

# AM41AN Coursework 1

Due by Friday the 13 May 2022 at 12:00 (electronic submission only).

## Instructions

In the present coursework you are required to perform a number of tasks and answer a number of questions that are presented in the following. All your findings (tasks results and answers to the questions) must be presented in the form of a scientific report. Such a report must be written in L<sup>A</sup>T<sub>E</sub>X, and it must contain an abstract, introduction, methods, results and conclusions.

The description of your tasks is as follows:

Download the file `data.dat` from the AM41AN web page in BlackBoard. This file contains one thousand registers, each form by a pair of coordinates  $(x_n, t_n)$ . The first coordinate  $x_n$  is a feature, or independent variable, the second  $t_n$  is the target, or dependent variable. Some level of noise has been added to the targets. A plot of  $t$  against  $x$  is presented in figure 1.

1. Your first task is to estimate the conditional expectation:

$$\langle t|x \rangle = \int dt \mathcal{P}(t|x) t,$$

where  $\mathcal{P}(t|x)$  is the conditional probability of  $t$  conditioned to  $x$ , by training the network presented in figure 2 through error back-propagation. This proposed network consists of three layers (input, hidden and output). The units in the input layer correspond to  $z_0^{(0)} = 1$  and  $z_1^{(0)} = x$ . The  $M + 1$  units in the hidden layer are  $z_0^{(1)} = 1$ , and  $z_k^{(1)} = \tanh \left( w_{k,0}^{(1)} z_0^{(0)} + w_{k,1}^{(1)} z_1^{(0)} \right)$  for all  $0 < k \leq M$ . The output unit is linear, i.e.  $z^{(2)} = w_0^{(2)} z_0^{(1)} + \sum_{k=1}^M w_k^{(2)} z_k^{(1)}$ . You are required to consider the cases with  $M = 7, 10$  and  $20$ . You may use the following error function:

$$E_0(\mathbf{w}; \{(x_n, t_n)\}) = \sum_{n=1}^{1000} \frac{1}{2} \left( t_n - z^{(2)}(x_n) \right)^2. \quad (1)$$

2. Make a comment in your report on how this problem could have been solved using radial basis functions.

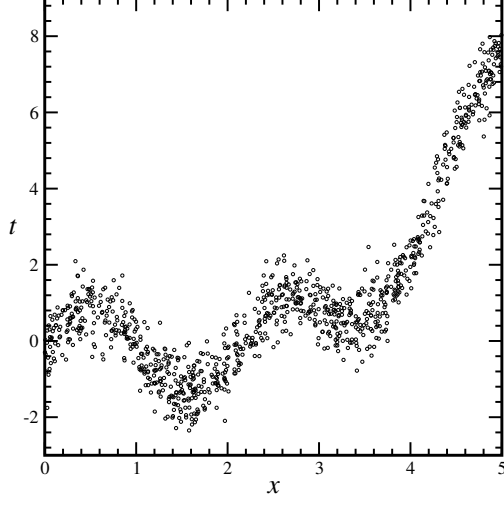


Figure 1: Plot of the points in the data set.

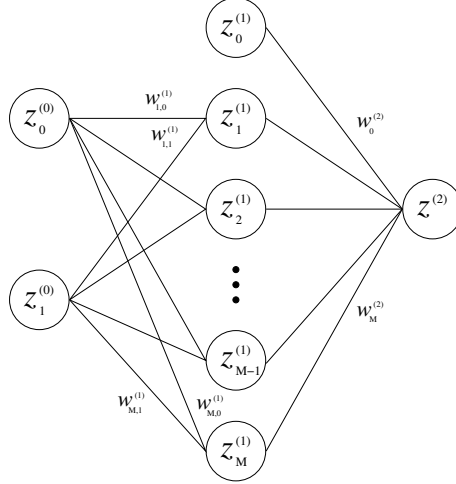


Figure 2: Architecture of the proposed network. The units in the input layer correspond to  $z_0^{(0)} = 1$  and  $z_1^{(0)} = x$ . The  $M + 1$  units in the hidden layer are  $z_0^{(1)} = 1$ , and  $z_k^{(1)} = \tanh \left( w_{k,0}^{(1)} z_0^{(0)} + w_{k,1}^{(1)} z_1^{(0)} \right)$  for all  $0 < k \leq M$ . The output unit is linear, i.e.  $z^{(2)} = w_0^{(2)} z_0^{(1)} + \sum_{k=1}^M w_k^{(2)} z_k^{(1)}$ .

