آرین شکلر - ۴۰۰۲۳۲۹۴   تکلیف شماره ۴ درس شبکه‌های عصبی

E12.2. First we will compute the direction of the initial step without batching:

$$a_1 = \text{logsig}(wp_1+b) = \frac{1}{1+\exp\{-(0\times(-2)+0.5)\}} = 0.6225$$

$$e_1 = t_1 - a_1 = 0.8 - 0.6225 = 0.1775$$

Assuming MSE performance index:

$$s_1^1 = -2\dot{f}(n)\,e_1 = -2a_1(1-a_1)e_1 = -2(0.6225)(1-0.6225)(0.1775) = -0.0834$$

We know that in steepest descent, the direction of the step is set to the negative of the gradient at that step. We know that:

$$\nabla w^m = s^m (a^{m-1})^T$$

$$\nabla b^m = s^m$$

because the presented network has only one layer, the direction w.r.t. weights is:

$$-\nabla w_1^1 = -s_1 p_1 = -(-0.0834)(-2) = -0.1668$$

and w.r.t. biases is:

$$-\nabla b_1^1 = -s_1 = -(-0.0834) = 0.0834$$

therefore the direction of the initial step in the (w,b) plane is: $\begin{bmatrix} -0.1668 \\ 0.0834 \end{bmatrix}$

Now we will consider batching. We will apply the second input to the network and calculate the directions w.r.t. the new input. Then we average the direction w.r.t. this input and the one we calculated previously for the final direction in batch mode.

$$a_2 = \text{logsig}(wp_2+b) = \frac{1}{1+\exp\{-(0\times 2+0.5)\}} = 0.6225$$

$$e_2 = t_2 - a_2 = 1 - 0.6225 = 0.3775$$

$$s_2^1 = -2\dot{f}(n_2)e_2 = -2a_2(1-a_2)(e_2) = -2(0.6225)(0.3775)(0.3775) = -0.1774$$

$\rightarrow$ direction w.r.t. weights: $-\nabla w_2^1 = -s_2 p_2 = -(-0.1774)(2) = 0.3548$

$\rightarrow$ direction w.r.t. biases: $-\nabla b_2^1 = -s_2 = -(-0.1774) = 0.1774$

$\rightarrow$ dir: $\begin{bmatrix} 0.3548 \\ 0.1774 \end{bmatrix}$

Now we can average the two directions for the final direction:

$$\begin{bmatrix} \text{dir } W \\ \text{dir } b \end{bmatrix} : \frac{1}{2}\left(\begin{bmatrix} -0.1668 \\ 0.0834 \end{bmatrix} + \begin{bmatrix} 0.3548 \\ 0.1774 \end{bmatrix}\right) = \begin{bmatrix} 0.0940 \\ 0.1304 \end{bmatrix}$$

If we compare this with the direction from the non-batch mode, we see that a middle ground has been found for the directions w.r.t. all of the inputs which should lead to a more stable learning process.

E12.5.

A result of P12.2 shows that if all eigen values of W is complex, learning is stable where:

$$W = \begin{bmatrix} 0 & I \\ -\gamma I & T \end{bmatrix} \quad \text{and} \quad T = [(1+\gamma)I - (1-\gamma)\alpha A] \text{ and } A = \nabla^2 f(x)$$

It is shown that if the following inequality holds, all eigenvalues of W are complex and the learning process is stable:

$$|(1+\gamma) - (1-\gamma)\alpha\lambda_i| < 2\sqrt{\gamma}$$

to ensure this holds we try for all λ_i's from the Hessian matrix, A.

Here we have $F(x) = \frac{1}{2}x^T \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} x + [1 \ \ 2]x + 2$

$$\rightarrow A = \begin{bmatrix} 3 & 1 \\ 1 & 3 \end{bmatrix} \rightarrow \begin{cases} \lambda_1 = 2 \\ \lambda_2 = 4 \end{cases}$$

i. for $\alpha = 1$, $\gamma = 0$

$$\rightarrow \begin{cases} |(1+0) - (1-0)(1)(2)| \overset{?}{<} 2\sqrt{0} \rightarrow 1 \not< 0 \\ |(1+0) - (1-0)(1)(4)| \overset{?}{<} 2\sqrt{0} \rightarrow 3 \not< 0 \end{cases}$$

$\rightarrow$ learning is <u>not</u> stable for this combination of α and γ.

ii. for $\alpha = 1$, $\gamma = 0.6$

$$\rightarrow \begin{cases} |(1+0.6) - (1-0.6)(1)(2)| \overset{?}{<} 2\sqrt{0.6} \rightarrow 0.8 \overset{\checkmark}{<} 1.5492 \\ |(1+0.6) - (1-0.6)(1)(4)| \overset{?}{<} 2\sqrt{0.6} \rightarrow 0 \overset{\checkmark}{<} 1.5492 \end{cases}$$

learning is <u>stable</u> for this combination of α and γ.

