The important quantities - that can be used in more efficient schemes than gradient descent - are, with $\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2_{\mathbf{W}} = (\mathbf{x}_i - \mathbf{t}_\alpha)^T \mathbf{M} (\mathbf{x}_i - \mathbf{t}_\alpha)$ and $\mathbf{M} = \mathbf{W}^T\mathbf{W}$:

- for the c_α

$$\frac{\partial H[f^*]}{\partial c_\alpha} = -2 \sum_{i=1}^{N} \Delta_i G(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2_{\mathbf{W}}) \quad ; \tag{16}$$

- for the centers t_α

$$\frac{\partial H[f^*]}{\partial \mathbf{t}_\alpha} = 4 c_\alpha \sum_{i=1}^{N} \Delta_i G'(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2_{\mathbf{W}}) \mathbf{M} (\mathbf{x}_i - \mathbf{t}_\alpha) \tag{17}$$

- and for $\mathbf{M}$

$$\frac{\partial H[f^*]}{\partial \mathbf{M}} = -2 \sum_{\alpha=1}^{n} c_\alpha \sum_{i=1}^{N} \Delta_i G'(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2_{\mathbf{W}}) Q_{i,\alpha} \tag{18}$$

  where $Q_{i,\alpha} = (\mathbf{x}_i - \mathbf{t}_\alpha)(\mathbf{x}_i - \mathbf{t}_\alpha)^T$ is a dyadic product and G' is the first derivative of G.

Remarks

1. Instead of equation (18) for $\mathbf{M}$ the following equation can be used for $\mathbf{W}$:

$$\frac{\partial H[f^*]}{\partial \mathbf{W}} = -4\mathbf{W} \sum_{\alpha=1}^{n} c_\alpha \sum_{i=1}^{N} \Delta_i G'(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2_{\mathbf{W}}) Q_{i,\alpha} \qquad (19)$$

2. From equation (18) the matrix $\mathbf{M}$ is guaranteed to remain symmetric in a deterministic gradient descent scheme, since the right hand-side of the equation is symmetric (because the $Q_{i,\alpha}$ are correlation matrices and a linear combination of symmetric matrices is symmetric). Of course, the initial value must be a symmetric matrix and in the stochastic update scheme, the noise term must not break the symmetry. The matrix $\mathbf{M}$ must satisfy ditional constraint of remaining positive definite (since the scalar product $\mathbf{x}^T\mathbf{M}\mathbf{x}$ must be non-negative). We conjecture that equations (16), (17) and (18) conserve the positive definiteness of M *if* G is positive definite.

3. Equation (16) has a simple interpretation: the correction is equal to the sum over the examples of the products between the error on that example and the "activity" of the "unit" that represents with its center that example. Notice that $H[f^*]$ is quadratic in the coefficients c_α, and if the centers and the matrix $\mathbf{M}$ are kept fixed, it can be shown (Poggio and Girosi, 1989) that the optimal coefficients are given by

$$\mathbf{c} = (G^T\, G + \lambda g)^{-1} G^T \mathbf{y} \qquad (20)$$

where we have defined $(\mathbf{y})_i = y_i$, $(\mathbf{c})_\alpha = c_\alpha$, $(G)_{i\alpha} = G(\mathbf{x}_i; \mathbf{t}_\alpha)$ and $(g)_{\alpha\beta} = G(\mathbf{t}_\alpha; \mathbf{t}_\beta)$. If λ is let go to zero, the matrix on the right side of equation (20) converges to the pseudoinverse of G (Albert, 1972), and if the Green's function is radial the approximation method of Broomhead and Lowe (1988) is recovered.

