CME 216, ME 343 - Spring 2020

Eric Darve, ICME

In the previous lecture, we explain the first step in the back-propagation algorithm.

We will derive the general formula for all layers.

Let us find the derivative with respect to $\boldsymbol{W}^{(n-1)}$ to get started.

We have:

$$\boldsymbol{y} = \boldsymbol{W}^{(n)}\phi \odot \boldsymbol{W}^{(n-1)}\boldsymbol{a}^{(n-2)}$$

Let's differentiate with respect to the (i, j) component of $W^{(n-1)}$.

Let's denote E_{ij} a matrix full of zeros with a 1 at index (i, j).

Then

$$\frac{\partial y}{\partial [W^{(n-1)}]_{ij}} = W^{(n)} \, \psi^{(n-1)} \, E_{ij} \, a^{(n-2)}$$

$\psi^{(n-1)}$ is a diagonal matrix with entries

$$[\psi^{(n-1)}]_{ii} = [\phi' \odot W^{(n-1)} a^{(n-2)}]_i$$

The scalar $W^{(n)}\,\psi^{(n-1)}\,E_{ij}\,a^{(n-2)}$ can be re-written using vectors:

$$\frac{\partial y}{\partial [W^{(n-1)}]_{ij}} = [W^{(n)}\psi^{(n-1)}]_i \; [a^{(n-2)}]_j$$

$W^{(n)}\psi^{(n-1)}$ is a row vector

$a^{(n-2)}$ is a column vector

Let's denote

$$\delta^{(n)} = \psi^{(n-1)}[\boldsymbol{W}^{(n)}]^T$$

Using matrix notations we get

$$\frac{\partial y}{\partial W^{(n-1)}} = \delta^{(n)} \left[a^{(n-2)}\right]^T$$

$\delta^{(n)}$ is a column vector

$\left[a^{(n-2)}\right]^T$ is a row vector.

$\frac{\partial y}{\partial W^{(n-1)}}$ is a matrix

Let us briefly repeat this for $\boldsymbol{W}^{(n-2)}$ to see how this works.

We now have:

$$\boldsymbol{y} = \boldsymbol{W}^{(n)}\phi \odot \boldsymbol{W}^{(n-1)}\phi \odot \boldsymbol{W}^{(n-2)}\boldsymbol{a}^{(n-3)}$$

$$\frac{\partial y}{\partial [W^{(n-2)}]_{ij}} =$$

$$= [W^{(n)} \psi^{(n-1)} W^{(n-1)} \psi^{(n-2)}]_i \, [a^{(n-3)}]_j$$

In matrix form:

$$\frac{\partial y}{\partial [W^{(n-2)}]} = \delta^{(n-1)} \, [a^{(n-3)}]^T$$

$$\delta^{(n-1)} = \psi^{(n-2)} [W^{(n-1)}]^T \psi^{(n-1)} [W^{(n)}]^T$$

We can now define the general formula.

There is a recurrence relation for $\delta^{(k)}$.

$$\delta^{(k)} = \psi^{(k-1)} \left[W^{(k)}\right]^T \delta^{(k+1)}, \qquad \delta^{(n+1)} = 1$$

$$\left[\psi^{(k-1)}\right]_{ii} = \left[\phi' \odot W^{(k-1)} a^{(k-2)}\right]_i$$

Equation for gradient with respect to matrix $W^{(k)}$

$$\frac{\partial y}{\partial W^{(k)}} = \delta^{(k+1)} \left[a^{(k-1)}\right]^T$$

Explicit expression for the bias

We previously said that the bias can be represented in this framework by adding a 1 at the end of the activation $a^{(k)}$.

With this trick, we can immediately find an explicit expression for the gradient with respect to the bias.

Gradient with respect to the bias

$$\frac{\partial y}{\partial b^{(k)}} = \delta^{(k+1)}$$

The implementation can be done in two passes called the forward and backward passes.

In the following, for completeness we will reintroduce the biases.

Forward pass

We compute all the activations.

k goes from 1 to $n-1$.

$$a^{(k)} = \phi \odot (W^{(k)} a^{(k-1)} + b^{(k)})$$

$$a^{(0)} = x$$

$\phi \odot$ means that ϕ is applied element-wise to $W^{(k)} a^{(k-1)} + b^{(k)}$.

We also need to save the values of the derivative:

$$[\psi^{(k)}]_{ii} = [\phi' \odot (W^{(k)} a^{(k-1)} + b^{(k)})]_i$$

$\psi^{(k)}$ is a diagonal matrix.

Backward pass

For k going from n to 2:

$$\delta^{(k)} = \psi^{(k-1)} \left[W^{(k)}\right]^T \delta^{(k+1)}$$

$$\delta^{(n+1)} = 1$$

Derivatives: from $k = n$ to $k = 1$

$$\frac{\partial y}{\partial W^{(k)}} = \delta^{(k+1)} \left[a^{(k-1)}\right]^T$$

$$\frac{\partial y}{\partial b^{(k)}} = \delta^{(k+1)}$$