E 12.8. $F(x) = x_1^2 + 2x_2^2$, $x_0 = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$

Params: $\alpha = 1$, $\gamma = 0.2$, $\eta = 1.5$, $\rho = 0.5$, $\xi = 5\%$

First we will evaluate the function at the initial guess:

$$F\left(\begin{bmatrix} 0 \\ -1 \end{bmatrix}\right) = 2$$

Now we calculate the gradient:

$$\nabla F(x) = \begin{bmatrix} \frac{\partial F(x)}{\partial x_1} \\ \frac{\partial F(x)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 \\ 4x_2 \end{bmatrix} , \quad g_0 = \nabla F(x) \Big|_{x = x_0} = \begin{bmatrix} 0 \\ -4 \end{bmatrix}$$

with $\alpha_0 = 1$ the first tentative step is calculated as follows:

$$\Delta x_0 = \gamma \Delta x_{-1} - (1-\gamma)\alpha g_0 = 0.2 \begin{bmatrix} 0 \\ 0 \end{bmatrix} - 0.8(1)\left(\begin{bmatrix} 0 \\ -4 \end{bmatrix}\right) = \begin{bmatrix} 0 \\ 3.2 \end{bmatrix}$$

$$\rightarrow x_1 = x_0 + \Delta x_0 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} + \begin{bmatrix} 0 \\ 3.2 \end{bmatrix} = \begin{bmatrix} 0 \\ 2.2 \end{bmatrix}$$

in order to see if we should keep this update or discard it, we test the value of the function at this potential x_1:

$$F(x_1) = F\left(\begin{bmatrix} 0 \\ 2.2 \end{bmatrix}\right) = 9.68$$

$F(x_1) > F(x_0) + \xi F(x_0) = 2.1$ $\rightarrow$ we reject this step, reduce the learning rate and set momentum to zero.

$\rightarrow x_1 = x_0$, $F(x_1) = F(x_0) = 2$, $\alpha = \rho\alpha = 0.5(1) = 0.5$, $\gamma = 0$

Now we recalculate the tentative step with momentum set to zero:

$$\Delta x_1 = -\alpha g_0 = -0.5\begin{bmatrix} 0 \\ -4 \end{bmatrix} = \begin{bmatrix} 0 \\ 2 \end{bmatrix}$$

$$\rightarrow x_2 = x_1 + \Delta x_1 = \begin{bmatrix} 0 \\ -1 \end{bmatrix} + \begin{bmatrix} 0 \\ 2 \end{bmatrix} = \begin{bmatrix} 0 \\ 1 \end{bmatrix}$$

evaluating $F(x)$ at x_2: $F(x_2) = 2$ $\rightarrow$ we accept the weight update but keep the momentum and learning rate the same.

$\rightarrow$ we calculate the gradient at x_2 : $g_2 = \nabla F(x) \Big|_{x = x_2} = \begin{bmatrix} 0 \\ 4 \end{bmatrix}$

$\rightarrow$ the third tentative step is calculated:

$$\Delta x_2 = -\alpha g_2 = -0.5\begin{bmatrix} 0 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ -2 \end{bmatrix}$$

$X_3 = X_2 + \Delta X_2 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -2 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$, $F(X_3) = 2$

→ as we can see if keep this update we will get stuck in a loop therefore we discard it and update the learning rate.

$X_3 = X_2$, $F(X_3) = F(X_2)$, $\alpha = \rho\alpha = 0.25$, $\gamma = 0$

we calculate the ~~third~~ fourth tentative step:

$$\Delta X_3 = -\alpha g_3 = -0.25 \begin{bmatrix} 0 \\ 4 \end{bmatrix} = \begin{bmatrix} 0 \\ -1 \end{bmatrix}$$

$$\rightarrow X_4 = X_3 + \Delta X_3 = \begin{bmatrix} 0 \\ 1 \end{bmatrix} + \begin{bmatrix} 0 \\ -1 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$$

evolute $F(X)$ at X_4, $F(\begin{bmatrix} 0 \\ 0 \end{bmatrix}) = 0$ → less than $F(X_3)$ → weight update is accepted, learning rate is increased, and momentum is reset

→ $X_4 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$, $\alpha = \eta\alpha = 1.5 \times 0.25 = 0.375$, $\gamma = 0.2$

---

note: the algorithm has already converged at $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$ because $\nabla F(X) \big|_{X = \begin{bmatrix} 0 \\ 0 \end{bmatrix}} = \begin{bmatrix} 0 \\ 0 \end{bmatrix}$.

