Notation Glossary:

z_j^l	Pre-Activation of j^{th} neuron in the l^{th} layer
a_j^l	(Post)-Activation of j^{th} neuron in the l^{th} layer
x_k	Input (0^{th}) layer activations
$A()$	General element-wise activation function
w_{jk}^l	Weight matrix of l^{th} layer
b_j^l	Bias vector of l^{th} layer
g_{jk}^l	Jacobian between layer l and $l-1$
J_{jk}^l	Jacobian between input (0^{th}) layer and layer l
H_{jk}^l	Diagonal Hessian between input (0^{th}) layer and layer l
$\odot$	Direct product over repeated index rather than summation

Network Jacobian:

In following derivations assume the standard feed forward relationships:

$$a_j^l = A(z_j^l) \tag{1}$$

$$z_j^l = \sum_k w_{jk}^l a_k^{l-1} + b_j^l \tag{2}$$

Note that the network's output corresponds to the L^{th} layer activations a_j^L. Begin by considering the layer to layer Jacobian:

$$\frac{\partial a_j^l}{\partial a_k^{l-1}} = \sum_q \frac{\partial a_j^l}{\partial z_q^l} \frac{\partial z_q^l}{\partial a_k^{l-1}} \tag{3}$$

$$= \sum_q \frac{\partial a_j^l}{\partial z_q^l} w_{qk}^l \tag{4}$$

$$g_{jk}^l = A'(z_j^l) \odot w_{jk}^l \tag{5}$$

Going 1 more layer back:

$$\frac{\partial a_j^l}{\partial a_k^{l-2}} = \sum_q \left[(A'(z_j^l) \odot w_{jq}^l)(A'(z_q^{l-1}) \odot w_{qk}^{l-1}) \right] \tag{6}$$

$$= \sum_q g_{jq}^l g_{qk}^{l-1} \tag{7}$$

This relationship can be recurred to obtain:

$$J_{jk}^l \equiv \frac{\partial a_j^l}{\partial x_k} = \sum_{\alpha,\beta...} g_{j\alpha}^l g_{\alpha\beta}^{l-1} ... g_{\mu k}^1 \tag{8}$$

Network Hessian:

The general l^{th} layer Hessian can be expressed as:

$$H_{jkm}^l \equiv \frac{\partial^2 a_j^l}{\partial x_m \partial x_k} = \frac{\partial J_{jk}^l}{\partial x_m} \tag{9}$$

For constructing Physics Loss functions, often only the diagonal Hessian is needed. Let:

$$\sum_k \delta_{mk} H^l_{jkm} \equiv H^l_{jk} \tag{10}$$

A further useful property of the full Hessian is its symmetry in the 2^{nd} and 3^{rd} indices (by property of partial derivatives):

$$H_{jkm} = H_{jmk} \tag{11}$$

We can also note the useful base cases:

$$J^0_{jk} = \delta_{jk} \tag{12}$$

$$H^0_{jkm} = 0 \tag{13}$$

To derive the full Hessian we consider a recursive method and start from:

$$J^l_{jk} \equiv \sum_q g^l_{jq} J^{l-1}_{qk} = \sum_q \left[A'(z^l_j) \odot w^l_{jq} J^{l-1}_{qk} \right] \tag{14}$$

Considering the Hessian as the derivative of the Jacobian:

$$H^l_{jkm} = \frac{\partial}{\partial x_m} \sum_q \left[A'(z^l_j) \odot w^l_{jq} J^{l-1}_{qk} \right] \tag{15}$$

$$= \sum_q \left[A''(z^l_j) \odot \frac{\partial z^l_j}{\partial x_m} \odot w^l_{jq} J^{l-1}_{qk} + A'(z^l_j) \odot w^l_{jq} \frac{\partial J^{l-1}_{qk}}{\partial x_m} \right] \tag{16}$$

We can recover several familiar terms from this expression, starting with

$$\frac{\partial J^{l-1}_{qk}}{\partial x_m} \equiv H^{l-1}_{qkm}, \tag{17}$$

and likewise, simplifying the first term:

$$\frac{\partial z^l_j}{\partial x_m} = \frac{\partial}{\partial x_m} \sum_p \left[w^l_{jp} a^{l-1}_p + b^l_j \right] = \sum_p w^l_{jp} \frac{\partial a^{l-1}_p}{\partial x_m} = \sum_p w^l_{jp} J^{l-1}_{pm} = \frac{J^l_{jm}}{A'(z^l_j)} \tag{18}$$

