Statistical Mechanics, Neural Networks, and Artificial Intelligence: *Using Powerful Brain Strategies to Improve AI*

DRAFT Chapter 7:
Backpropagation Part 2:
Adapting the Input-to-Hidden Connection Weights

Alianna J. Maren
Northwestern University School of Professional Studies
Master of Science in Data Science Program

Draft: 2020-01-06

7.1 Dependence of SSE on All Output Nodes

As a next step, we want to compute the dependence of the summed squared error on the input-to-hidden weights.

The key difference between this calculation and the previous one is that earlier, when we were interested in the dependence of the summed squared error (SSE) on a specific hidden-to-output connection weight $v_{h,o}$, we only needed to think about the squared error at output node O_o. The other output nodes did not impact the dependence of the SSE on this output node.

In contrast, this time, when we want to consider the dependence of the SSE on a given input-to-hidden connection weight $w_{i,h}$, we now have to think about the influence of the squared errors from *each* of the output nodes. This is because the input-to-hidden connection weight $w_{i,h}$ influences the activation (resulting output) of hidden node H_h. The activation of this hidden node, however, impacts all of the output nodes. Therefore, we need to consider the SSE at all the output nodes, and their back-propagated influence on H_h, and from that hidden node to $w_{i,h}$. This is shown in Figure 7.1.

We thus desire

$$\frac{\partial SSE}{\partial w_{i,h}} = \sum_{o=1}^{O} \Big[\frac{\partial SE}{\partial E_o} \frac{\partial E_o}{\partial w_{i,h}} \Big] = \sum_{o=1}^{O} \Big[\frac{\partial SE}{\partial E_o} \frac{\partial E_o}{\partial A_o} \frac{\partial A_o}{\partial w_{i,h}} \Big] \tag{7.1}$$

Thus far, we are identical with the previous initial step in backpropagation, given in Eqn. **??**, with the exception that the last term involves the dependence of the output activation on a specific input-to-hidden weight, instead of hidden-to-output weight. Also, we are summing over the full set of squared errors at each of the output nodes.

Before we go further, we're going to write the dependence of the activation (actual output) at output node O_o on the input-to-hidden connection weight connecting input node I_i to hidden node H_h. That is, we will write the expression for

$$\frac{\partial A_o}{\partial w_{i,h}} \tag{7.2}$$

We know already that the activation (actual output) of the output node o is a result of the transfer function being applied to the summed inputs to that node.

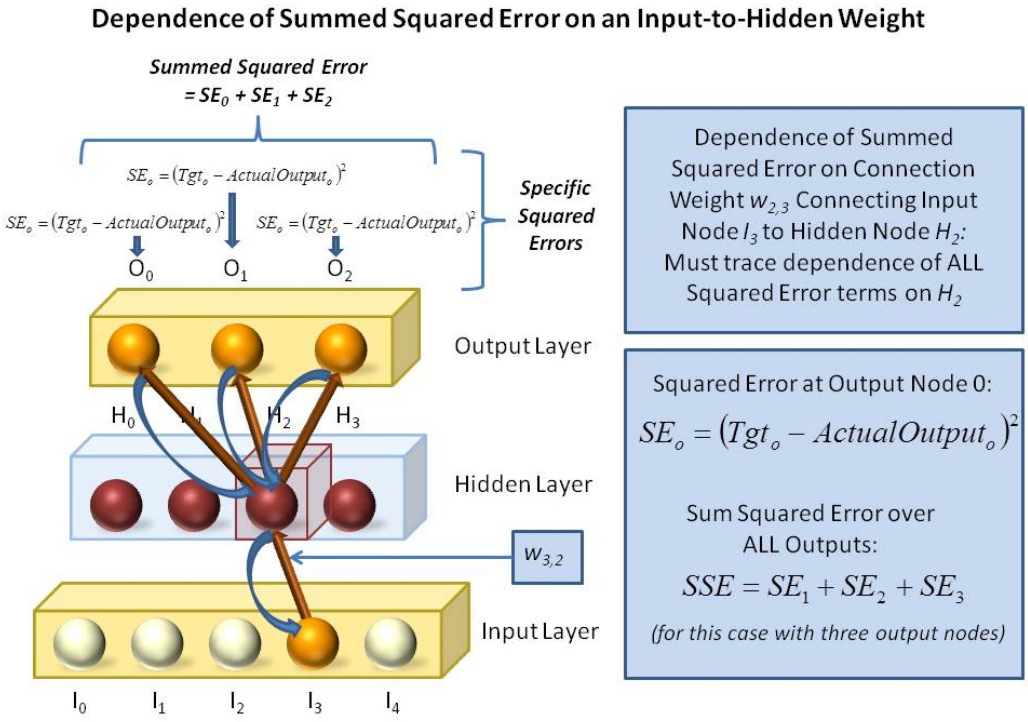

Figure 7.1: Dependence of the Summed Squared Error (SSE) on the input-to-hidden connection weight $w_{i,h}$ requires tracing the dependence of the squared error at each output node through hidden node h and from there to the connection weight $w_{i,h}$. This is shown for the case where the connection weight is between input node 3 and hidden node 2, $w_{3,2}$, where the nodes in each layer are numbered in the manner of Python code, beginning with the zeroth node in each layer.