E 12.11. Because the line along which we need to minimize is already given to us we do not need to calculate the gradient of $F(X)$ nor do we need an explicit initial guess.

i. To determine the interval first we calculate $F(X)$ at $\begin{bmatrix} 0 \\ 0 \end{bmatrix}$

$$\rightarrow F(\begin{bmatrix} 0 \\ 0 \end{bmatrix}) = 0$$

$$b_1 = \varepsilon = 0.5 , F(b_1) = F(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + 0.5 \begin{bmatrix} -1 \\ 0 \end{bmatrix}) = F(\begin{bmatrix} -0.5 \\ 0 \end{bmatrix}) = -3/8 = -0.375$$

$$b_2 = 2\varepsilon = 1 , F(b_2) = F(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + 1 \begin{bmatrix} -1 \\ 0 \end{bmatrix}) = F(\begin{bmatrix} -1 \\ 0 \end{bmatrix}) = -\frac{1}{2} = -0.5$$

$$b_3 = 4\varepsilon = 2 , F(b_3) = F(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + 2 \begin{bmatrix} -1 \\ 0 \end{bmatrix}) = F(\begin{bmatrix} -2 \\ 0 \end{bmatrix}) = 0$$

→ The function increases between two consecutive evaluation → minimum must occur at $[0.5, 2]$

ii. we will take one step of the golden section search to reduce the interval.

$$c_1 = a_1 + (1-\tau)(b_1 - a_1) = 0.5 + (0.382)(2-0.5) = 1.073$$

$$d_1 = b_1 - (1-\tau)(b_1 - a_1) = 2 - (0.382)(2-0.5) = 1.472$$

$$F_{a_1} = -0.375 \quad , \quad F_{b_1} = 0$$

$$F_{c_1} = F\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + 1.073 \begin{bmatrix} -1 \\ 0 \end{bmatrix}\right) = F\left(\begin{bmatrix} -1.073 \\ 0 \end{bmatrix}\right) = -0.4973$$

$$F_{d_1} = F\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix} + 1.472 \begin{bmatrix} -1 \\ 0 \end{bmatrix}\right) = F\left(\begin{bmatrix} -1.472 \\ 0 \end{bmatrix}\right) = -0.3886$$

$$\rightarrow F_{c_1} < F_{d_1} \Rightarrow a_2 = a_1 \, , \; b_2 = d_1$$

$$\rightarrow [0.5 \, , \, 2] \xrightarrow[\text{to}]{\text{reduced}} [0.5 \, , \, 1.472]$$

This procedure continues until convergence.

E12.14. $w^1(0) = \begin{bmatrix} -0.27 \\ -0.41 \end{bmatrix}$, $b^1(0) = \begin{bmatrix} -0.48 \\ -0.13 \end{bmatrix}$, $w^2(0) = \begin{bmatrix} 0.09 & -0.17 \end{bmatrix}$, $b^2(0) = \begin{bmatrix} 0.48 \end{bmatrix}$

$$P_1 = 1 \quad , \quad t_1 \approx 1.7 \quad , \qquad P_2 = 0 \quad , \quad t_2 = 1$$

First we will propagate the inputs through the network and calculate the errors.

$$a_1 = \text{logsig}(w_1 P + b_1) \quad , \quad a_2 = \text{purelin}(w_2 a_1 + b_2)$$

$$\rightarrow P_1 = 1 \rightarrow a_2^1 \approx 0.45 \, , \; e_1 = 1.7 - 0.45 = 1.25 \quad \Big| \quad a_1^1 \approx \begin{bmatrix} 0.32 \\ 0.37 \end{bmatrix}$$

$$\rightarrow P_2 = 0 \rightarrow a_2^2 \approx 0.43 \, , \; e_2 = 1 - 0.43 = 0.57 \quad \Big| \quad a_1^2 \approx \begin{bmatrix} 0.38 \\ 0.47 \end{bmatrix}$$

we can now initialize and backpropagate the marquardt sensitivities.

