A Proofs

Example A.1. An *affine layer* $\mathrm{aff} : \mathbb{R}^{d_1 \times d_0 + d_1} \times \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_1 \times N}$ is given by the formula

$$\mathrm{aff}(A, b, X) := AX + b1_N^T. \tag{16}$$

A routine calculation shows that

$$J\mathrm{aff}(A, b, X) = A \otimes \mathrm{Id}_N, \tag{17}$$

while

$$D\mathrm{aff}(A, b, X) = \big(\mathrm{Id}_{d_1} \otimes X^T, \mathrm{Id}_{d_1} \otimes 1_N\big). \tag{18}$$

More generally, if $P : \mathbb{R}^p \to \mathbb{R}^{d_1 \times d_0}$ denotes any continuously differentiable map, then one obtains a *P-parameterised affine layer*

$$\mathrm{aff}_P(w, b, X) := \mathrm{aff}(P(w), b, X) = P(w)X + b1^T. \tag{19}$$

One has

$$J\mathrm{aff}_P(w, b, X) = P(w) \otimes \mathrm{Id}_N \tag{20}$$

and, by the chain rule,

$$D\mathrm{aff}_P(w, b, X) = \big((\mathrm{Id}_{d_1} \otimes X^T)DP(w), \mathrm{Id}_{d_1} \otimes 1\big), \tag{21}$$

where $DP(w) \in \mathbb{R}^{d_0 N \times p}$ is the derivative of P at w. Common examples include *ϵ-weight normalisation* $\mathrm{wn}(w) := (\epsilon + \|w\|_{row}^2)^{-\frac{1}{2}} w$ [32] and convolutions, which send convolutional kernels to associated Toeplitz matrices. We will also consider *ϵ-entry normalisation* $\mathrm{en}(w) := (\epsilon + w^2)^{-\frac{1}{2}} w$, with operations applied entrywise.

Example A.2. A *(parameter-free) elementwise nonlinearity* $\Phi : \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_0 \times N}$ defined by a continuously differentiable function $\phi : \mathbb{R} \to \mathbb{R}$ is given by applying ϕ to every component of a matrix $X \in \mathbb{R}^{d_0 \times N}$. Extension to the parameterised case is straightforward.

Example A.3. A *(parameter-free) batch normalisation (BN) layer* $\mathrm{bn} : \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_0 \times N}$ is given by the formula

$$\mathrm{bn}(X) := \frac{X - \mathbb{E}[X]}{\sqrt{\epsilon + \sigma[X]^2}}, \tag{22}$$

where $\epsilon > 0$ is some fixed hyperparameter and $\mathbb{E}$ and σ denote the row-wise mean and standard deviation. The parameterised BN layer from [17], with scaling and bias parameters γ and β respectively, is given simply by postcomposition $\mathrm{aff}_{\mathrm{diag}}(\gamma, \beta, \cdot) \circ \mathrm{bn}$ with a diag-parameterised affine layer (Example A.1).

Example A.4. A *residual block* $f : \mathbb{R}^p \times \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_1 \times N}$ can be defined given any other layer (or composite thereof) $g : \mathbb{R}^p \times \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_1 \times N}$ by the formula

$$f(\theta, X) := IX + g(\theta, X), \tag{23}$$

where $I : \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_1 \times N}$ is some linear transformation. In practice, I is frequently the identity map [16]; our main theorem will concern the case where I has all singular values equal to 1.

Proof of Theorem 2.4. Using the fact that $D\ell = (D\gamma \circ F) \cdot DF$, we compute:

$$\begin{aligned}
\|\nabla \ell(\theta)\|^2 &= \langle DF(\theta)^T \nabla \gamma(F(\theta)), DF(\theta)^T \nabla \gamma(F(\theta))\rangle \\
&= \langle \nabla \gamma(F(\theta)), DF(\theta) DF(\theta)^T \nabla \gamma(F(\theta))\rangle \\
&\geq \lambda(DF(\theta)) \|\nabla \gamma(F(\theta))\|^2 \\
&\geq \mu \lambda(DF(\theta)) \bigg(\gamma(F(\theta)) - \inf_{\theta'} \gamma(F(\theta'))\bigg),
\end{aligned}$$

where the first inequality follows from the standard estimate $\langle v, AA^T v\rangle \geq \lambda_{min}(AA^T)\|v\|^2$, and the final inequality follows from the fact that γ is μ-PŁ over the set $\{F(\theta) : \theta \in \mathbb{R}^p\}$. □

Our proof of Proposition 4.3 requires the following standard lemma.

