APPENDIX

6.1. Proof of Hessian Chain Rule

Our goal is to compute the Hessian with respect to the parameters in the layer k. By the chain rule

$$D_{w^{(k)}} L = D_{z^{(n)}} L \bullet D_{w^{(k)}} z^{(n)} \tag{16}$$

Note that the second tensor is of shape $[d_n, d_{k+1}, d_k]$ (rank 3!), the contraction is over the dimension of $z^{(n)}$. Again by the chain and product rules

$$D^2_{w^{(k)}} L = \underbrace{D^2_{z^{(n)}} L \bullet D_{w^{(k')}} z^{(n)} \bullet D_{w^{(k)}} z^{(n)}}_{H_1} + \underbrace{D_{z^{(n)}} L \bullet D^2_{w^{(k)}} z^{(n)}}_{H_2} \tag{17}$$

In the component H_1 the dot-products contract indices $z^{(n)}$ (note that D^2 is symmetric and the terms D are same, hence the order of pairing dimensions of $w^{(k)}$ does not matter). As for the second component H_2, it is a product of tensors of rank 1 and 5. In order to further simplify, we are going to show that H_2 negligible compared to H_1. the intuition is as follows: in H_1 the contribution comes from gradients $D_{w^{(k)}}$ while in H_2 from second-order derivatives $D^2_{w^{(k)}}$; we consider activations such that $f(u) = au + O(u^3)$ and therefore for small u second-derivatives are near zero but first derivatives are not, and their contributions dominate.

In the analysis below we assume that weights are sufficiently small, and biasses are zero (or of much smaller variance compared to weights). Let $u^{(k)} = w^{(k)} \cdot z^{(k)} + b^{(k)}$ be the output before activation at the k-th layer.

Due to Equation (17), our goal is to evaluate first and second derivatives of $z^{(n)}$ with respect to weights $w^{(k)}$, under the assumption that inputs $z^{(i)}$ are sufficiently small. Consider how $z^{(k+1)}$ depends on $w^{(k)}$. By the chain rules

$$D_{w^{(k)}} z^{(k+1)} = D_{u^{(k)}} f(u^{(k)}) \bullet D_{w^{(k)}} u^{(k)} \tag{18}$$

$$D^2_{w^{(k)}} z^{(k+1)} = D^2_{u^{(k)}} f(u^{(k)}) \bullet D_{w^{(k)}} u^{(k)} \bullet D_{w^{(k)}} u^{(k)} \tag{19}$$

Note that $D_{u^{(k)}} f(u^{(k)})$ and $D^2_{u^{(k)}} f(u^{(k)})$ are diagonal tensors because f is applied element-wise. More precisely

$$\left[D^2_{u^{(k)}} f(u^{(k)})\right]_{i,j,j'} = \delta_{i,j}\delta_{i,j'} \cdot f''(u_i^{(k)}) \tag{20}$$

where $\delta_{\cdot,\cdot}$ is the Kronecker delta which is one where indices match and zero otherwise. Moreover,

$$\left[D_{w^{(k)}} u^{(k)}\right]_{j,p,q} = \frac{\partial}{\partial w^{(k)}_{p,q}} (w^{(k)} \cdot z^{(k)} + b^{(k)})_j = \delta_{j,p} \cdot z_q^{(k)} \tag{21}$$

Thus

$$\left[D^2_{w^{(k)}} z^{(k+1)}\right]_{i,p,q,p',q'} = \delta_{i,p}\delta_{i,p'} \cdot f''(u_i^{(k)}) \cdot z_q^{(k)} \cdot z_{q'}^{(k)} \tag{22}$$

When this tensor acts, as a bi-linear form, on a tensor $g = g_{p,q}$ we therefore obtain

$$\left[D^2_{w^{(k)}} z^{(k+1)} \bullet g \bullet g\right]_i = f''(u_i^{(k)}) \sum_{q,q'} z_q^{(k)} z_{q'}^{(k)} g_{i,q} g_{i,q'} \tag{23}$$

$$= f''(u_i^{(k)}) \left(\sum_q z_q^{(k)} g_{i,q}\right)^2 \tag{24}$$

Since our assumption on activations implies $f''(u) = O(f''' \cdot u)$ for real u, we see this is of order $O(f'''\|u^{(k)}\|\|z^{(k)}\|^2) \cdot \|g\|^2$.