4. Equation (17) is similar to task-dependent clustering (Poggio and Girosi, 1989). This can be best seen by assuming that Δ_i are constant: then the gradient descent updating rule makes the centers move as a function of the majority of the data, that is of the position of the clusters. In this case a technique similar to the k-means algorithm is recovered

(MacQueen, 1967; Moody and Darken, 1989). Fquating $\frac{\partial H[f^*]}{\partial t_\alpha}$ to zero we notice that, when the matrix $\mathbf{M}$ is set to the identity matrix, the optimal centers $\mathbf{t}_\alpha$ satisfy the following set of nonlinear equations:

$$\mathbf{t}_\alpha = \frac{\sum_i P_i^\alpha \mathbf{x}_i}{\sum_i P_i^\alpha} \quad \alpha = 1, \ldots, n$$

where $P_i^\alpha = \Delta_i G'(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2)$. The optimal centers are then a weighted sum of the data points. The weight P_i^α of the data point i for a given center $\mathbf{t}_\alpha$ is high if the interpolation error Δ_i is high there *and* the radial basis function centered on that knot changes quickly in a neighborhood of the data point. This observation suggests faster update schemes, in which a suboptimal position of the centers is first found and then the c_α are determined, similarly to the algorithm developed and tested successfully by Moody and Darken (1989).

5. Equation (19) (by assuming that $\sum_{\alpha=1}^n c_\alpha \Delta_i G'(\|\mathbf{x}_i - \mathbf{t}_\alpha\|^2_{\mathbf{W}})$ is asymptotically constant (!!)) contains the quantity $\sum_{i=1}^N Q_{i,\alpha}$ which is an estimate of the correlation matrix of all the examples relative to t_α (modulus a normalization factor). Let us define $C_{m,\alpha}$ as the $d \times m$ matrix whose columns are the vectors of the examples $\mathbf{x}_1 - \mathbf{t}_\alpha, ..., \mathbf{x}_m - \mathbf{t}_\alpha$. Then $\sum_{i=1}^N Q_{i,\alpha}$ can be written as $\sum_{i=1}^N Q_{i,\alpha} = C_{N,\alpha} C_{N,\alpha}^T$ and is the $d \times d$ correlation matrix (d being the number of components of $\mathbf{x}$). Interestingly, in this case, equation (19), when inserted in the gradient descent equation, has the form:

$$\dot{\mathbf{W}} = -\mathbf{W}Q$$

which has the solution

$$\mathbf{W}(t) = \mathbf{W}(0)e^{-Qt} = \mathbf{W}(0)\sum_{j=1}^{N} e^{-\lambda_j t}\mathbf{e}_j\mathbf{e}_j^T$$

where $\mathbf{e}_j$ are the eigenvectors of Q and λ_j are the associated eigenvalues. All eigenvectors will decay to 0, the ones with the largest eigenvalues fastest. Since in the full equation the other terms such as Δ_i will keep $\mathbf{W}$ from decaying to 0, we may expect that $\mathbf{W}$ will converge to a matrix

with rows that are similar to the eigenvectors of Q with the smallest eigenvalues. In other words, the equation should converge to rows of $\mathbf{W}$ that span the space orthogonal to the space spanned by the principal components of the input examples (i.e. the eigenvectors of Q with the largest eigenvalues). In this case, the matrix M is a projection operator that projects $\mathbf{x}$ into a space orthogonal to the space of the principal components. The principal components are the singular vectors of $\mathbf{X}$, with the property that they span a nested set of optimal subspaces. This interpretation of the gradient descent equation is just a rough indication of what may happen, because of the very strong underlying assumptions. It turns out that in the object recognition case (Poggio and Edelmann, 1990), the interpretation is perfectly consistent with what one expects, given the (linear) computational theory underlying the problem (Basri and Ullmann, 1990; see also the appendix in Edelmann and Poggio, 1990). Under orthographic projection, the vectors representing views of the same object span a linear subspace with a low dimension. Let us assume, according to the above discussion, that $\mathbf{W}$ projects a new input vector into a space orthogonal to the one spanned by the principal components extracted from many views of the object (the "examples"). Then, if the new input is another view of the same object, the result will be close to zero for all units. In the case of the Gaussian, for instance, this means that each unit will be maximally activated and by suitable choice of $\mathbf{c}$ any desired output may be synthesized. On the other hand, if the new input is the view of a different object, the result of operating on it with $\mathbf{W}$ will be different from zero and possibly large enough to give a very small activity of the unit making it impossible to synthesize a desired output by an appropriate choice of the $\mathbf{c}$ (the output will be zero or close to it). In this case, the appropriate $\mathbf{W}$ will solve the problem with just one center (since the problem is linear). Notice that if $\mathbf{W}$ is symmetric (i.e. if $\mathbf{W}$ is the square root of $\mathbf{M}$), it has the same eigenvectors of $\mathbf{M}$, and $\mathbf{M}$ and $\mathbf{W}$ have the same null space.