Lemma A.5. *Let $\{g_i : \mathbb{R}^p \to \mathbb{R}^{m_i \times m_{i-1}}\}_{i=1}^n$ be a family of matrix-valued functions. If, with respect to some submultiplicative matrix norm, each g_i is bounded by b_i and Lipschitz with constant c_i on a set $S \subset \mathbb{R}^p$, then their pointwise matrix product $\theta \mapsto \prod_{i=1}^n g_i(\theta)$ is also bounded and Lipschitz on S, with bound $\prod_{i=1}^n b_i$ and Lipschitz constant $\sum_{i=1}^n c_i\big(\prod_{j\neq i} b_j\big)$.* □

Proof of Lemma A.5. We prove the lemma by induction. When $n = 2$, adding and subtracting a copy of $g_1(\theta)g_2(\theta')$ and using the triangle inequality implies that $\|g_1g_2(\theta) - g_1g_2(\theta')\|$ is bounded by

$$\|g_1(\theta)(g_2(\theta) - g_2(\theta'))\| + \|(g_1(\theta) - g_1(\theta'))g_2(\theta')\|.$$

Applying submultiplicativity of the matrix norm and the bounds provided by the b_i and c_i gives

$$\|g_1g_2(\theta) - g_1g_2(\theta')\| \leq (b_1c_2 + b_2c_1)\|\theta - \theta'\|.$$

Now suppose we have the result for $n = k$. Writing $\prod_{i=1}^{k+1} g_i$ as $g_1 \prod_{i=2}^{k+1} g_i$ and applying the above argument, the induction hypothesis tells us that $\prod_{i=1}^{k+1} g_i$ is indeed bounded by $\prod_{i=1}^{k+1} b_i$ and Lipschitz with Lipschitz constant $\sum_{i=1}^{k+1} c_i\big(\prod_{j\neq i} b_j\big)$. The result follows. □

Proof of Proposition 4.3. By Proposition 4.2, it suffices to show that for each $1 \leq l \leq L$, the function

$$\vec{\theta} \mapsto \prod_{j=l+1}^{L} Jf_j\big(\theta_j, f_{<j}(\vec{\theta}, X)\big) Df_l\big(\theta_l, f_{<l}(\vec{\theta}, X)\big) \tag{24}$$

is bounded and Lipschitz on S. To show this, we must first prove that each map $\vec{\theta} \mapsto f_{<j}(\vec{\theta}, X)$ is bounded and Lipschitz on S. This we prove by induction.

By hypothesis, $\vec{\theta} \mapsto f_1(\vec{\theta}, X) = f_1(\theta_1, X)$ is bounded and Lipschitz on S. Suppose now that for $j > 1$, one has $\vec{\theta} \mapsto f_{<j}(\vec{\theta}, X)$ bounded and Lipschitz on S. Then the range of $S \ni \theta \mapsto f_{<j}(\vec{\theta}, X)$ is a bounded subset of $\mathbb{R}^{d_j \times N}$. By hypothesis on f_j, it then follows that $\theta \mapsto f_{<j+1}(\vec{\theta}, X) = f_j\big(\theta_j, f_{<j}(\vec{\theta}, X)\big)$ is bounded and Lipschitz on S.

The hypothesis on the Jf_j and Df_j now implies that the maps $\vec{\theta} \mapsto Jf_j\big(\theta_j, f_{<j}(\vec{\theta}, X)\big)$, $l + 1 \leq j \leq L$, and $\vec{\theta} \mapsto Df_l\big(\theta_l, f_{<l}(\vec{\theta}, X)\big)$ are all bounded and Lipschitz on S. In particular, as a product of bounded and Lipschitz functions, the map given in Equation (24) is also bounded and Lipschitz on S. Therefore DF is bounded and Lipschitz on S. □

Proof of Corollary 4.4. By hypothesis, $F(S)$ is a bounded subset of $\mathbb{R}^{d_L \times N}$. Continuity of γ then implies that $\gamma(F(S))$ is a bounded subset of $\mathbb{R}$, so that $F(S)$ is contained in a sublevel set of γ. The result now follows from the hypotheses. □