3. Once you have solved the regression task set above, consider the following regularized error function:

$$\tilde{E}(\mathbf{w}) = E_0(\mathbf{w}) + \frac{\nu}{2} \mathbf{w}^T \mathbf{w}. \quad (2)$$

Explain what are the expected effects of the new added term  $\frac{\nu}{2} \mathbf{w}^T \mathbf{w}$ , and how these effects depend on the value of  $\nu$ . Use the spectrum of the Hessian matrix (in the following notation many subscripts and superscripts have being dropped, see Eqs. (8) to (10) below)

$$[\mathbf{H}]_{j,k} = \frac{\partial^2 E_0}{\partial w_j \partial w_k}$$

to relate the weights minimising the cost with regularization (2),  $\tilde{\mathbf{w}}_\nu$ , to the cost without regularization (1),  $\mathbf{w}^*$ , i.e.

$$\tilde{\mathbf{w}}_\nu = \left( \sum_k \frac{\lambda_k}{\lambda_k + \nu} \mathbf{u}_k \mathbf{u}_k^T \right) \mathbf{w}^*, \quad (3)$$

where  $\mathbf{H} \mathbf{u}_k = \lambda_k \mathbf{u}_k$ . Make a plot of the error function  $E_0(\tilde{\mathbf{w}}_\nu)$  against  $\nu \in (0, 1)$  for the networks with  $M = 7, 10$  and  $20$ .

## Hints

As mentioned in the Instructions, you must present your work in the form of a report. A template for the report is provided in the file `cw_report_template.pdf`

In the Method section you must demonstrate that the error back-propagation equations for this network are:

$$\frac{\partial E_n}{\partial w_m^{(2)}} = \delta^{(2)} z_m^{(1)} \quad (4)$$

$$\frac{\partial E_n}{\partial w_{m,j}^{(1)}} = \delta_m^{(1)} z_j^{(0)} \quad (5)$$

with

$$\delta^{(2)} = z^{(2)} - t \quad (6)$$

$$\delta_m^{(1)} = \delta^{(2)} w_m^{(2)} \left[ 1 - \left( z_m^{(1)} \right)^2 \right]. \quad (7)$$

You must also demonstrate that the entries of the Hessian satisfy the fol-

lowing relationships:

$$\frac{\partial^2 E_n}{\partial w_{n,k}^{(1)} \partial w_{m,j}^{(1)}} = \left\{ \kappa_n^{(1)} \kappa_m^{(1)} - 2\delta_{n,m} z_m^{(1)} \delta_m^{(1)} \right\} z_k^{(0)} z_j^{(0)} \quad (8)$$

$$\frac{\partial^2 E_n}{\partial w_n^{(2)} \partial w_{m,j}^{(1)}} = \left\{ z_n^{(1)} \kappa_m^{(1)} + \delta_{n,m} q_m^{(1)} \right\} z_j^{(0)} \quad (9)$$

$$\frac{\partial^2 E_n}{\partial w_n^{(2)} \partial w_m^{(2)}} = z_n^{(1)} z_m^{(1)}, \quad (10)$$

with

$$\kappa_m^{(1)} = w_m^{(2)} \left[ 1 - \left( z_m^{(1)} \right)^2 \right] \quad (11)$$

$$q_m^{(1)} = \delta^{(2)} \left[ 1 - \left( z_m^{(1)} \right)^2 \right]. \quad (12)$$

The error function (1) derives from a particular model of noise. You must explain (in the Methods section) what model is this, and how such a model is associated with a sum-of-squares type of error function.

Your error back-propagation program must implement the following learning algorithm

$$\left[ w_{n,k}^{(1)} \right]_{\ell+1} = \left[ w_{n,k}^{(1)} \right]_{\ell} - \eta_{\ell} \frac{\partial E_{\ell}}{\partial w_{n,k}^{(1)}} \quad (13)$$

$$\left[ w_k^{(2)} \right]_{\ell+1} = \left[ w_k^{(2)} \right]_{\ell} - \eta_{\ell} \frac{\partial E_{\ell}}{\partial w_k^{(2)}}, \quad (14)$$

where the sub-index  $\ell$  indicates the iteration and  $\eta_{\ell}$  is the learning rate. I have found that a learning rate satisfying:

$$\eta_{\ell} = \frac{0.1}{\sqrt{\ell + 1}}, \quad (15)$$

works fine. Feel free to experiment with other learning rates (you are not required to report your findings on this particular exploration).

