A Main Theorem Proof

To reduce notation clutter we drop layer index l and re-state the theorem:

Theorem 3.1. *Let* $G(\boldsymbol{m}, \boldsymbol{y}; \boldsymbol{\theta}_d) = \mathrm{Attn}(\boldsymbol{m}, \boldsymbol{y}, \boldsymbol{m})$*, assuming that* $\|\partial\mathcal{L}/\partial G\| = \Theta(1)$*, then* $\Delta G \triangleq G\left(\boldsymbol{m} - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{m}}, \boldsymbol{y}; \boldsymbol{\theta}_d - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_d}\right) - G(\boldsymbol{m}, \boldsymbol{y}; \boldsymbol{\theta}_d)$ *satisfies* $\|\Delta G\| = \Theta(\eta/L_d)$ *when:*

$$\|v\|^2\|w\|^2 + \|w\|^2\|m_i\|^2 + \|v\|^2\|m_i\|^2 = \Theta(1/L_d)$$

for all $i = 1, \ldots, n$*.*

Proof. Since we are only considering the magnitude of the update, it is sufficiently instructive to study the case where $d = d' = 1$. In this case the projection matrices reduce to scalars k, q, v, $w \in \mathbb{R}$, and $\boldsymbol{m}$ is a $n \times 1$ vector. Recall that for a single query y the attention block is defined as follows:

$$G(\boldsymbol{m}, y; \boldsymbol{\theta}_d) = \mathrm{softmax}\left(\frac{1}{\sqrt{d}}yqk\boldsymbol{m}^T\right)\boldsymbol{m}vw$$

Let $s_i = \frac{e^{\frac{km_iqy}{\sqrt{d}}}}{\sum_{j=1}^n e^{\frac{km_jqy}{\sqrt{d}}}}$ and $\delta_{ij} = 0$ if $i = j$ and 0 otherwise, we have:

$$\frac{\partial G}{\partial k} = \frac{1}{\sqrt{d}}vwqy\sum_{i=1}^n m_is_i\left(m_i - \sum_{j=1}^n s_jm_j\right)$$

$$\frac{\partial G}{\partial y} = \frac{1}{\sqrt{d}}vwqk\sum_{i=1}^n m_is_i\left(m_i - \sum_{j=1}^n s_jm_j\right)$$

$$\frac{\partial G}{\partial q} = \frac{1}{\sqrt{d}}vwyk\sum_{i=1}^n m_is_i\left(m_i - \sum_{j=1}^n s_jm_j\right)$$

$$\frac{\partial G}{\partial v} = w\sum_{i=1}^n s_im_i$$

$$\frac{\partial G}{\partial w} = v\sum_{i=1}^n s_im_i$$

$$\begin{aligned}\frac{\partial G}{\partial m_i} &= vws_i + vw\sum_{j=1}^n \frac{\partial s_j}{\partial m_i}x_j \\ &= vws_i + vw\sum_{j=1}^n m_js_j(\delta_{ji} - s_i)\frac{1}{\sqrt{d}}kqy \\ &= vws_i + \frac{1}{\sqrt{d}}vwkqys_i(m_i - \sum_{j=1}^n m_js_j)\end{aligned}$$

Combining these expressions we get that the total change ΔG is given by:

$\Delta G =$

$$-\eta\frac{\partial\mathcal{L}}{\partial G}\left(\frac{v^2w^2}{d}\left(\sum_{i=1}^n s_im_i\left(m_i - \sum_{j=1}^n s_jm_j\right)\right)^2(q^2y^2 + q^2k^2 + y^2k^2) + \left(\sum_{i=1}^n s_im_i\right)^2(w^2 + v^2)\right.$$

$$\left.+ v^2w^2\sum_{i=1}^n s_i^2\left(1 + \frac{1}{d}k^2q^2y^2(m_i - \sum_{j=1}^n s_jm_j)^2 + \frac{1}{\sqrt{d}}kqy(m_i - \sum_{j=1}^n s_jm_j)\right)\right)$$

