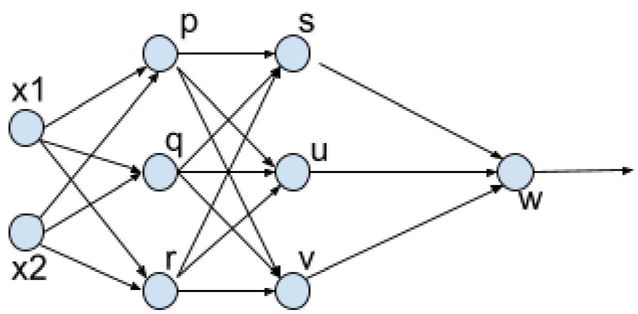

We assume the above network has sigmoid activation: $\sigma(x) = 1/(1 + e^{-x})$.

What is the loss of the above network?
Start with the representation of a datapoint x in the first and second layers and then write the final least squares loss (similar to what we did for the single layer network).

What are the gradient updates of the above network?
We have unknown weights in the final layer $w = (w_1, w_2, w_3)$, weights in the second to last layer $s = (s_1, s_2, s_3), u = (u_1, u_2, u_3), v = (v_{1,}, v_2, v_3)$, and weights in the first layer $p = (p_1, p_2), q = (q_1, q_2), r = (r_1, r_2)$.

To make our calculations easier to understand and perhaps rewrite them as matrix products we let $h = (h_1, h_2, h_3)$ be the representation of x in the first layer of the network and let $z = (z_1, z_2, z_3)$be the representation of x in the second layer.

We have $h_1 = \sigma(p^T x)$, $h_2 = \sigma(q^T x)$, $h_3 = \sigma(r^T x)$. Then we have $z_1 = \sigma(s_1 \sigma(p^T x\ ) + s_2\ \sigma(q^T x) + s_3 \sigma(r^T x))$ which I can also write as $z_1 = \sigma(s_1 h_1 + s_2\ h_2 + s_3 h_3)$. Similarly we can calculate $z_2$and z_3.

This means I can write the final loss f as $f = ((w_1, w_2, w_3)^T (z_1, z_2, z_3) - y)^2$.

<u>Final output gradient:</u>

For the gradient updates we have

$df/dw_1 = 2\sqrt{(f)} z_1$ => same as $df/dw_1 = 2((w_1, w_2, w_3)^T (z_1, z_2, z_3) - y)\ z_1$

Thus we can write df/dw as

$$df/dw\ = (2((w_1, w_2, w_3)^T (z_1, z_2, z_3) - y))(\ z_1, z_2, z_3)$$

<u>Second to last layer gradient:</u>

For the second to last layer gradient updates we need df/ds, df/du, and df/dv. Let us calculate df/ds_1.

We have already defined the coordinates of z above. For example $z_1 = \sigma(s_1\sigma(p^T x\ ) + s_2\ \sigma(q^T x) + s_3\sigma(r^T x))$. We can rewrite z1 as $z_1 = \sigma(s_1 h_1 + s_2\ h_2 + s_3 h_3)$ where $h_1 = \sigma(p^T x),\ h_2 = \sigma(q^T x),\ and\ h_3 = \sigma(r^T x)$.

Now we are in a better shape to calculate $df/ds_1 = (df/dz_1)(dz_1/ds_1)$

We have $df/dz_1 = 2\sqrt{(f)}w_1$and we have
$dz_1/ds_1 = d\sigma/ds_1 = \sigma(s_1 h_1 + s_2\ h_2 + s_3 h_3)(1 - \sigma(s_1 h_1 + s_2\ h_2 + s_3 h_3))h_1 = z_1(1 - z_1)h_1$. Thus
$df/ds_1 = 2\sqrt{(f)}w_1(z_1(1 - z_1)h_1)$since $d\sigma/df(x) = \sigma(f(x))(1 - \sigma(f(x))df/dx$

Similarly we can get $df/ds_2$and df/ds_3.

First layer gradient:

The final step is to do $df/dp_1 = (df/dz_1)(dz_1/dh_1)(dh_1/dp_1)$. We already have some components worked out:

$df/dz_1 = 2\sqrt{(f)}w_1$

$dz_1/dh_1 = z_1(1 - z_1)s_1$

$dh_1/dp_1 = \sigma(p^T x)(1 - \sigma(p^T x))x_1 = h_1(1 - h_1)x_1$