To prove Theorem 4.5 it will be convenient to recall some tensor calculus. If $f : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{m_1 \times m_2}$ is a matrix-valued, differentiable function of a matrix-valued variable, its derivative Df can be regarded as a map $\mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{m_1 \times m_2 \times n_1 \times n_2}$ whose components are given by

$$Df^{i_1,i_2}_{j_1,j_2}(X) = \frac{\partial f^{i_1}_{i_2}}{\partial x^{j_1}_{j_2}}(X), \qquad X \in \mathbb{R}^{n_1 \times n_2}$$

where $1 \leq i_\alpha \leq m_\alpha$ and $1 \leq j_\alpha \leq n_\alpha$ are the indices, $\alpha = 1, 2$. It is easily deduced from the chain rule of ordinary calculus that if $f : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{m_1 \times m_2}$ and $g : \mathbb{R}^{m_1 \times m_2} \to \mathbb{R}^{l_1 \times l_2}$ are differentiable, then $g \circ f$ is differentiable with derivative $(Dg \circ f) \cdot Df : \mathbb{R}^{n_1 \times n_2} \to \mathbb{R}^{l_1 \times l_2 \times n_1 \times n_2}$, where here $\cdot$ denotes contraction over the $m_1 \times m_2$ indices. The following lemmata then follow from routine calculation.

Lemma A.6. *Let bn $: \mathbb{R}^{d \times N} \to \mathbb{R}^{d \times N}$ be an ϵ-batchnorm layer. Then one can write bn $= v \circ m$, where $v, m : \mathbb{R}^{d \times N} \to \mathbb{R}^{d \times N}$ are given respectively by*

$$v(Y) = (N\epsilon + \|Y\|^2_{row})^{-\frac{1}{2}} \sqrt{N} Y, \tag{25}$$

$$m(X) = X - \frac{1}{N} X 1_{N \times N}. \tag{26}$$

One has

$$\frac{\partial v^i_j}{\partial y^k_l} = \delta^i_k \sqrt{N}(N\epsilon + \|y^i\|^2)^{-\frac{1}{2}}\left(\delta^j_l - (N\epsilon + \|y^i\|^2)^{-1} y^i_l y^i_j\right) \tag{27}$$

and

$$\begin{aligned}\frac{\partial^2 v^i_j}{\partial y^m_n \partial y^k_l} =& \delta^i_k \delta^i_m \sqrt{N}(N\epsilon + \|y^i\|^2)^{-\frac{3}{2}} \times \\ & \times \big(3(N\epsilon + \|y^i\|^2)^{-1} y^i_n y^i_l y^i_j \\ & - (\delta^j_l y^i_n + \delta^l_n y^i_j + \delta^j_n y^i_l)\big),\end{aligned} \tag{28}$$

with

$$\frac{\partial m^i_j}{\partial x^k_l} = \delta^i_k(\delta^j_l - N^{-1}). \tag{29}$$

and all second derivatives of m being zero. □

Lemma A.7. *Let wn : $\mathbb{R}^{d_1 \times d_0} \to \mathbb{R}^{d_1 \times d_0}$ be an ϵ-weight normalised parameterisation (Example A.1). Then one has*

$$\frac{\partial wn^i_j}{\partial w^k_l} = \delta^i_k(\epsilon + \|w^i\|^2)^{-\frac{1}{2}}\left(\delta^j_l - (\epsilon + \|w^i\|^2)^{-1} w^i_l w^i_j\right) \tag{30}$$

and

$$\begin{aligned}\frac{\partial wn^i_j}{\partial w^m_n \partial w^k_k} =& \delta^i_k \delta^i_m (\epsilon + \|w^i\|^2)^{-\frac{3}{2}} \times \\ & \times \big(3(\epsilon + \|w^i\|^2)^{-1} w^i_n w^i_l w^i_j \\ & - (\delta^j_l w^i_n + \delta^l_n w^i_j + \delta^j_n w^i_l)\big).\end{aligned} \tag{31}$$