By the assumption of the Theorem $\|\eta\frac{\partial\mathcal{L}}{\partial G}\| = \Theta(\eta)$, so we need to bound the term inside the main parentheses by $\Theta(\frac{1}{L})$. Note that $s_i \geq 0$ and $\sum s_i = 1$, which implies that each summation with s and m is $\Theta(m)$. The desired magnitude $\Theta(\frac{1}{L})$ is smaller than 1 so terms with lower power are leading: $v^2w^2, w^2m_i^2, v^2m_i^2$. The result follows. $\square$

B Derivation of Sufficient Conditions

In Section 3.2 we set the goal to make model update bounded in magnitude independent of model depth:

> **GOAL:** $f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$ is updated by $\Theta(\eta)$ per optimization step as $\eta \to 0$. That is, $||\Delta f|| = \Theta(\eta)$, where $\Delta f \triangleq f\left(\boldsymbol{x} - \eta\frac{\partial\mathcal{L}}{\boldsymbol{x}}, \boldsymbol{y} - \eta\frac{\partial\mathcal{L}}{\boldsymbol{y}}; \boldsymbol{\theta} - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}}\right) - f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta})$.

To achieve this, we study the forward and backward passes. Given the encoder f_e and decoder f_d, the Transformer model can be written as $f(\boldsymbol{x}, \boldsymbol{y}; \boldsymbol{\theta}) = f_d(\boldsymbol{m}, \boldsymbol{y}; \boldsymbol{\theta}_d)$ where $\boldsymbol{m} = f_e(\boldsymbol{x}; \boldsymbol{\theta}_e)$ is the memory output of the encoder. The total change after model update is then given by:

$$\Delta f = \Delta f_d \stackrel{\text{def}}{=} f_d\left(\tilde{\boldsymbol{m}}, \boldsymbol{y} - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{y}}; \boldsymbol{\theta}_d - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_d}\right) - f_d(\boldsymbol{m}, \boldsymbol{y}; \theta_d)$$

where $\tilde{\boldsymbol{m}} = f_e\left(\boldsymbol{x} - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{x}}; \boldsymbol{\theta}_e - \eta\frac{\partial\mathcal{L}}{\partial\boldsymbol{\theta}_e}\right)$ is the updated memory. Analogous to Zhang et al. (2019b), without loss of generality, we make the following assumptions to simplify derivations:

1. All relevant weights are positive with magnitude less than 1.
2. Encoder and decoder have the same number of layers N, with $L_e = 2N$ and $L_d = 3N$ blocks in the encoder and decoder respectively.
3. Embedding dimension d is 1 and the size of the input encoder sequence is n.
4. Derivative of f with respect to the loss function $\frac{\partial\mathcal{L}}{\partial f_d}$ is of order $\Theta(1)$

Forward Pass The Transformer encoder consists of L_e residual blocks $G_1, \ldots, G_{L_e}$ alternating between self-attention and MLP blocks. Let $\boldsymbol{x}_1 = \boldsymbol{x}$ and $\boldsymbol{x}_{l+1} = \boldsymbol{x}_l + G_l(\boldsymbol{x}_l, \boldsymbol{\theta}_{el})$ denote the output of the l-th block such that $\boldsymbol{m} = \boldsymbol{x}_{L_e}$. When l is odd, G_l is a self-attention block with parameters $\boldsymbol{\theta}_{el} = \{k_{el}, q_{el}, v_{el}, w_{el}\}$, and when l in even G_l is an MLP with parameters $\boldsymbol{\theta}_{el} = \{v_{el}, w_{el}\}$. We have:

$$\begin{aligned}
\boldsymbol{x}_{l+1} &\stackrel{\Theta}{=} \boldsymbol{x}_l + v_{el}w_{el}\boldsymbol{x}_l \\
\boldsymbol{x}_l &\stackrel{\Theta}{=} \boldsymbol{x}\left(1 + \sum_{i=1}^{l} v_{ei}w_{ei}\right) \\
\boldsymbol{m} &\stackrel{\Theta}{=} \boldsymbol{x}\left(1 + \sum_{l=1}^{L_e} v_{el}w_{el}\right)
\end{aligned}$$