$$A_o = \mathscr{F}(NdInpt_o) \tag{7.3}$$

Now, what we need is to work through the dependence of A_o on the connection weight $w_{i,h}$ using this relation.

$$\frac{\partial A_o}{\partial w_{i,h}} = \frac{\partial \mathscr{F}(NdInpt_o)}{\partial w_{i,h}} = \frac{\partial \mathscr{F}(NdInpt_o)}{\partial (NdInpt_o)} \frac{\partial (NdInpt_o)}{\partial w_{i,h}} \tag{7.4}$$

We have previously found the dependence of the transfer function $\mathscr{F}$ on the node input; we already know that

$$\frac{\partial \mathscr{F}(NdInpt_o)}{\partial NdInpt_o} = \alpha \mathscr{F}_o \big[1 - \mathscr{F}_o\big] \tag{7.5}$$

We can substitute this into Eqn. 7.4 to obtain

$$\frac{\partial A_o}{\partial w_{i,h}} = \frac{\partial \mathscr{F}(NdInpt_o)}{\partial w_{i,h}} = \alpha \mathscr{F}_o \big[1 - \mathscr{F}_o\big] \frac{\partial (NdInpt_o)}{\partial w_{i,h}} \tag{7.6}$$

Now, we need to find

$$\frac{\partial (NdInpt_o)}{\partial w_{i,h}} \tag{7.7}$$

To do this, we recall from Eq. **??** that

$$NdInpt_o = \sum_{h=1}^{H} v_{h,o} * H_h. \tag{7.8}$$

We substitute this into Eqn. 7.7, while changing the index on the summation from h to q, to obtain

$$\frac{\partial (NdInpt_o)}{\partial w_{i,h}} = \frac{\partial \Big(\sum\limits_{q=1}^{H} v_{q,o} * H_q \Big)}{\partial w_{i,h}} \tag{7.9}$$

Since the dependence holds only in the case where $q = h$, we can simplify this as

$$\frac{\partial (NdInpt_o)}{\partial w_{i,h}} = v_{h,o} \frac{\partial H_h}{\partial w_{i,h}} \tag{7.10}$$

We can write an expression for the activation of the hidden node H_h as

$$H_h = \mathscr{F}(\sum_{i=1}^{I} w_{i,h} * Input_i) = \mathscr{F}(NdInput_h) \tag{7.11}$$

$$H_h = \mathscr{F}(\sum_{i=1}^{I} w_{i,h} * Input_i) = \mathscr{F}(NdInput_h) \tag{7.12}$$

We can write the node inputs to hidden node h as

$$NdInpt_h = \sum_{i=1}^{I} w_{i,h} * Input_i. \tag{7.13}$$

We can substitute from Eqn. 7.13 into Eqn. 7.12 to obtain

$$H_h = \mathscr{F}(NdInput_h) = \mathscr{F}\Big(\sum_{i=1}^{I} w_{i,h} * Input_i\Big) \tag{7.14}$$

We can substitute this expression for H_h into Eqn. 7.8 to obtain

$$NdInpt_o = \sum_{h=1}^{H} v_{h,o} * H_h = \sum_{h=1}^{H} v_{h,o} * \mathscr{F}\Big(\sum_{i=1}^{I} w_{i,h} * Input_i\Big). \tag{7.15}$$

Now, let's use this expression to figure out the dependence of the node inputs at output node o on the input-to-hidden connection weight $w_{i,h}$. Specifically, going back to Eqn. 7.7, we introduce a substitution based on Eqn. 7.15 and write

$$\frac{\partial(NdInpt_o)}{\partial w_{i,h}} = \frac{\partial\Big[\sum_{h=1}^{H} v_{h,o} * \mathscr{F}\Big(\sum_{i=1}^{I} w_{i,h} * Input_i\Big)\Big]}{\partial w_{i,h}} \tag{7.16}$$

The hidden-to-output connection weights $v_{h,o}$ are a constant with regard to the specific input-to-hidden connection weight $w_{i,h}$, and so we can take these terms (along with their summation) outside of the partial derivative. We also want to distinguish between the sum over all possible hidden nodes and the specific one with which we want to compute a dependence, and so we change the running index on the sum over the hidden nodes to be q. Thus we write

$$\frac{\partial(NdInpt_o)}{\partial w_{i,h}} = \sum_{q=1}^{H} v_{q,o} \frac{\partial\Big[\mathscr{F}\Big(\sum\limits_{i=1}^{I} w_{i,q} * Input_i\Big)\Big]}{\partial w_{i,h}} \tag{7.17}$$