Putting these expressions together we yield a recursion relation for the full Hessian:

$$H^l_{jkm} = \frac{A''(z^l_j)}{A'(z^l_j)} \odot J^l_{jm} \sum_q w^l_{jq} J^{l-1}_{qk} + A'(z^l_j) \sum_q w^l_{jq} H^{l-1}_{qkm} \tag{19}$$

$$= \frac{A''(z^l_j)}{\left[A'(z^l_j)\right]^2} \odot J^l_{jm} \odot J^l_{jk} + A'(z^l_j) \sum_q w^l_{jq} H^{l-1}_{qkm} \tag{20}$$

Finally, we can contract over $m \,\&\, k$ to obtain the diagonal Hessian of the l^{th} layer activations (with respect to the inputs x_k:

$$H^l_{jk} = \frac{A''(z^l_j)}{\left[A'(z^l_j)\right]^2} \odot (J^l_{jk})^2 + A'(z^l_j) \sum_q w^l_{jq} H^{l-1}_{qk} \tag{21}$$

Although the above forms are probably the more succinct and interpretable, the divisions by A' prove problematic when activation functions have stationary points. Undoing some of the substitutions we can return to more computationally stable expressions:

$$H^l_{jkm} = A''(z^l_j)\left[\sum_q w^l_{jq} J^{l-1}_{qm}\right]\left[\sum_q w^l_{jq} J^{l-1}_{qk}\right] + A'(z^l_j)\sum_q w^l_{jq} H^{l-1}_{qkm} \tag{22}$$

$$H^l_{jk} = A''(z^l_j)\left[\sum_q w^l_{jq} J^{l-1}_{qk}\right]^2 + A'(z^l_j)\sum_q w^l_{jq} H^{l-1}_{qk} \tag{23}$$

Jacobian and Hessian Loss Derivatives:

If we construct a Physics Loss using Jacobian and/or diagonal Hessian terms, we will need to propagate the according loss derivative through the network (standard backpropagation). This will require: $\frac{\partial J^L_{jk}}{\partial a^L_m}$ and $\frac{\partial H^L_{jk}}{\partial a^L_m}$. It turns out that both terms can be obtained quite easily using the chain rule:

$$\frac{\partial J^L_{jk}}{\partial a^L_m} = \sum_q \frac{\partial x_q}{\partial a^L_m}\frac{J^L_{jk}}{\partial x_q} \tag{24}$$

$$= \sum_q \frac{H^L_{jkq}}{J^L_{mq}} \tag{25}$$

Similarly for the Hessian:

$$\frac{\partial H^L_{jk}}{\partial a^L_m} = \sum_q \frac{T^L_{jkq}}{J^L_{mq}} \tag{26}$$

Where we have let $T^L_{jkq} \equiv \frac{\partial H^L_{jk}}{\partial x_q}$ which turns out to be reasonably straightforward to derive recursively. Bearing in mind the issues of computational stability, we begin the derivation from Eqn. (22):

$$T^l_{jkm} = \frac{\partial}{\partial x_m}\left[A''(z^l_j)\left[\sum_q w^l_{jq} J^{l-1}_{qk}\right]^2 + A'(z^l_j)\sum_q w^l_{jq} H^{l-1}_{qk}\right] \tag{27}$$

$$= A'''(z^l_j)\frac{\partial z^l_j}{\partial x_m}\left[\sum_q w^l_{jq} J^{l-1}_{qk}\right]^2 + 2A''(z^l_j)\sum_q w^l_{jq} H^{l-1}_{qkm} \tag{28}$$

$$+ A''(z^l_j)\frac{\partial z^l_j}{\partial x_m}\sum_q w^l_{jq} H^{l-1}_{qk} + A'(z^l_j)\sum_q w^l_{jq} T^{l-1}_{qkm} \tag{29}$$

$$= \left[\sum_q w^l_{jq} J^{l-1}_{qm}\right]\left[A'''(z^l_j)\left[\sum_q w^l_{jq} J^{l-1}_{qk}\right]^2 + A''(z^l_j)\sum_q w^l_{jq} H^{l-1}_{qk}\right] \tag{30}$$

$$+ 2A''(z^l_j)\sum_q w^l_{jq} H^{l-1}_{qkm} + A'(z^l_j)\sum_q w^l_{jq} T^{l-1}_{qkm} \tag{31}$$

Where the relationship in Eqn. (18) has been used to substitute the $\frac{\partial z^l_j}{\partial x_m}$ terms.