# A. Deferred Proofs

## A.1. Proof of Theorem 1

Before delving into the proof, we introduce notation that will admit a more compact presentation of formulae. For $1 \leq a \leq b \leq N$, we denote:

$$\prod_{a}^{j=b} W_j \ := W_b W_{b-1} \cdots W_a$$
$$\prod_{j=a}^{b} W_j^\top := W_a^\top W_{a+1}^\top \cdots W_b^\top$$

where $W_1 \ldots W_N$ are the weight matrices of the depth-$N$ linear network (Equation 2). If $a > b$, then by definition both $\prod_a^{j=b} W_j$ and $\prod_{j=a}^b W_j^\top$ are identity matrices, with size depending on context, *i.e.* on the dimensions of matrices they are multiplied against. Given any square matrices (possibly scalars) $A_1, A_2, \ldots, A_m$, we denote by $diag(A_1 \ldots A_m)$ a block-diagonal matrix holding them on its diagonal:

$$diag(A_1 \ldots A_m) = \begin{bmatrix} A_1 & 0 & 0 & 0 \\ 0 & \ddots & 0 & 0 \\ 0 & 0 & A_m & 0 \\ 0 & 0 & 0 & 0 \end{bmatrix}$$

As illustrated above, $diag(A_1 \ldots A_m)$ may hold additional, zero-valued rows and columns beyond $A_1 \ldots A_m$. Conversely, it may also trim (omit) rows and columns, from its bottom and right ends respectively, so long as only zeros are being removed. The exact shape of $diag(A_1 \ldots A_m)$ is again determined by context, and so if $B$ and $C$ are matrices, the expression $B \cdot diag(A_1 \ldots A_m) \cdot C$ infers a number of rows equal to the number of columns in $B$, and a number of columns equal to the number of rows in $C$.

Turning to the actual proof, we disregard the trivial case $N = 1$, and begin by noticing that Equation 3, along with the definition of $W_e$ (Equation 5), imply that for every $j = 1 \ldots N$:

$$\frac{\partial L^N}{\partial W_j}(W_1, \ldots, W_N) = \prod_{i=j+1}^{N} W_i^\top \cdot \frac{dL^1}{dW}(W_e) \cdot \prod_{i=1}^{j-1} W_i^\top$$

Plugging this into the differential equations of gradient descent (Equation 6), we get:

$$\dot{W}_j(t) = -\eta\lambda W_j(t) - \eta \prod_{i=j+1}^{N} W_i^\top(t) \cdot \frac{dL^1}{dW}(W_e(t)) \cdot \prod_{i=1}^{j-1} W_i^\top(t) \quad , \ j = 1 \ldots N \tag{16}$$

For $j = 1 \ldots N-1$, multiply the $j$'th equation by $W_j^\top(t)$ from the right, and the $j+1$'th equation by $W_{j+1}^\top(t)$ from the left. This yields:

$$W_{j+1}^\top(t)\dot{W}_{j+1}(t) + \eta\lambda \cdot W_{j+1}^\top(t) W_{j+1}(t) = \dot{W}_j(t) W_j^\top(t) + \eta\lambda \cdot W_j(t) W_j^\top(t)$$

Taking the transpose of these equations and adding to themselves, we obtain, for every $j = 1 \ldots N-1$:

$$\begin{aligned} W_{j+1}^\top(t)\dot{W}_{j+1}(t) + \dot{W}_{j+1}^\top(t) W_{j+1}(t) + \\ 2\eta\lambda \cdot W_{j+1}^\top(t) W_{j+1}(t) = \\ \dot{W}_j(t) W_j^\top(t) + W_j(t)\dot{W}_j^\top(t) + \\ 2\eta\lambda \cdot W_j(t) W_j^\top(t) \end{aligned} \tag{17}$$

Denote for $j = 1 \ldots N$:

$$C_j(t) := W_j(t) W_j^\top(t) \quad , \quad C'_j(t) := W_j^\top(t) W_j(t)$$

Equation 17 can now be written as:

$$\dot{C}'_{j+1}(t) + 2\eta\lambda \cdot C'_{j+1}(t) = \dot{C}_j(t) + 2\eta\lambda \cdot C_j(t) \quad , \ j = 1 \ldots N-1$$

Turning to Lemma 1 below, while recalling our assumption for time $t_0$ (Equation 7):

$$C'_{j+1}(t_0) = C_j(t_0) \quad , \ j = 1 \ldots N-1$$

we conclude that, throughout the entire time-line:

$$C'_{j+1}(t) = C_j(t) \quad , \ j = 1 \ldots N-1$$

Recollecting the definitions of $C_j(t), C'_j(t)$, this means:

$$W_{j+1}^\top(t) W_{j+1}(t) = W_j(t) W_j^\top(t) \ , \ j = 1 \ldots N-1 \tag{18}$$