The decoder computation is similar with the addition of encoder-attention blocks:

$$\begin{aligned}
\boldsymbol{y}_2 &= \boldsymbol{y}_1 + G_1(\boldsymbol{y}_1; \boldsymbol{\theta}_{d1}) \\
\boldsymbol{y}_3 &= \boldsymbol{y}_2 + G_2(\boldsymbol{m}, \boldsymbol{y}_2; \boldsymbol{\theta}_{d2}) \\
\boldsymbol{y}_4 &= \boldsymbol{y}_3 + G_3(\boldsymbol{y}_3; \boldsymbol{\theta}_{d3}) \\
&\vdots \\
f_d(\boldsymbol{m}, \boldsymbol{y}; \boldsymbol{\theta}_d) &= \boldsymbol{y}_{L_d} + G_{L_d}(\boldsymbol{y}_{L_d}; \boldsymbol{\theta}_{dL_d})
\end{aligned}$$

where $\boldsymbol{y}_1 = \boldsymbol{y}$. When $l\%3 \neq 0$, G_l is an attention block with parameters $\boldsymbol{\theta}_{dl} = \{k_{dl}, q_{dl}, v_{dl}, w_{dl}\}$. Otherwise, G_l is an MLP with parameters $\boldsymbol{\theta}_{dl} = \{v_{dl}, w_{dl}\}$. We have:

$$\begin{aligned}
\boldsymbol{y}_2 &\overset{\Theta}{=} \boldsymbol{y}_1 + v_{d1} w_{d1} \boldsymbol{y}_1 \\
\boldsymbol{y}_3 &\overset{\Theta}{=} \boldsymbol{y}_2 + v_{d2} w_{d2} \boldsymbol{m} \\
\boldsymbol{y}_4 &\overset{\Theta}{=} \boldsymbol{y}_3 + v_{d3} w_{d3} \boldsymbol{y}_3 \\
&\vdots \\
f(\boldsymbol{m}, \boldsymbol{y}; \boldsymbol{\theta}_d) &\overset{\Theta}{=} \boldsymbol{y}_{L_d} + v_{dL_d} w_{dL_d} \boldsymbol{x}_{L_d}
\end{aligned}$$

from which it follows that $\boldsymbol{y}_l \overset{\Theta}{=} \boldsymbol{y} \left(1 + \sum_{\substack{i=1 \\ i\%2 \neq 2}}^{l} v_{di} w_{di}\right) + \boldsymbol{m} \sum_{\substack{i=1 \\ i\%2 = 2}}^{l} v_{di} w_{di}$.

Backward Pass With $\boldsymbol{\theta}_E = \{\boldsymbol{x}, \boldsymbol{\theta}_e\}$ and $\boldsymbol{\theta}_D = \{\boldsymbol{x}, \boldsymbol{\theta}_d\}$ denoting full encoder and decoder parameters (including input embeddings), by Taylor expansion we have:

$$\begin{aligned}
\Delta f &= \frac{\partial f}{\partial \boldsymbol{\theta}_D} \Delta \boldsymbol{\theta}_D + \frac{\partial f}{\partial \boldsymbol{\theta}_E} \Delta \boldsymbol{\theta}_E + O\left(||\Delta \boldsymbol{\theta}_D||^2 + ||\Delta \boldsymbol{\theta}_E||^2\right) \\
&= \frac{\partial f}{\partial \boldsymbol{\theta}_d} \Delta \boldsymbol{\theta}_d + \frac{\partial f}{\partial \boldsymbol{\theta}_e} \Delta \boldsymbol{\theta}_e + \frac{\partial f}{\partial \boldsymbol{x}} \Delta \boldsymbol{x} + \frac{\partial f}{\partial \boldsymbol{y}} \Delta \boldsymbol{y} + O\left(\eta^2\right) \\
&= -\eta \frac{\partial f_d}{\partial \boldsymbol{\theta}_d} \frac{\partial f_d}{\partial \boldsymbol{\theta}_d}^T \frac{\partial \mathcal{L}}{\partial f_d}^T - \eta \frac{\partial f_d}{\partial f_e} \frac{\partial f_e}{\partial \boldsymbol{\theta}_e} \frac{\partial f_e}{\partial \boldsymbol{\theta}_e}^T \frac{\partial f_d}{\partial f_e}^T \frac{\partial \mathcal{L}}{\partial f_d}^T - \eta \frac{\partial f_d}{\partial \boldsymbol{y}} \frac{\partial f_d}{\partial \boldsymbol{y}}^T \frac{\partial \mathcal{L}}{\partial f_d}^T \\
&\quad - \eta \frac{\partial f_d}{\partial f_e} \frac{\partial f_e}{\partial \boldsymbol{x}} \frac{\partial f_e}{\partial \boldsymbol{x}}^T \frac{\partial f_d}{\partial f_e}^T \frac{\partial \mathcal{L}}{\partial f_d}^T + O(\eta^2)
\end{aligned} \tag{1}$$

Note that to reach our goal, it is sufficient for each of the terms to be of order $\Theta(\eta)$. We derive necessary conditions to achieve that by studying each partial derivative in Equation 1 and its contribution to Δf. By assumption 4 we have that $\frac{\partial \mathcal{L}}{\partial f_d} \overset{\Theta}{=} 1$. From the additive block-based architecture of the encoder:

$$f_e(\boldsymbol{x}; \boldsymbol{\theta}_e) = \boldsymbol{x}_1 + G_1(\boldsymbol{x}_1; \boldsymbol{\theta}_{e1}) + G_2(\boldsymbol{x}_2; \boldsymbol{\theta}_{e2}) + \ldots + G_{L_e}(\boldsymbol{x}_{L_e}; \boldsymbol{\theta}_{eL_e})$$

we have that:

$$\begin{aligned}
\frac{\partial f_e}{\partial \boldsymbol{x}} &= \frac{\partial \boldsymbol{x}_2}{\partial \boldsymbol{x}} + \frac{\partial G_2(\boldsymbol{x}_2; \boldsymbol{\theta}_{e2})}{\partial \boldsymbol{x}_2} \frac{\partial \boldsymbol{x}_2}{\partial \boldsymbol{x}} + \ldots + \frac{\partial G_{L_e}(\boldsymbol{x}_{L_e}; \boldsymbol{\theta}_{eL_e})}{\partial \boldsymbol{x}_{L_e}} \cdots \frac{\partial \boldsymbol{x}_2}{\partial \boldsymbol{x}} \\
&\overset{\Theta}{=} 1 + \frac{\partial G_1(\boldsymbol{x}; \boldsymbol{\theta}_{el})}{\partial \boldsymbol{x}}
\end{aligned}$$

so derivative magnitude is independent of the model depth. Following analogous derivation for $\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}$ we get that for each layer l:

$$\begin{aligned}
\frac{\partial f_e}{\partial \boldsymbol{\theta}_{el}} &= \frac{\partial G_l(\boldsymbol{x}_l; \boldsymbol{\theta}_{e1})}{\partial \boldsymbol{\theta}_{el}} \\
&\quad + \frac{\partial G_{l+1}(\boldsymbol{x}_{l+1}; \boldsymbol{\theta}_{e(l+1)})}{\partial \boldsymbol{x}_{l+1}} \frac{\partial G_l(\boldsymbol{x}_l; \boldsymbol{\theta}_{e1})}{\partial \boldsymbol{\theta}_{el}} \\
&\quad + \ldots \\
&\quad + \frac{\partial G_{L_e}(\boldsymbol{x}_{L_e}; \boldsymbol{\theta}_{eL_e})}{\partial \boldsymbol{x}_{L_e}} \cdots \frac{\partial G_l(\boldsymbol{x}_l; \boldsymbol{\theta}_{e1})}{\partial \boldsymbol{\theta}_{el}} \\
&\overset{\Theta}{=} \frac{\partial G_l(\boldsymbol{x}_l; \boldsymbol{\theta}_{el})}{\partial \boldsymbol{\theta}_{el}}
\end{aligned}$$