$$\tilde{S}_1^2 = -\dot{F}^2(n_1^2) = -1 \qquad \Big| \; F^2 : \text{purelin}$$

$$\tilde{S}_1^1 = \dot{F}^1(n_1^1)(W^2)^T \tilde{S}_1^2 = \begin{bmatrix} f^1(n_{1,1}^1)(1 - f^1(n_{1,1}^1)) & 0 \\ 0 & f^2(n_{1,2}^1)(1 - f^2(n_{1,2}^1)) \end{bmatrix} \begin{bmatrix} 0.09 \\ -0.17 \end{bmatrix} [-1]$$

$$= \begin{bmatrix} -1.3125 & 0 \\ 0 & -0.8316 \end{bmatrix} \begin{bmatrix} 0.09 \\ -0.17 \end{bmatrix} [-1] = \begin{bmatrix} 0.1181 \\ -0.1414 \end{bmatrix}$$

$$\tilde{s}_2^2 = -\dot{F}^2(n_2^2) = -1$$

$$\tilde{s}_2^1 = \dot{F}^1(n_2^1)(W^2)^T\tilde{s}_2^2 = \begin{bmatrix} -0.7104 & 0 \\ 0 & -0.7469 \end{bmatrix}\begin{bmatrix} 0.09 \\ -0.17 \end{bmatrix}[-1] = \begin{bmatrix} 0.064 \\ -0.025 \end{bmatrix}$$

$$\rightarrow \quad \tilde{S}^1 = \left[\tilde{s}_1^1 \mid \tilde{s}_2^1\right] = \begin{bmatrix} 0.1187 & 0.064 \\ -0.1414 & -0.025 \end{bmatrix}$$

$$\rightarrow \quad \tilde{S}^2 = \left[\tilde{s}_1^2 \mid \tilde{s}_2^2\right] = \begin{bmatrix} -1 & -1 \end{bmatrix}$$

we can now compute the jacobian:

$$J(x) = \begin{bmatrix} \frac{\partial e_{1,1}}{\partial w_{1,1}^1}, & \frac{\partial e_{1,1}}{\partial w_{1,2}^1}, & \frac{\partial e_{1,1}}{\partial b_1^1}, & \frac{\partial e_{1,1}}{\partial b_2^1}, & \frac{\partial e_{1,1}}{\partial w_{1,1}^2}, & \frac{\partial e_{1,1}}{\partial w_{2,1}^2}, & \frac{\partial e_{1,1}}{\partial b_1^2} \\ \frac{\partial e_{1,2}}{\partial w_{1,1}^1}, & \frac{\partial e_{1,2}}{\partial w_{1,2}^1}, & \frac{\partial e_{1,2}}{\partial b_1^1}, & \frac{\partial e_{1,2}}{\partial b_2^1}, & \frac{\partial e_{1,2}}{\partial w_{1,1}^2}, & \frac{\partial e_{1,2}}{\partial w_{2,1}^2}, & \frac{\partial e_{1,2}}{\partial b_1^2} \end{bmatrix}$$

$$[J]_{1,1} = \frac{\partial e_{1,1}}{\partial w_{1,1}^1} = \frac{\partial e_{1,1}}{\partial n_{1,1}^1} \times \frac{\partial n_{1,1}^1}{\partial w_{1,1}^1} = \tilde{S}_{1,1}^1 \times \frac{\partial n_{1,1}^1}{\partial w_{1,1}^1} = \tilde{S}_{1,1}^1 \times P_1 = 0.1187 \times 1 = 0.1187$$

$$[J]_{1,2} = \frac{\partial e_{1,1}}{\partial w_{1,2}^1} = \tilde{S}_{1,2}^1 \times P_1 = -0.1414 \times 1 = -0.1414$$

$$[J]_{1,3} = \frac{\partial e_{1,1}}{\partial b_1^1} = \tilde{S}_{1,1}^1 = 0.1187 \quad , \quad [J]_{1,4} = \frac{\partial e_{1,1}}{\partial b_1^2} = \tilde{S}_{1,2}^1 = -0.1414$$

$$[J]_{1,5} = \frac{\partial e_{1,1}}{\partial w_{1,1}^2} = \tilde{S}_{1,1}^2 \times a_{1,1}^1 = -1 \times [0.32] = -0.32$$

$$[J]_{1,6} = \frac{\partial e_{1,1}}{\partial w_{1,1}^2} = \tilde{S}_{1,1}^2 \times a_{2,1}^1 = -1 \times [0.37] = -0.37$$

$$[J]_{1,7} = \frac{\partial e_{1,1}}{\partial b_1^2} = \tilde{S}_{1,1}^2 = -1$$

The second row can be calculated similarly:

$[J]_{2,1} = 0.064 \times 0 = 0 \mid [J]_{2,2} = -0.025 \times 0 = 0 \mid [J]_{2,3} = 0.064 \mid [J]_{2,4} = -0.025$

$[J]_{2,5} = -1 \times [0.38] = -0.38 \mid [J]_{2,6} = -1 \times [0.47] = -0.47 \mid [J]_{2,7} = -1$

$$\rightarrow \quad J(x) = \begin{bmatrix} 0.1181 & -0.1414 & 0.1187 & -0.1414 & -0.32 & -0.37 & -1 \\ 0 & 0 & 0.064 & -0.025 & -0.38 & -0.47 & -1 \end{bmatrix}$$