6. One may think intuitively that it is desirable that $\mathbf{W}$ is space dependent, that is $\mathbf{W} = \mathbf{W}(\mathbf{x})$. This assumption, however, seems rather meaningless from the point of view of regularization theory. As a consequence, we believe that it is wrong to assume $\mathbf{W} = \mathbf{W}(\mathbf{x})$ in a scheme

such as HyperBF. On the other hand, it makes theoretically sense to use different HyperBF networks for different subsets of the domain of the given multivariate function, each one possibly with a different $\mathbf{W}$. We do not have any theory, however, of how to partition appropriately the domain of the function. An alternative approach, that also makes sense, is local linear approximation. In this case one finds a set of local charts, somewhat similarly to computing $\mathbf{W}(\mathbf{x})$.

5.1 A practical algorithm

It seems natural to try to find a reasonable initial value for the parameters $\mathbf{c}, \mathbf{t}_\alpha, \mathbf{M}$, to start the minimization process. In the absence of more specific prior information the following heuristics seems reasonable.

- Set the number of centers and set the centers' positions to positions suggested by cluster analysis of the data (or more simply to a subset of the examples' positions).
- Set the rows of $\mathbf{W}$ to be vectors orthogonal to the eigenvectors with largest eigenvalues of $\sum_\alpha \sum_i Q_{i,\alpha}$.
- Use matrix pseudo-inversion to find the c_α.
- Use the $\mathbf{t}_\alpha$, $\mathbf{M} = \mathbf{W}^T\mathbf{W}$ and c_α found so far as initial values for gradient descent equations.

It should be noticed than an even more general strategy makes sense in some cases. Suppose that the system can be made to operate satisfactorily with the steps above or perhaps just with the first step. Suppose also that the system can continue to accumulate examples while operating. An example could be an autonomous vehicle that can improve, say, the model of its dynamics by collecting appropriate example pairs while operating. Then it makes sense to perform dimensionality reduction and to move the centers as outlined above. As an additional step one may try to eliminate features that receive little weight, if possible, and then to add other features while keeping the previously found centers. This is equivalent to adding centers of higher dimensionality. Another iteration of moving centers, finding norm weights, eliminating features and centers then takes place.

Experiments with movable centres and movable weights have been performed in the context of object recognition (Poggio and Edelman, 1990; Edelman and Poggio, 1990) and approximation of multivariate functions (Caprile, Girosi and Poggio, 1990) and in both cases the results are promising.

6 Remarks

1. Equation (19) is similar to an operation of (task-dependent) dimensionality reduction (Duda and Hart, 1973) whereas equation (17) is similar to a clustering process.

2. It is conceivable that learning the weights of the norm is even more important than learning the centers and that in many cases it may be preferable to set the centers to a representative subset of the data and to keep them fixed thereafter.

3. A specific matrix $\mathbf{W}$ corresponds to a specific metric in the multidimensional input space: $\mathbf{W}$ projects the input vector into the subspace spanned by its rows. In the case of the rows of $\mathbf{W}$ spanning the space orthogonal to the principal components of the inputs, $\mathbf{W}$ assigns a metric ellipsoid with the largest axes (corresponding to a large σ in the Gaussian) along the principal components and the small axis (corresponding to a small σ in the Gaussian) orthogonal to it: thus even vectors that are far away (in the ordinary euclidean metric) are close in this metric if they lie in the hyperplane of the principal components and even close vectors (in the ordinary metric) are far away in the metric induced by $\mathbf{W}$ if they are orthogonal to the principal components.