And it follows that the magnitude of $\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}$ is bound by:

$$\frac{\partial f_e}{\partial \boldsymbol{\theta}_e} \overset{\Theta}{=} \left(\frac{\partial G_1(\boldsymbol{x}_1; \boldsymbol{\theta}_{e1})}{\partial \boldsymbol{\theta}_{e1}}, \frac{\partial G_2(\boldsymbol{x}_2; \boldsymbol{\theta}_{e2})}{\partial \boldsymbol{\theta}_{e2}}, \cdots, \frac{\partial G_{L_e}(\boldsymbol{x}_{L_e}; \boldsymbol{\theta}_{eL_e})}{\partial \boldsymbol{\theta}_{eL_e}}\right)$$

with the corresponding inner product:

$$\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}^T \overset{\Theta}{=} \sum_{l=1}^{L_e} \frac{\partial G_l(\boldsymbol{x}_l;\boldsymbol{\theta}_{el})}{\partial \boldsymbol{\theta}_{el}}\frac{\partial G_l(\boldsymbol{x}_l;\boldsymbol{\theta}_{el})}{\partial \boldsymbol{\theta}_{el}}^T \tag{2}$$

Similar analysis for the decoder gives:

$$\frac{\partial f_d}{\partial \boldsymbol{\theta}_d}\frac{\partial f_d}{\partial \boldsymbol{\theta}_d}^T \overset{\Theta}{=} \sum_{\substack{l=1\\ l\%3\neq 2}}^{L_d} \frac{\partial G_l(\boldsymbol{y}_l;\boldsymbol{\theta}_{dl})}{\partial \boldsymbol{\theta}_{dl}}\frac{\partial G_l(\boldsymbol{y}_l;\boldsymbol{\theta}_{dl})}{\partial \boldsymbol{\theta}_{dl}}^T + \sum_{\substack{l=1\\ l\%3=2}}^{L_d} \frac{\partial G_l(\boldsymbol{m},\boldsymbol{y}_l;\boldsymbol{\theta}_{dl})}{\partial \boldsymbol{\theta}_{dl}}\frac{\partial G_l(\boldsymbol{m},\boldsymbol{y}_l;\boldsymbol{\theta}_{dl})}{\partial \boldsymbol{\theta}_{dl}}^T \tag{3}$$

Finally, the order of the term $\frac{\partial f_d}{\partial f_e}\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}^T\frac{\partial f_d}{\partial f_e}^T$ in Equation 1 depends on $\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}^T$ and $\frac{\partial f_d}{\partial f_e}\frac{\partial f_d}{\partial f_e}^T$. Since encoder and decoder are linked by memory, we have:

$$\frac{\partial f_d}{\partial f_e}\frac{\partial f_d}{\partial f_e}^T \overset{\Theta}{=} \sum_{\substack{l=1\\ l\%3=2}}^{L_d} \frac{\partial G_l(\boldsymbol{m},\boldsymbol{y}_l;\boldsymbol{\theta}_{dl})}{\partial \boldsymbol{m}}\frac{\partial G_l(\boldsymbol{m},\boldsymbol{y}_l;\boldsymbol{\theta}_{dl})}{\partial \boldsymbol{m}}^T \tag{4}$$

