1. In a PLN with K clusters, $y_p = A_k \cdot x_p$, where the kth cluster is closest to $x_p$. We want to calculate derivatives of this mapping.

(a) Let $a_k(m,n)$ denote an element of $A_k$ and let $g(i,n)$ denote the partial derivative of $y_{pi}$ with respect to $x_{pn}$. Find $g(i,n)$ if $x_p$ is closest to the kth cluster.

(b) How many $G$ matrices are there for this network?

(c) If $x_n$ (for all values of $p$) has no useful information for calculating $y_i$ (for all $p$), what value should we expect $g(i,n)$ to have?

(d) We want to calculate weights $w(n)$ (for $n$ between 1 and $N$) for a weighted distance measure, using coefficients from all of the $A_k$ matrices. Using ideas from part (c), give a reasonable expression for $w(n)$.

2. Given training patterns $(x_p, t_p)$, the piecewise linear network (PLN) is designed by (1) clustering the vectors $x_p$ to produce K mean vectors $m_k$ and (2) designing a different linear mapping for each cluster. The PLN’s output vector is $y_p' = A_k \cdot x_p$, where the distance $d(x_p, m_k)$ is the smallest. Let $d(x_p, m_k)$ denote the weighted distance between $x_p$ and $m_k$, expressed as

$$d(x_p, m_k) = \sum_{n=1}^{N} w(n)[x_p(n) - m_k(n)]^2$$

(a) What is the storage capacity of the PLN if we assume that each cluster has exactly the maximum number of patterns that can be memorized?

(b) Find $P_{ab}/M$ in terms of $N$, $K$, and $M$.

(c) Suppose that we want our training algorithm to maximize the actual number of memorized patterns. If $N_v >> C_{PLN}$, what is the actual number of memorized patterns if the patterns are evenly distributed among the K clusters?

(d) Continuing part (c), how should the clustering algorithm distribute the patterns among the clusters to maximize the actual pattern memorization?
3. Consider a PLN that uses the normal, squared Euclidean distance measure. In order to make the mapping continuous, we can construct a matrix $B$ as

$$B = \sum_{k=1}^{K} w_k A_k$$

using the distances $d_k = d(x, m_k)$ and $d_{\text{max}} = \max \{d_k\}$. The output vector is then $y = B \cdot x$, where $B$ is a function of $x$.

(a) What value should $w_k$ approach as $x$ approaches $m_k$?
(b) Using $d_k$, give an expression for $w_k$ that satisfies the requirement of part (a).
(c) How many multiplies are required to calculate all of the $d_k$ and $w_k$, given $x$?
(d) Give an expression for $M_{\text{PLN}}$.

4. The radial basis function (RBF) network can be analyzed using polynomial basis functions (PBFs). Let $\text{net}_k$ denote the usual RBF net function for the $k$th center. Assume that $N_v$ is much larger than $K$, the number of cluster centers or hidden units.

(a) If the hidden unit activations are of the form

$$\phi(\text{net}) = \text{net}^2$$

what can we say about the $x_n^2$ terms at the output nodes, for different values of $k$?
(b) Give an expression for $P_{ab}$, assuming that the network has no direct connections from input to output, and no output layer thresholds.
(c) For the activations of part (a), find an upper limit on the number of effective free parameters in the network.
(d) For the activations of part (a), how many patterns can we guarantee that the RBF network can memorize? What is an upper bound on $C_{\text{rbf}}$?

5. An RBF network with cascade connectivity and no thresholds is to be designed, using $N_v$ training patterns. Let each of the $N_v$ input vectors be the center vector of a hidden unit. The distances calculated as net values for the $k$th hidden unit are $d(k,1), d(k,2), \ldots, d(k,k), d(k,k+1), \ldots, d(k,N_v)$ where $d(k,m)$ denotes the squared Euclidean distance from the $k$th input vector to the $m$th input vector. The Schmidt Procedure is to be used, with ordering, so that the most effective center vectors can be found.

(a) How many unique distances $d(k,m)$ exist? Don’t count zero-valued distances.
(b) Given the $d(k,m)$ and the corresponding hidden unit activations $O(k,m)$, give an expression for $r(m,n)$, for hidden unit activations. How many multiplies are required to calculate one $r(m,n)$ if $m$ is not equal to $n$?
(c) Given the $d(k,m)$ and $O(k,m)$, how many multiplies are required to calculate the cross-correlation matrix $C$?
6. We want to use BP to train a radial basis function (RBF) network with standard cascade connectivity and no thresholds. Its **weighted** distance measure, activation output, and error function are

\[ d(x_p,m_k) = \sum_{n=1}^{N} w(n)[x_p(n) - m_k(n)]^2, \]

\[ O_p(k) = e^{-\beta d(x_p,m_k)}, \]

\[ y_p(i) = \sum_{k=1}^{N_h} w_o(i,k)O_p(k) \]

\[ E = \frac{1}{N_v} \sum_{p=1}^{N_v} \sum_{i=1}^{M} [t_p(i) - y_p(i)]^2 \]

(a) Give a two-line expression for \( \partial E/\partial \beta_m \) in the form

\[ \frac{\partial E}{\partial \beta_m} = \frac{?}{N_v} \sum_{p=1}^{N_v} \sum_{i=1}^{M} [t_p(i) - y_p(i)] \frac{\partial ?}{\partial ?}, \]

\[ \frac{\partial ?}{\partial ?} = ? \]

(b) Suppose that Newton’s algorithm is to be used to train the \( N_h \) by 1 vector \( \beta \). Give the Hessian matrix element \( h(m,k) \) in Gauss-Newton form where

\[ h(m,k) = \frac{\partial^2 E}{\partial \beta_m \partial \beta_k} \]

Remember, the Gauss-Newton approximation is

\[ \frac{\partial^2 f^2(x,y)}{\partial x \partial y} = 2 \frac{\partial f(x,y)}{\partial x} \cdot \frac{\partial f(x,y)}{\partial y} \]