Claim 1 (Magnitude of second derivative of weights)**.**

$$D^2_{w^{(k)}} z^{(k+1)} \bullet g \bullet g = O(f'''\|u^{(k)}\|\|z^{(k)}\|^2) \cdot \|g\|^2. \tag{25}$$

which is of order $O(f''' \cdot c^3)$ where c is the constant from our 'relatively small inputs' assumption.

Next, observe that the roles of $z^{(k)}$ and $w^{(k)}$ in $u^{(k)}$ are symmetric. Thus we have a similar result with respect to $z^{(k)}$.

Claim 2 (Magnitude of second derivative of inputs)**.**

$$D^2_{z^{(k)}} z^{(k+1)} \bullet g \bullet g = O(f'''\|u^{(k)}\|\|w^{(k)}\|^2) \cdot \|g\|^2 \tag{26}$$

which is of order $O(f''' \cdot c)$ where c is the constant from our 'relatively small inputs' assumption.

We need to prove that this propagates to higher-level outputs z^i, where $i > k$. This is intuitive, considering now that $z^{(i)}$ is a function of $z^{(k+1)}$ with no dependencies on $w^{(k)}$. To prove it formally look at the second-order chain rule

$$D^2_{w^{(k)}} z^{(i)} = D^2_{z^{(k+1)}} z^{(i)} \bullet D_{w^{(k)}} z^{(k+1)} \bullet D_{w^{(k)}} z^{(k+1)} + D_{z^{(k+1)}} z^{(i)} \bullet D^2_{w^{(k)}} z^{(k+1)} \tag{27}$$

Now the second term is clearly $O(f'''c^3)$ by the first claim. As for the first term, the first tensor is of order $O(f'''c)$ while the two others are $O(f'c)$. The dot-product gives the bound $O(f'''c^3)$.

Claim 3. *For every $i > k$ it holds $D^2_{w^{(k)}} z^{(i)} = O(f''c^3)$.*

Summing up, we can ignore second-derivatives with respect to weights, and this is accurate except third-order terms in the magnitude of $z^{(i)}$. In particular, we can ignore the effect of H_2.

6.2. Factorizing the Hessian Quadratic Form

Consider any potential update vector g for weights $w^{(k)}$, it has to be of same shape as $D_{w^{(k)}} L$ or quivalently $w^{(k)}$, that

is $[d_{k+1}, d_k]$. Our goal is to evaluate the hessian quadratic form on g. Ignoring the smaller part H_2 we are left with H_1 which gives

$$D^2_{w^{(k)}} L \bullet g \bullet g = D^2_{z^{(n)}} L \bullet D_{w^{(k)}} z^{(n)} \bullet D_{w^{(k)}} z^{(n)} \bullet g \bullet g \tag{28}$$

where g is contracted on all indices together with $w^{(k)}$. To emphasize this we can regroup, obtaining

Claim 4 (Hessian quadratic form)**.** *The hessian quadratic form for an update g equals*

$$D^2_{w^{(k)}} L \bullet g \bullet g = D^2_{z^{(n)}} L \bullet \left(D_{w^{(k)}} z^{(n)} \bullet g\right) \bullet \left(D_{w^{(k)}} z^{(n)} \bullet g\right) \tag{29}$$

We work further to simplify rank-3 tensors. By the chain rule

$$D_{w^{(k)}} z^{(n)} \bullet g = D_{z^{(k+1)}} z^{(n)} \bullet D_{w^{(k)}} z^{(k+1)} \bullet g \tag{30}$$

Let $u^{(k)} = w^{(k)} \cdot z^{(k)} + b^{(k)}$ be the output before activation. By the chain rule

$$D_{w^{(k)}} z^{(k+1)} = D_{u^{(k)}} z^{(k+1)} \bullet D_{w^{(k)}} u^{(k)} \tag{31}$$