Equations 2, 3 and 4 cover all the major terms in the total change Δf, so we focus on them to derive the target bound. Expanding the terms in Equation 2 and applying Theorem 3.1 we get the following:

$$\begin{aligned}
\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}^T &\overset{\Theta}{=} \sum_{l=1}^{L_e} \frac{\partial G_l(\boldsymbol{x}_l;\boldsymbol{\theta}_{el})}{\partial \boldsymbol{\theta}_{el}}\frac{\partial G_l(\boldsymbol{x}_l;\boldsymbol{\theta}_{el})}{\partial \boldsymbol{\theta}_{el}}^T + \frac{\partial G_l(\boldsymbol{x}_l;\boldsymbol{\theta}_{el})}{\partial \boldsymbol{x}_l}\frac{\partial G_l(\boldsymbol{x}_l;\boldsymbol{\theta}_{el})}{\partial \boldsymbol{x}_i}^T \\
&\overset{\Theta}{=} \sum_{l=1}^{L_e} (v_{el}^2 + w_{el}^2)\boldsymbol{x}_l\boldsymbol{x}_l^T + v_{el}^2 w_{el}^2 \mathbf{1}_{m\times m} \\
&\overset{\Theta}{=} \sum_{l=1}^{L_e} (v_{el}^2 + w_{el}^2)\left(1 + \sum_{i=1}^{l} v_{ei}w_{ei}\right)^2 \boldsymbol{x}\boldsymbol{x}^T + v_{el}^2 w_{el}^2 \mathbf{1}_{m\times m}
\end{aligned} \tag{5}$$

Similarly, expanding Equation 3 we get:

$$\frac{\partial f_d}{\partial \boldsymbol{\theta}_d}\frac{\partial f_d}{\partial \boldsymbol{\theta}_d}^T \overset{\Theta}{=} \sum_{\substack{l=1\\ l\%3\neq 2}}^{L_d} \left((v_{dl}^2 + w_{dl}^2)\boldsymbol{y}_l\boldsymbol{y}_l^T + v_{dl}^2 w_{dl}^2 \mathbf{1}_{n\times n}\right) + \sum_{\substack{l=1\\ l\%3=2}}^{3N} \left((v_{dl}^2 + w_{dl}^2)\boldsymbol{m}^T\boldsymbol{m} + v_{dl}^2 w_{dl}^2\right)\mathbf{1}_{n\times n} \tag{6}$$

And finally for Equation 4 we have:

$$\frac{\partial f_d}{\partial f_e}\frac{\partial f_d}{\partial f_e}^T \overset{\Theta}{=} \sum_{\substack{l=1\\ l\%3=2}}^{L_d} v_{dl}^2 w_{dl}^2 \mathbf{1}_{n\times n} \tag{7}$$

To achieve the target goal it is sufficient to make Equations 5, 6 and 7 of order $\Theta(1)$. Assuming that all weights are initialized to the same order of magnitude ($v_{el} = \Theta(v_e)$, $w_{el} = \Theta(w_e)$ etc., for all l), the sufficient condition for Equation 5 can be derived as follows:

$$\begin{aligned}
1 &\overset{\Theta}{=} \sum_{l=1}^{L_e} (v_{el}^2 + w_{el}^2)\left(1 + \sum_{i=1}^{l} v_{ei}w_{ei}\right)^2 x^2 + v_{el}^2 w_{el}^2 \\
&\overset{\Theta}{=} L_e\left((v_e^2 + w_e^2)\left(1 + \sum_{i=1}^{l} v_e w_e\right)^2 x^2 + v_e^2 w_e^2\right) \\
&\overset{\Theta}{=} L_e\left(\|v_e\|^2\|x\|^2 + \|w_e\|^2\|x\|^2 + \|v_e\|^2\|w_e\|^2\right)
\end{aligned} \tag{8}$$

Similar derivation for Equation 6 gives:

$$L_d(\|v_d\|^2\|w_d\|^2 + \|v_d\|^2\|y\|^2 + \|w_d\|^2\|y\|^2 + \|v_d\|^2\|w_d\|^2 + \|v_d\|^2\|m\|^2 + \|w_d\|^2\|m\|^2) \overset{\Theta}{=} 1 \tag{9}$$