Similarly, if en : $\mathbb{R}^{d_1 \times d_0} \to \mathbb{R}^{d_1 \times d_0}$ is an ϵ-entry-normalised parameterisation, then

$$\frac{\partial en^i_j}{\partial w^k_l} = \delta^i_k \delta^j_l \epsilon(\epsilon + (w^i_j)^2)^{-\frac{3}{2}} \tag{32}$$

and

$$\frac{\partial en^i_j}{\partial w^n_m \partial w^k_l} = -\delta^i_n \delta^j_m \delta^i_k \delta^j_l 3\epsilon(\epsilon + (w^i_j)^2)^{-\frac{3}{2}} w^i_j \tag{33}$$

Proof of Theorem 4.5. (1) follows from continuity of the nonlinearity and its derivative, implying boundedness of both over bounded sets in $\mathbb{R}^{d\times N}$.

(2) and (3) follow from a similar argument to the following argument for batch norm, which we give following Lemma A.6. Specifically, for the composite $f := \text{bn} \circ \text{aff} : \mathbb{R}^{d_1 \times d_0} \times \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_1 \times N}$ defined by an ϵ-BN layer and an affine layer, we will prove that over any set $B \subset \mathbb{R}^{d_0 \times N}$ consisting of matrices X whose covariance matrix is nondegenerate, one has f, Df and Jf all globally bounded and Lipschitz. Indeed, v (Equation (25)) is clearly globally bounded, while Dv (Equation (27)) is globally bounded, decaying like $\|Y\|^{-1}_{row}$ out to infinity, and $D^2 v$ (Equation (28)) is globally bounded, decaying like $\|Y\|^{-2}_{row}$ out to infinity. Consequently,

$$\begin{aligned}\text{bn} \circ \text{aff} &= v \circ (m \circ \text{aff}), \\ D(\text{bn} \circ \text{aff}) &= (Jv \circ m \circ \text{aff}) \cdot (Jm \circ \text{aff}) \cdot D\text{aff}, \\ J(\text{bn} \circ \text{aff}) &= (Jv \circ m \circ \text{aff}) \cdot (Jm \circ \text{aff}) \cdot J\text{aff},\end{aligned}$$

and similarly the derivatives of $D(\text{bn} \circ \text{aff})$ and $J(\text{bn} \circ \text{aff})$ are all globally bounded over $\mathbb{R}^{d_1 \times d_0} \times B$. The hypothesis that B consist of matrices with nondegenerate covariance matrix is needed here because while $Jv \circ m \circ \text{aff}$ decays like $\|A(X - \mathbb{E}[X])\|^{-1}_{row}$ out to infinity, the row-norm $\|(A(X - \mathbb{E}[X]))^i)\|^2 = (A^i(X - \mathbb{E}[X])(X - \mathbb{E}[X])^T(A^i)^T) = \|A^i\|^2_{\text{Cov}(X)}$ can only be guaranteed to increase with A if $\text{Cov}(X)$ is nondegenerate. Thus, for instance, without the nondegeneracy hypothesis on $\text{Cov}(X)$, $A \mapsto J(\text{bn} \circ \text{aff})(A, X)$ grows unbounded like $J\text{aff}(A, X) = A \otimes \text{Id}_N$ in any direction of

degeneracy of $\mathrm{Cov}(X)$. Nonetheless, with the nondegenerate covariance assumption on elements of B, bn $\circ$ aff satisfies the hypotheses of Proposition 4.3 over $\mathbb{R}^{d_1 \times d_0} \times B$.

(2) and (3) now follow from essentially the same boundedness arguments as for batch norm, using Lemma A.7 in the place of Lemma A.6. However, since the row norms in this case are always defined by the usual Euclidean inner product on row-vectors, as opposed to the possibly degenerate inner product coming from the covariance matrix of the input vectors, one does not require any hypotheses aside from boundedness on the set B. Thus entry- and weight-normalised affine layers satisfy the hypotheses of Proposition 4.3.

Finally, (5) follows from the above arguments. More specifically, if $g : \mathbb{R}^p \times \mathbb{R}^{d \times N} \to \mathbb{R}^{d \times N}$ is any composite of layers of the above form, then g satisfies the hypotheses of Proposition 4.3. Consequently, so too does the residual block $f(\theta, X) := X + g(\theta, X)$, for which $Jf(\theta, X) = \mathrm{Id}_d \otimes \mathrm{Id}_N + Jg(\theta, X)$ and $Df(\theta, X) = Dg(\theta, X)$.