4. In the case of N examples, $n = N$ fixed centers and $\mathbf{M} = I$, there are enough data to constrain the N coefficients c_α to be found. Moving centers add another nd parameters (d is the number of input components) and the matrix $\mathbf{M}$ another $\frac{d^2+d}{2}$ independent parameters. Thus the number of examples N must be sufficiently large to constrain adequately the free parameters – n d-dimensional centers, n coefficients c_α and $\frac{d^2+d}{2}$ independent entries of the matrix $\mathbf{M}$. Thus

$$N >> n + nd + \frac{d^2 + d}{2}.$$

5. In the case of Gaussian basis functions, learning the entries of a diagonal $\mathbf{W}$ is equivalent to learning the variances of each two-dimensional (or one-dimensional) Gaussian receptive field for each center. It is clear that sets of units with different scales (see section 3.1) correspond to sets of units with different $\mathbf{W}$.

Acknowledgements We thank Shimon Ullman for useful discussions with one of us (T.P.) that led to an understanding of the role of $\sum_{i=1}^{N} Q_{i,\alpha}$ in the object recognition example. Lew Tucker and Barbara Moore had suggested weights similar to $\mathbf{W}$ long before T.P. finally understood what they meant. We also thanks B. Caprile, C. Furlanello and E. Grimson for useful suggestions and discussions.

References

[1] A. Albert. *Regression and the Moore-Penrose Pseudoinverse.* Academic Press, New York, 1972.

[2] R. Basri and S. Ullman. Recognition by linear combinations of models. A.I. Memo No. 1152, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[3] M. Bertero. Regularization methods for linear inverse problems. In C. G. Talenti, editor, *Inverse Problems.* Springer-Verlag, Berlin, 1986.

[4] D.S. Broomhead and D. Lowe. Multivariable functional interpolation and adaptive networks. *Complex Systems*, 2:321–355, 1988.

[5] B. Caprile, F. Girosi, and T. Poggio. Hyperbf networks: techniques and experiments. A.I. Memo (in preparation), Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[6] J. Demmel. The geometry of ill-conditioning. *J. Complexity*, 3:201–229, 1987.

[7] R. O. Duda and P. E. Hart. *Pattern Classification and Scene Analysis.* Wiley, New York, 1973.

[8] S. Edelman and T. Poggio. Bringing the grandmother back into the picture: a memory-based view of object recognition. A.I. Memo (to appear), Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1990.

[9] F. Girosi and T. Poggio. Networks and the best approximation property. A.I. Memo 1164, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[10] S. Kirkpatrick, C.D. Gelatt, and M.P. Vecchi. Optimization by simulated annealing. *Science*, 220:219–227, 1983.

[11] J. MacQueen. Some methods of classification and analysis of multivariate observations. In L.M. LeCam and J. Neyman, editors, *Proc. 5th Berkeley Symposium on Math., Stat., and Prob.*, page 281. U. California Press, Berkeley, CA, 1967.

[12] S.G. Mikhlin. *The problem of the minimum of a quadratic functional.* Holden-Day, San Francisco, CA, 1965.

[13] J. Moody and C. Darken. Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1(2):281–294, 1989.

[14] V.A. Morozov. *Methods for solving incorrectly posed problems.* Springer-Verlag, Berlin, 1984.

[15] T. Poggio and S. Edelman. A network that learns to recognize 3D objects. *Nature*, 343:263–266, 1990.

[16] T. Poggio and F. Girosi. A theory of networks for approximation and learning. A.I. Memo No. 1140, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, 1989.

[17] T. Poggio and F. Girosi. A theory of networks for learning. *Science*, 247:978–982, 1990.

[18] T. Poggio and F. Girosi. Hyperbf: A powerful approximation technique for learning. In Patrick H. Winston and Sarah A. Shellard, editors, *Artificial Intelligence at MIT: Expanding Frontiers, Vol. 1.* M.I.T. Press, Cambridge, MA, 1990a.

[19] I. Stakgold. *Green's functions and boundary problems.* John Wiley and Sons, New York, 1979.

[20] A. N. Tikhonov. Solution of incorrectly formulated problems and the regularization method. *Soviet Math. Dokl.*, 4:1035–1038, 1963.

[21] A. N. Tikhonov and V. Y. Arsenin. *Solutions of Ill-posed Problems.* W. H. Winston, Washington, D.C., 1977.