Iterating the equations (13) and (14) only once will not produce a suitable set of parameters. I would suggest you start your simulation ( $\ell = 0$ ) with sufficiently small values of the parameters  $\left[ w_{n,k}^{(1)} \right]_0$  and  $\left[ w_k^{(2)} \right]_0$  ( $O(10^{-2})$ ), after iterating equations (13) and (14) one thousand times (the size of the data set) you should reset your learning rate to its initial value ( $\eta_0 = 0.1$ ), scramble your data set (to avoid giving excessive importance to the first elements of the original data set) and start a second run of 1000 iterations (epoch) taking as initial parameter values the last values of the previous run. You may need up to 500 epochs to obtain a suitable result. The description of the implementation of the error back-propagation method (parameters, initial conditions, number of epochs used until convergence) should be described in the Results section.

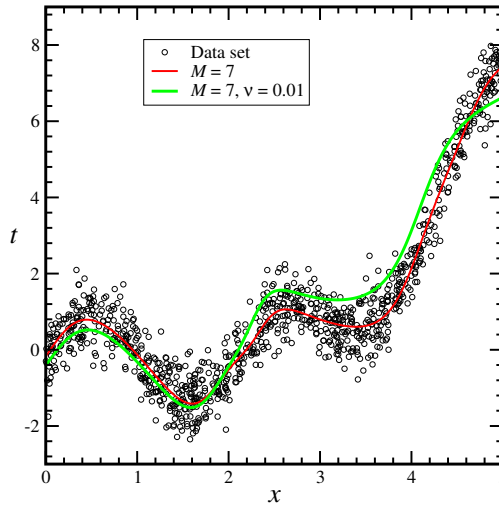


Figure 3: Comparison of the data points (black circles) with the output of a network with 7 hidden nodes without regularization (red curve), and with regularization (with  $\nu = 0.01$ ).

You must present your results in the form of graphs, in which you must compare the data set against the implementation of your trained and regularized networks.

You must also present the numerical values of the parameters  $w_{n,k}^{(1)}$  and  $w_k^{(2)}$ . Not two independent implementations of these methods should produce the same set of parameters, therefore these numbers will be used to check possible collusion.

Your results must be presented also in the form of graphs. For instance, to show the match between your results and the data set you may present a graph like the one in figure 3.

To show the performance of the regularization process you may present a graph similar to the one in figure 4. Your Conclusions must refer to the results you have obtained. Once you have finished with all your sections you may write the Abstract in which you make a brief summary of all the other sections.

The marking scheme is as follows:

- 10 marks for presentation.
- 5 marks for the Abstract.
- 25 marks for Introduction (well described and defined problem, in mathematical terms).
- 25 marks for Methods (all the developments of the quantities you need and the description of the network's architecture must be presented here).

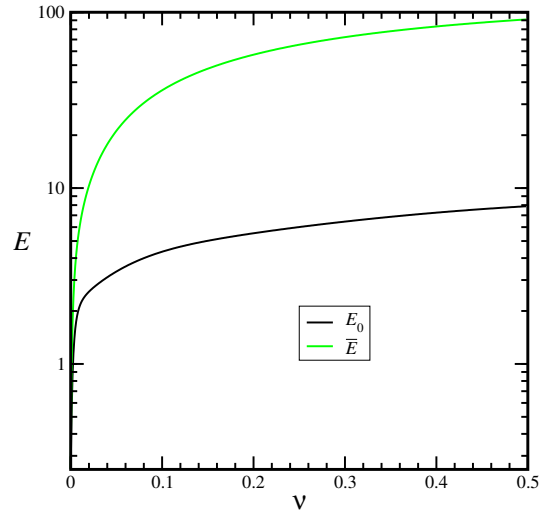


Figure 4: Comparison of quadratic error with (green curve) and without regularization term ( $\nu\Omega(\mathbf{w}) = \nu\mathbf{w} \cdot \mathbf{w}$ ).

- 20 marks for Results. Graphs with your results and comments on the learning rate and initial values used in your program must be mentioned here. You must also present in a table format the numerical values of the parameters you have found. No marks will be granted if you fail to present your parameters.
- 15 marks for Conclusions.