□

Proof of Proposition 4.6. In the notation of Proposition 4.2, the product $DF(\vec{\theta})DF(\vec{\theta})^T$ is the sum of the positive-semidefinite matrices $D_{\theta_l}F(\vec{\theta})D_{\theta_l}F(\vec{\theta})^T$. Therefore $\lambda(DF(\vec{\theta})) \geq \sum_l \lambda(D_{\theta_l}F(\vec{\theta}))$. The result now follows from the inequality $\lambda(AB) \geq \lambda(A)\lambda(B)$ applied inductively using Equation (11). Note that $\lambda(AB) \geq \lambda(A)\lambda(B)$ is either trivial if one or both of A and B have more rows than columns (in which case the right hand side is zero), and follows from the well-known inequality $\sigma(AB) \geq \sigma(A)\sigma(B)$ for the smallest singular values if both A and B have at least as many columns as rows. □

Theorem 4.7 follows from the following two lemmata.

Lemma A.8. *Let $g : \mathbb{R}^p \times \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_1 \times N}$ be a layer for which there exists $\delta > 0$ such that $\|Jg(\theta, X)\|_2 < (1 - \delta)$ for all θ and X. Let $I : \mathbb{R}^{d_0 \times N} \to \mathbb{R}^{d_1 \times N}$ be a linear map whose singular values are all equal to 1. Then the residual block $f(\theta, X) := IX + g(\theta, X)$ has $\sigma(Jf(\theta, X)) > \delta$ for all θ and X.*

Proof. Observe that

$$Jf(\theta, X) = I \otimes \mathrm{Id}_N + Jg(\theta, X). \tag{34}$$

The result then follows from Weyl's inequality: all singular values of $I \otimes \mathrm{Id}_N$ are equal to 1, so that

$$\sigma\big(Jf(\theta, X)\big) \geq 1 - \|Jg(\theta, X)\|_2 > \delta$$

for all θ and X. □

Lemma A.9. *Let $P : \mathbb{R}^p \to \mathbb{R}^{d_1 \times d_0}$ be a parameterisation. Then*

$$\sigma\big(D\mathit{aff}_P(w, X)\big) \geq \sigma(X)\sigma\big(DP(w)\big) \tag{35}$$

for all $w \in \mathbb{R}^p$ and $X \in \mathbb{R}^{d_0 \times N}$.

Proof. Follows from Equation (21) and the inequality $\sigma(AB) \geq \sigma(A)\sigma(B)$. □

Proof of Theorem 4.7. Hypothesis 1 in Theorem 4.7 says that the residual branches of the f_l, $l \geq 2$, satisfy the hypotheses of Lemma A.8, so that $\sigma(Jf_l(\theta_l, f_{<l}(\vec{\theta}, X))) > 0$ for all $l \geq 2$. By the assumption that $d_{l-1} \geq d_l$, this means that $\lambda(Jf_l(\theta_l, f_{<l}(\vec{\theta}, X))) = \sigma(Jf_l(\theta_l, f_{<l}(\vec{\theta}, X)))^2 > 0$. On the other hand, hypothesis 2 together with Lemma A.9 implies that $\lambda(Df_1(\theta_1, X)) \geq \sigma(Df_1(\theta_1, X))^2 > 0$. The result now follows from Proposition 4.6. □

.

Proof of Theorem 5.1. By Theorem 4.5, all layers satisfy the Hypotheses of Proposition 4.3 and so by Corollary 4.4, the associated loss function is globally Lipschitz, with Lipschitz constant some $\beta > 0$. Take $\eta > 0$ to be any number smaller than $2\beta^{-1}$; thus the loss can be guaranteed to be decreasing with every gradient descent step.

We now show that the network satisfies the hypotheses of Theorem 4.7. The dimension constraints in item (1) are encoded directly into the definition of the network, while the operator-norm of each of the residual branches, as products of $P(w)$ and $D\Phi$ matrices, are globally bounded by 1 by our hypotheses on these factors. For item (2), our data matrix is full-rank since it consists of linearly independent data, while by definition we have $p_1 = d_1 d_0$ with $Df_1 = D\text{aff}_{\text{en}}$ being everywhere full-rank since ϵ-entry-normalisation is a diffeomorphism onto its image for any $\epsilon > 0$. Its hypotheses being satisfied by our weight-normalised residual network, Theorem 4.7 implies the parameter-function map F associated to $\{f_l\}_{l=1}^L$ and X satisfies $\lambda(DF(\vec{\theta})) = \sigma(DF(\vec{\theta}))^2 > 0$ for all parameters $\vec{\theta}$. There are now two cases to consider.

