1. A sigmoidal MLP has 8 inputs, 12 units in the first hidden layer, 6 units in the second hidden layer, and 3 outputs. It is fully connected. As usual, thresholds in the hidden and output layers are handled by adding a 9th input that equals 1.
   (a) Give the lower and upper bounds on pattern storage.
   (b) How many absolute free parameters are there in the network?
   (c) What degree functional link net would have at least as great a pattern storage as the MLP’s lower bound?
   (d) What degree functional link net would have at least as great a pattern storage as the MLP’s upper bound?

2. In a group of $N_{v1}$ training patterns, each of the $N_u$ unique input vectors occurs $K$ times, each time with a different desired output vector. Therefore, $N_{v1} = K \cdot N_u$.
   For the $k$th unique input vector $x_k$, the corresponding desired output vectors are $t_{km}$ for $m$ between 1 and $K$. In a second group of $N_{v2}$ training patterns, input vectors and desired output vectors are unique.
   (a) When a very powerful training algorithm trains an MLP with $P_{ab}$ absolute free parameters for patterns in the first group, let the output training error be denoted as $E_1$. For what value of $P_{ab}$ will should $E_1$ reach its minimum?
   (b) In terms of $K$ and $t_{km}$, what output $y_k$ will the network of part (a) produce every time $x_k$ is the input vector?
   (c) If the same training algorithm trains an MLP with $P_{ab}$ absolute free parameters, for patterns in the second group, what training error $E_2$ can we expect when $P_{ab}/M = N_{v2}$?
   (d) Putting both groups together, we have $N_v$ training patterns, where $N_v = N_{v1} + N_{v2}$. If we train a third network on all $N_v$ patterns, for what value of $P_{ab}$ can $E$ reach its minimum value? Give the resulting value of $E$ in terms of symbols $E_1$, $E_2$, $N_{v1}$, $N_{v2}$, and $N_v$.

3. In the proof of theorem 4.2, we need to show that $E[(t(i) - y_{opt}(i)) \cdot (y_{opt}(i) - y(i))] = 0$. Using $y(i)$’s Schmidt procedure basis functions for $y_{opt}(i)$ and $t(i)$ as well, we have
   \[ y(i) = \sum_{k=1}^{L} w_o'(k)X'_k, \quad y_{opt}(i) = \sum_{k=1}^{L+K} w_o'(k)X'_k, \quad t(i) = e + \sum_{k=1}^{m} w_o'(k)X'_k \]
   where, $e$ is zero-mean noise statistically independent of the basis functions.
   (a) Plug the expressions above into $E[(t(i) - y_{opt}(i)) \cdot (y_{opt}(i) - y(i))]$ and simplify, without evaluating the expected value operator.
   (b) Applying the expected value operator, why is the final answer zero?
4. In a classifier with $L$ basis functions and $M$ classes, suppose the $M \times L$ “optimal” output weight matrix $W^{\text{opt}}$ is known, as well as the $L \times L$ basis function autocorrelation matrix $R$.

(a) Find $C^{\text{opt}}$ in terms of $R$ and $W^{\text{opt}}$.

(b) We know that $C^{\text{opt}}$ can be expressed as

$$C^{\text{opt}} = \frac{1}{N_c} \sum_{p=1}^{N_c} X_p (t_p)^T$$

If $c(i)$ is the $i$th column of $C^{\text{opt}}$, and the $p$th element of the column vector $t(i)$ is $t_p(i)$, find the $n$th row $p$th column element of the $L \times N_c$ matrix $A$ so that

$$c(i) = A \cdot t(i)$$

(c) If we solve the linear equations above for the desired outputs $t_p(i)$, under what circumstances is the vector $t(i)$ unique?

5. When neural classifier desired outputs are coded, performance is usually bad. However, it may be possible for them to have $(N_c - 1)$ desired outputs, and still have good performance. For the $N_c = 3$ case, let the desired uncoded MLP classifier outputs be $t(i) = \delta(i - i_c)$ where $i_c$ is the correct class. Let $y(1)$, $y(2)$, and $y(3)$ denote the uncoded MLP classifier outputs. Suppose that two coded desired outputs are used, which are $t'(1) = t(1) - t(2)$ and $t'(2) = t(2) - t(3)$.

(a) Given the coded outputs $y'(1)$ and $y'(2)$, how can we construct $y'(3)$? What are the three possible values of $t'(i)$?

(b) Given $y'(1)$, $y'(2)$, and $y'(3)$, when do we decide class 1? class 2? class 3?

(c) Realistically, we can model $y'(1)$ as $t'(1) + e(1)$ and $y'(2)$ as $t'(2) + e(2)$, where $e(1)$ and $e(2)$ are noise. We can also model $y'(3)$ as $t'(3) + e(3)$. If $e(1)$ and $e(2)$ are statistically independent noise and vary from -.8 to +.8, how will $e(3)$ vary?

(d) In part (c), can $y'(1)$ or $y'(2)$ ever have the incorrect sign? Can $y'(3)$ ever have the incorrect sign?

(e) Considering your answer in part (d), is this new set of coded outputs as good as using uncoded outputs?
6. A support vector machine (SVM) classifier has $N$ inputs, $N_{sv}$ support vectors (SVs or hidden units), one output, and one output threshold $b$. The hidden units are already initialized and are not trained.

(a) How many training patterns can be memorized by this network structure, if the usual MSE $E$ is minimized using OWO? Hint: Use $C_L$.

(b) If the output weights are $w(k)$ for $1 \leq k \leq N_{sv}$, what term $E_1$ can added to $E$ before training, so that only $N_{sv}$ patterns are memorized?

(c) Given $N_v$ training patterns and a known value of $N_{sv}$ that is less than $N_v$, we want to use LM to minimize the error function $E_t = E + E_1$ to train an SVM. However, we don’t know which group of $N_{sv}$ patterns are the support vectors. In terms of $N_v$ and $N_{sv}$, how many networks, at most, do we train before finding the SVM?

(d) Continuing part (c), how do we know when we’ve found the SVM?
Reference Material

**Theorem 4.2:** When a network’s output $y$ minimizes the training MSE with respect to its weights, this error can be decomposed as:

$$E[||t - y||^2] = E[||t - y_{opt}||^2] + E[||y_{opt} - y||^2].$$

Here $y_{opt}$, which is the output of a network similar to that of $y$ but more complex, minimizes $E[||t - y_{opt}||^2]$.

Given a set of $N$ elements, the number of subsets of size $M$ is

$$\binom{N}{M} = \frac{N!}{(N-M)!M!}$$