And for Equation 7 we have:

$$L_d(\|v_d\|^2\|w_d\|^2) \overset{\Theta}{=} 1 \tag{10}$$

C Encoder Initialization

Recall that $L_e = 2N$ and $L_d = 3N$, substituting these into gradient expressions for the encoder and decoder we get:

$$\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}^T \overset{\Theta}{=} 2N((v_e^2 + w_e^2)\,(1 + 2Nv_e w_e)^2\,\boldsymbol{x}\boldsymbol{x}^T + v_e^2 w_e^2 \mathbf{1}_{m\times m})$$
$$\frac{\partial f_d}{\partial \boldsymbol{\theta}_d}\frac{\partial f_d}{\partial \boldsymbol{\theta}_d}^T \overset{\Theta}{=} 2N\left((v_d^2 + w_d^2)\boldsymbol{y}\boldsymbol{y}^T + v_d^2 w_d^2 \mathbf{1}_{n\times n}\right) + N((v_d^2 + w_d^2)\boldsymbol{m}^T\boldsymbol{m} + v_d^2 w_d^2)\mathbf{1}_{n\times n}$$
$$\frac{\partial f_d}{\partial f_e}\frac{\partial f_d}{\partial f_e}^T \overset{\Theta}{=} 3Nv_d^2 w_d^2 \mathbf{1}_{n\times n}$$

Note that if $\|v_e\|\|w_e\| < \Theta(1/N)$ then $\|\boldsymbol{m}\| \overset{\Theta}{=} \|\boldsymbol{x}\|$. With this in mind, we let $\|v_d\| \overset{\Theta}{=} \|w_d\| \overset{\Theta}{=} \|\boldsymbol{y}\| \overset{\Theta}{=} \|\boldsymbol{x}\| \overset{\Theta}{=} (9N)^{-\frac{1}{4}}$, which by design gives:

$$\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}^T \overset{\Theta}{=} 2N((v_e^2 + w_e^2)(9N)^{-\frac{1}{4}} + v_e^2 w_e^2)\mathbf{1}_{m\times m}$$
$$\frac{\partial f_d}{\partial \boldsymbol{\theta}_d}\frac{\partial f_d}{\partial \boldsymbol{\theta}_d}^T \overset{\Theta}{=} 2N\left((3(9N)^{-1}\right) + N(3(9N)^{-1})\mathbf{1}_{n\times n} \overset{\Theta}{=} \mathbf{1}_{n\times n}$$
$$\frac{\partial f_d}{\partial f_e}\frac{\partial f_d}{\partial f_e}^T \overset{\Theta}{=} 3N(9N)^{-1} \overset{\Theta}{=} \mathbf{1}_{n\times n}$$

We then solve for the magnitude of v_e and w_e that achieves $\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}\frac{\partial f_e}{\partial \boldsymbol{\theta}_e}^T \overset{\Theta}{=} \mathbf{1}_{n\times n}$. Assuming that $\|v_e\| = \|w_e\|$ due to symmetry, we obtain $\|v_e\| = \|w_e\| = \left(\frac{\sqrt{22}-2}{6}\right)^{\frac{1}{2}} N^{-\frac{1}{4}} \approx 0.67N^{-\frac{1}{4}}$.

D Training Hyper-Parameters

Parameters	**IWSLT'14**$_{small}$ **De-En**	**WMT'18**$_{base}$ **Fi-En**	**WMT'17**$_{base}$ **En-De**	**WMT'17**$_{deep}$ **En-De**	**WMT'17**$_{big}$ **En-De**
Starting learning rate	0.0005	0.0006	0.0007	0.0004	0.0004
Decay steps	4000	4000	4000	4000	4000
Dropout	0.5	0.4	0.2	0.4	0.4
Batch size (tokens)	4k	80k	25k	25k	25k
Max updates	300k	90k	1M	500k	500k
Mixed precision	No	No	No	Yes	Yes

Table 1: Hyper-parameters for T-Fixup models on each dataset.