In the first case, the gradient descent trajectory never leaves some ball of finite radius in $\mathbb{R}^{d_1 \times d_0}$, the parameter space for the first layer. In any such ball, recalling that the first layer's parameterisation is entry-normalisation (Example A.1), the smallest singular value

$$\sigma(D\,\text{en}(w)) = \min_{1 \leq i \leq d_1 1 \leq j \leq d_0} \frac{\epsilon}{(\epsilon + (w_j^i)^2)^{\frac{3}{2}}} \tag{36}$$

of $D\,\text{en}(w)$ is uniformly lower bounded by some positive constant. Thus by Lemmas A.9 and A.8[7], the smallest singular value of DF is also uniformly lower bounded by a positive constant in any such ball. It follows from Theorem 2.4 that the loss satisfies the PŁ-inequality over such a ball, so that gradient descent converges in this case at a linear rate to a global minimum.

The second and only other case that must be considered is when for each $R > 0$, there is some time T for which the weight norm $\|w_t\|$ of the parameters in the first layer is greater than R for all $t \geq T$. That is, the parameter trajectory in the first layer is unbounded in time. In this case, inspection of Equation (36) reveals that the smallest singular value of DF *cannot* be uniformly bounded below by a positive constant over all of parameter space. Theorem 2.4 then says that there is merely a sequence $(\mu_t)_{t \in \mathbb{N}}$, with μ_t proportional to $\sigma(D\,\text{en}(w_t))$, for which

$$\ell_t - \ell^* \leq \prod_{i=0}^{t} (1 - \mu_i \alpha)(\ell_0 - \ell^*), \tag{37}$$

where $\alpha = \eta(1 - 2\beta^{-1}\eta) > 0$. To guarantee convergence in this case, therefore, it suffices to show that $\prod_{t=0}^{\infty}(1 - \mu_t \alpha) = 0$; equivalently, it suffices to show that the infinite series

$$\sum_{t=0}^{\infty} \log(1 - \mu_t \alpha) \tag{38}$$

diverges.

The terms of the series (38) form a sequence of negative numbers which converges to zero. Hence, for the series (38) to diverge, it is *necessary* that μ_t decrease *sufficiently slowly* with time. By the integral test, therefore, it suffices to find an integrable function $m : [t_0, \infty) \to \mathbb{R}_{\geq 0}$ such that $\mu_t \geq m(t)$ for each integer $t \geq t_0$, for which the integral $\int_{t_0}^{\infty} \log(1 - m(t)\alpha)\, dt$ diverges.

We construct m by considering the worst possible case: where each gradient descent step is in exactly the same direction going out to ∞, thereby decreasing $\sigma(D\,\text{en}(w))$ at the fastest possible rate. By applying an orthogonal-affine transformation to $\mathbb{R}^{d_1 d_0}$, we can assume without loss of generality that the algorithm is initialised at, and consistently steps in the direction of, the first canonical basis vector e_1 in $\mathbb{R}^{d_1 d_0}$. Specifically, letting $\vec{\theta}$ be the vector of parameters for all layers following the first and $w \in \mathbb{R}^{d_1 d_0}$ the first layer parameters, for $r \in \mathbb{R}_{\geq 1}$ we may assume that

$$\nabla_w \ell(\vec{\theta}, re_1) = \partial_{w_1^1} \ell(\vec{\theta}, re_1) e_1, \tag{39}$$

with $\partial_{w_1^1} \ell(\vec{\theta}, re_1) \geq 0$ for all $(\vec{\theta}, r)$. Let γ denote the convex function defined by the cost c (cf. Equation (6)), and let $A(\vec{\theta}, re_1)$ denote the $d_1 d_0$-dimensional row vector

$$D\gamma\big(F(\vec{\theta}, re_1)\big) \prod_{l=2}^{L} Jf_l\big(\theta_l, f_{<l}\big((\vec{\theta}, re_1), X\big)\big)(\text{Id}_{d_1} \otimes X^T). \tag{40}$$

[7] See the supplementary material.