Note that $D_{w^{(k)}} u^{(k)}$ is a third-order tensor of shape $[d_{k+1}, d_{k+1}, d_k]$. Denote its elements by $M_{i',i,j}$. We have

$$\left[D_{w^{(k)}} u^{(k)}\right]_{i',i,j} = [i' = i] \cdot z^{(k)}_j \tag{32}$$

and we compute the dot product

$$\left[D_{w^{(k)}} u^{(k)} \bullet g\right]_{i'} = \sum_{i,j} \left[D_{w^{(k)}} u^{(k)}\right]_{i',i,j} g_{i,j} = \sum_j z^{(k)}_j g_{i',j} \tag{33}$$

which can be expressed compactly in terms of matrix multiplication as

$$D_{w^{(k)}} u^{(k)} \bullet g = g \cdot z^{(k)} \tag{34}$$

which is a vector of shape $[d_{k+1}]$. Using this in Equation (31) we obtain, in terms of matrix products

$$D_{w^{(k)}} z^{(k+1)} \bullet g = D_{u^{(k)}} z^{(k+1)} \cdot \left(D_{w^{(k)}} z^{(k+1)} \bullet g\right) = D_{u^{(k)}} z^{(k+1)} \cdot g_i \cdot z^{(k)} \tag{35}$$

Now, in terms of matrix products, Equation (30) becomes

$$D_{w^{(k)}} z^{(n)} \bullet g = D_{z^{(k+1)}} z^{(n)} \cdot D_{w^{(k)}} z^{(k+1)} \cdot g \cdot z^{(k)} \tag{36}$$

which is a vector of shape $[d_n]$. Finally note that $H \bullet v \bullet v = v^T \cdot H \cdot v$ where H is a symmetric matrix and v is vector. This proves

Claim 5 (Approximated hessian form)**.** *For sufficiently small inputs, the hessian quadratic form can be approximated as*

$$\mathbf{H}_{w^{(k)}}[g, g] \approx v^T \cdot \mathbf{H}_z \cdot v \tag{37}$$

where

$$v = D_{z^{(k+1)}} z^{(n)} \cdot D_{w^{(k)}} z^{(k+1)} \cdot g \cdot z^{(k)}. \tag{38}$$

This claim implies the first part of Theorem 1. The error estimate follows by follows by the discussion in the previous subsection.

6.3. Further Factorization

We have seen that the quadratic effects of $w^{(k)}$ can be ignored, thus it is enough to consider the simplified recursion

$$z^{(k+1)} \approx a \cdot \left(w^{(k)} \cdot z^{(k)} + b^{(k)}\right) \tag{39}$$

for a diagonal matrix a (which captures first-derivatives of activations), or equivalently:

Claim 6 (Second-order recursion for small inputs)**.** *For relatively small inputs the hessian can be computed under the simplified recursion*

$$z^{(k+1)} \approx \mathbf{J}^{(k)} \cdot z^{(k)} \tag{40}$$

where $\mathbf{J}^k = D_{z^{(k)}} z^{(k+1)}$. In particular the term H_2 can be ignored.

We now proceed to further factorize v. By linearization we obtain

$$z^{(k)} \approx D_{z^{(k-1)}} z^{(k)} \cdot z^{(k-1)} = \ldots = D_{z^{(k-1)}} z^{(k)} \tag{41}$$

Moreover, by the chain rule, output gradient facatorizes as

$$D_{z^{(k+1)}} z^{(n)} = D_{z^{(n-1)}} z^{(n)} \cdot z^{(k-1)} \cdot D_{z^{(n-2)}} z^{(n-1)} \cdot \ldots \cdot D_{z^{(k+1)}} z^{(k+2)} \tag{42}$$

regardless of linearizing assumptions. Combining these two observations gives

Claim 7 (Factorizing under linearization)**.** *Up to terms linear in $z^{(i)}$ for $i = k-1, \ldots, 1$ we have*

$$v \approx D_{z^{(n-1)}} z^{(n)} \cdot D_{z^{(n-2)}} z^{(n-1)} \cdot \ldots \cdot D_{z^{(k+1)}} z^{(k+2)} \cdot A \cdot g \cdot D_{z_{(k-1)}} z^{(k)} \cdot \ldots D_{z^{(0)}} z^{(1)} \cdot z^{(0)} \tag{43}$$

This proves Equation (7).