We know that the transfer function is being taken, in each case, at hidden node q, and we know that the derivative of this is given as

$$\begin{aligned}
\frac{\partial\Big[\mathscr{F}\Big(\sum\limits_{i=1}^{I} w_{i,q} * Input_i\Big)\Big]}{\partial w_{i,h}} &= \frac{\partial\mathscr{F}(NdInput_q)}{\partial NdInput_q}\frac{\partial NdInput_q}{\partial w_{i,h}} \\
&= \alpha\mathscr{F}_q(1-\mathscr{F}_q) * \frac{\partial NdInput_q}{\partial w_{i,h}} \\
&= \alpha\mathscr{F}_q(1-\mathscr{F}_q) * \frac{\partial\Big(\sum\limits_{i=1}^{I} w_{i,q} * Input_i\Big)}{\partial w_{i,h}} \\
&= \alpha\mathscr{F}_q(1-\mathscr{F}_q)Input_i
\end{aligned} \tag{7.18}$$

We note that this dependence occurs only when $w_{i,q} = w_{i,h}$ (all other terms drop out in the last sum), so that $q = h$, and we can rewrite the previous Eqn. 7.18 as

$$\frac{\partial\Big[\mathscr{F}\Big(\sum\limits_{i=1}^{I} w_{i,q} * Input_i\Big)\Big]}{\partial w_{i,h}} = \alpha\mathscr{F}_h(1-\mathscr{F}_h)Input_i \tag{7.19}$$

We return now to the initial equation for backpropagating the dependence of the summed squared error SSE on a given input-to-hidden connection weight, Eqn. 7.1, and substitute what we have gained in this last step to obtain

$$\begin{aligned}\frac{\partial SSE}{\partial w_{i,h}} &= \sum_{o=1}^{O}\left[\frac{\partial SE}{\partial E_o}\frac{\partial E_o}{\partial w_{i,h}}\right] = \sum_{o=1}^{O}\left[\frac{\partial SE}{\partial E_o}\frac{\partial E_o}{\partial A_o}\frac{\partial A_o}{\partial w_{i,h}}\right] \\ &= \sum_{o=1}^{O}\left[\frac{\partial SE}{\partial E_o}\frac{\partial E_o}{\partial A_o}\alpha v_{h,o}\mathscr{F}_h(1-\mathscr{F}_h)Input_i\right] \\ &= \alpha\sum_{o=1}^{O}\left[\frac{\partial SE}{\partial E_o}\frac{\partial E_o}{\partial A_o}v_{h,o}\mathscr{F}_h(1-\mathscr{F}_h)Input_i\right]\end{aligned} \tag{7.20}$$

We can quickly obtain the dependence of the squared error on the error (first term inside the summation) and the dependence of the error on the activation (second term inside the summation), using results previously obtained, referencing Eqn. **??** which had previously given us

$$\frac{\partial SSE}{\partial E_o} = \frac{1}{2}\frac{\partial E_{q=o}^2}{\partial E_o} = E_o \tag{7.21}$$

and also Eqn. **??**

$$\frac{\partial E_o}{\partial \mathscr{F}_o} = -1. \tag{7.22}$$

so that we can write

$$\begin{aligned}\frac{\partial SSE}{\partial w_{i,h}} &= \alpha\sum_{o=1}^{O}\left[\frac{\partial SE}{\partial E_o}\frac{\partial E_o}{\partial A_o}v_{h,o}\mathscr{F}_h(1-\mathscr{F}_h)Input_i\right] \\ &= \alpha\sum_{o=1}^{O}\left[E_o\frac{\partial E_o}{\partial A_o}v_{h,o}\mathscr{F}_h(1-\mathscr{F}_h)Input_i\right] \\ &= \alpha\sum_{o=1}^{O}\left[E_o(-1)v_{h,o}\mathscr{F}_h(1-\mathscr{F}_h)Input_i\right] \\ &= -\alpha\sum_{o=1}^{O}\left[E_o v_{h,o}\mathscr{F}_h(1-\mathscr{F}_h)Input_i\right]\end{aligned} \tag{7.23}$$

However, while the summation is over all output nodes, the terms involving the specific hidden node involved in the dependence of the SSE on the input-to-hidden connection weight $w_{i,h}$, together with the actual value of the input

node i, that is, $Input_i$, are constants relative to the output error terms. We can thus pull them out of the summation and write

$$\begin{aligned} \frac{\partial SSE}{\partial w_{i,h}} &= -\alpha \sum_{o=1}^{O} \Big[E_o v_{h,o} \mathscr{F}_h (1 - \mathscr{F}_h) Input_i \Big] \\ &= -\alpha \mathscr{F}_h (1 - \mathscr{F}_h) Input_i \sum_{o=1}^{O} v_{h,o} E_o \end{aligned} \tag{7.24}$$

As previously mentioned, this method was originally developed by Paul Werbos and presented in his Ph.D. dissertation at Harvard University [1].

Bibliography

[1] P. J. Werbos, *Beyond Regression: New Tools for Prediction and Analysis in the Behavioral Sciences*. PhD thesis, Harvard University, 1974.