Then, in this worst possible case, the single nonzero partial derivative defining the loss gradient with respect to w at the point $(\vec{\theta}, re_1)$ is given by

$$\partial_{w_1^1} \ell(\vec{\theta}, re_1)) = A(\vec{\theta}, re_1)_1 \, \frac{\epsilon}{(\epsilon + r^2)^{\frac{3}{2}}}, \tag{41}$$

where $A(\vec{\theta}, re_1)_1$ denotes the first component of the row vector $A(\vec{\theta}, re_1)$ (cf. Equation (21)). By Theorem 4.5, however, the magnitude of $A(\vec{\theta}, re_1)$ can be globally upper bounded by some constant C. Thus

$$\partial_{w_1^1} \ell(\vec{\theta}, re_1) \leq \frac{C\epsilon}{(\epsilon + r^2)^{\frac{3}{2}}} \leq \frac{C\epsilon}{r^3} \tag{42}$$

for all $\vec{\theta}$ and $r \geq 1$.

Let us therefore consider the Euler method, with step size η, applied over $\mathbb{R}_{\geq 1}$, starting from $r_0 = 1$, with respect to the vector field $V(r) = C\epsilon r^{-3}$. Labelling the iterates $(r_t)_{t \in \mathbb{N}}$, we claim that there exist constants γ_1, γ_2 and γ_3 such that $0 < r_t \leq \gamma_1 + \gamma_2 (t + \gamma_3)^{\frac{1}{4}}$ for all $t \in \mathbb{N}$. Indeed, observe that the solution to the flow equation $\dot{r}(t) = C\epsilon r(t)^{-3}$ is $r(t) = (4C\epsilon t + 1)^{\frac{1}{4}}$, so the claim follows if we can show that there exists a constant B such that $|r_t - r(t)| < B$ for all integer $t \geq 0$. However this follows from Theorem 10.6 of [14].

Now, since for each t, r_t is an upper bound for the magnitude of the parameter vector $w_t \in \mathbb{R}^{d_1 d_0}$, we see from Equation (36) that the smallest singular value $\sigma(D\text{en}(w_t))$ admits the lower bound

$$\sigma(D\,\text{en}(w_t)) \geq \frac{\epsilon}{(\epsilon + (\gamma_1 + \gamma_2 (t + \gamma_3)^{\frac{1}{4}})^2)^{\frac{3}{2}}} \tag{43}$$

for all $t \in \mathbb{N}$. Clearly, $(\epsilon + (\gamma_1 + \gamma_2 (t + \gamma_3)^{\frac{1}{4}})^2)^{\frac{3}{2}} = O(t^{\frac{3}{4}})$. Hence there exist $t_0 > 0$ and $\Gamma > 0$ such that

$$\mu_t \geq \frac{\Gamma}{t^{\frac{3}{4}}} \tag{44}$$

for all integer $t \geq t_0$. Then, setting $m(t) := \Gamma t^{-\frac{3}{4}}$, the integral

$$\int_{t_0}^{\infty} \log(1 - m(t)\alpha)\, dt = \int_{t_0}^{\infty} \log\left(\frac{t^{\frac{3}{4}} - \Gamma\alpha}{t^{\frac{3}{4}}}\right) dt \tag{45}$$

diverges. It follows that gradient descent converges as $t \to \infty$ to a global minimum. □

A.1 Experimental details

For all our experiments, the data was standardised channel-wise using the channel-wise mean and standard deviation over the training set.

On ImageNet, the models were trained using the default PyTorch ImageNet example[8], using SGD with weight decay of $1e-4$ and momentum of 0.9, batch size of 256, and random crop/horizontal flip data augmentation. The test accuracies obtained were 74.22 ± 0.14 for the standard network, and 74.82 ± 0.04 for the modified network.

On CIFAR10/100, our models[9] were trained using SGD with a batch size of 128 and random crop/horizontal flip data augmentation. We ran 10 trials over each of the learning rates 0.2, 0.1, 0.05 and 0.02. The only exception to this is for CIFAR10 with a learning rate of 0.2, where training diverged 6 our of 10 times on the original network, so we plotted only those 4 trials where training did not diverge. Mean and standard deviation test accuracies, as well as averaged-over-final epoch loss values, are given in Tables 1 and 2.

The plots at the optimal learning rate, 0.1 for CIFAR10 and 0.05 for CIFAR100, are in Figure 3, while we provide the plots for the other learning rates in Figures 4 and 5.

While Hypothesis 6.1 is clearly validated at all learning rates on CIFAR100, it holds to lessening extents on CIFAR10 as learning rate is decreased. Ultimately, the modified version ends up achieving a slightly *higher* loss at the end of training with the smallest learning rate, suggesting that Hypothesis 6.1 should not be invoked too far from initialisation.

[8] https://github.com/pytorch/examples/tree/main/imagenet

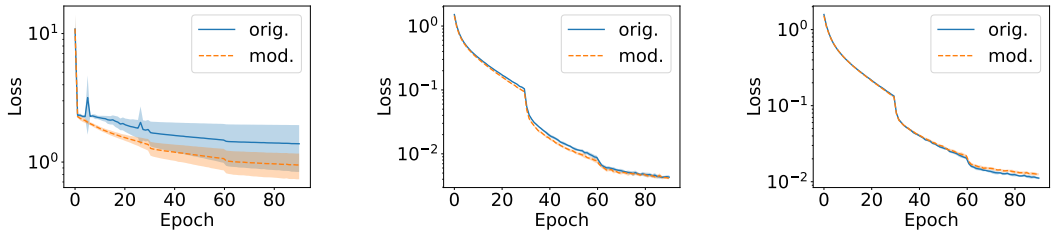
[9] minimally modifying https://github.com/kuangliu/pytorch-cifar

Table 1: ResNet18 on CIFAR10

	Original		Modified	
Learning rate	Test accuracy	Final loss	Test accuracy	Final loss
0.2	48.14 ± 19.95	1.3883 ± 0.5503	62.71 ± 6.14	0.9495 ± 0.2150
0.1	91.64 ± 0.24	0.0032 ± 0.0002	91.42 ± 0.61	0.0022 ± 0.0002
0.05	91.22 ± 0.25	0.0043 ± 0.0003	91.14 ± 0.25	0.0041 ± 0.0001
0.02	89.50 ± 0.32	0.0112 ± 0.0004	88.94 ± 0.35	0.0125 ± 0.0004

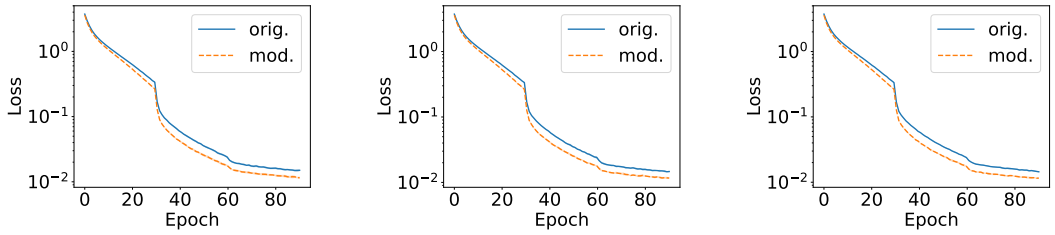
Table 2: ResNet18 on CIFAR100

	Original		Modified	
Learning rate	Test accuracy	Final loss	Test accuracy	Final loss
0.2	69.67 ± 0.58	0.0150 ± 0.0005	70.07 ± 0.53	0.0115 ± 0.0005
0.1	70.05 ± 0.43	0.0147 ± 0.0004	70.87 ± 0.42	0.0116 ± 0.0003
0.05	70.04 ± 0.48	0.0145 ± 0.0003	70.34 ± 0.56	0.0116 ± 0.0003
0.02	69.75 ± 0.56	0.0145 ± 0.0003	70.13 ± 0.55	0.0116 ± 0.0004



(a) $lr = 0.2$ (b) $lr = 0.05$ (c) $lr = 0.02$

Figure 4: Loss plots for ResNet18 on CIFAR10



(a) $lr = 0.2$ (b) $lr = 0.05$ (c) $lr = 0.02$

Figure 5: Loss plots for ResNet18 on CIFAR100