Regard $t$ now as fixed, and for every $j = 1 \ldots N$, let:

$$W_j(t) = U_j \Sigma_j V_j^\top \tag{19}$$

be a singular value decomposition. That is to say, $U_j$ and $V_j$ are orthogonal matrices, and $\Sigma_j$ is a rectangular-diagonal matrix holding non-decreasing, non-negative singular values on its diagonal. Equation 18 implies that for $j = 1 \ldots N-1$:

$$V_{j+1} \Sigma_{j+1}^\top \Sigma_{j+1} V_{j+1}^\top = U_j \Sigma_j \Sigma_j^\top U_j^\top$$

For a given $j$, the two sides of the above equation are both orthogonal eigenvalue decompositions of the same matrix. The square-diagonal matrices $\Sigma_{j+1}^\top \Sigma_{j+1}$ and $\Sigma_j \Sigma_j^\top$ are thus the same, up to a possible permutation of diagonal elements (eigenvalues). However, since by definition $\Sigma_{j+1}$ and $\Sigma_j$ have non-increasing diagonals, it must hold that $\Sigma_{j+1}^\top \Sigma_{j+1} = \Sigma_j \Sigma_j^\top$. Let $\rho_1 > \rho_2 > \cdots > \rho_m \geq 0$ be the distinct eigenvalues, with corresponding multiplicities $d_1, d_2, \ldots, d_m \in \mathbb{N}$. We may write:

$$\Sigma_{j+1}^\top \Sigma_{j+1} = \Sigma_j \Sigma_j^\top = diag(\rho_1 I_{d_1}, \ldots, \rho_m I_{d_m}) \tag{20}$$

where $I_{d_r}$, $1{\leq}r{\leq}m$, is the identity matrix of size $d_r \times d_r$. Moreover, there exist orthogonal matrices $O_{j,r} \in \mathbb{R}^{d_r,d_r}$, $1{\leq}r{\leq}m$, such that:

$$U_j = V_{j+1} \cdot diag(O_{j,1}, \ldots, O_{j,m})$$

$O_{j,r}$ here is simply a matrix changing between orthogonal bases in the eigenspace of $\rho_r$ – it maps the basis comprising $V_{j+1}$-columns to that comprising $U_j$-columns. Recalling that both $\Sigma_j$ and $\Sigma_{j+1}$ are rectangular-diagonal, holding only non-negative values, Equation 20 implies that each of these matrices is equal to $diag(\sqrt{\rho_1}\cdot I_{d_1}, \ldots, \sqrt{\rho_m}\cdot I_{d_m})$. Note that the matrices generally do not have the same shape and thus, formally, are not equal to one another. Nonetheless, in line with our $diag$ notation (see beginning of this subsection), $\Sigma_j$ and $\Sigma_{j+1}$ may differ from each other only in trailing, zero-valued rows and columns. By an inductive argument, all the singular value matrices $\Sigma_1, \Sigma_2, \ldots, \Sigma_N$ (see Equation 19) are equal up to trailing zero rows and columns. The fact that $\rho_1 \ldots \rho_m$ do not include an index $j$ in their notation is thus in order, and we may write, for every $j = 1 \ldots N{-}1$:

$$\begin{aligned} W_j(t) &= U_j \Sigma_j V_j^\top \\ &= V_{j+1} \cdot diag(O_{j,1}, \ldots, O_{j,m}) \cdot \\ &\qquad diag(\sqrt{\rho_1}\cdot I_{d_1}, \ldots, \sqrt{\rho_m}\cdot I_{d_m}) \cdot V_j^\top \end{aligned}$$

For the $N$'th weight matrix we have:

$$\begin{aligned} W_N(t) &= U_N \Sigma_N V_N^\top \\ &= U_N \cdot diag(\sqrt{\rho_1}\cdot I_{d_1}, \ldots, \sqrt{\rho_m}\cdot I_{d_m}) \cdot V_N^\top \end{aligned}$$

Concatenations of weight matrices thus simplify as follows:

$$\prod_j^{i=N} W_i(t) \prod_{i=j}^N W_i^\top(t) = \tag{21}$$
$$U_N \cdot diag\Big((\rho_1)^{N-j+1}\cdot I_{d_1}, \ldots, (\rho_m)^{N-j+1}\cdot I_{d_m}\Big) \cdot U_N^\top$$

$$\prod_{i=1}^j W_i^\top(t) \prod_1^{i=j} W_i(t) = \tag{22}$$
$$V_1 \cdot diag\Big((\rho_1)^{j}\cdot I_{d_1}, \ldots, (\rho_m)^{j}\cdot I_{d_m}\Big) \cdot V_1^\top$$
$$, \; j = 1 \ldots N$$

where we used the orthogonality of $O_{j,r}$, and the obvious fact that it commutes with $I_{d_r}$. Consider Equation 21 with $j = 1$ and Equation 22 with $j = N$, while recalling that by definition $W_e(t) = \prod_1^{i=N} W_j(t)$:

$$W_e(t)W_e^\top(t) = U_N \cdot diag\Big((\rho_1)^N I_{d_1}, \ldots, (\rho_m)^N I_{d_m}\Big) \cdot U_N^\top$$
$$W_e^\top(t)W_e(t) = V_1 \cdot diag\Big((\rho_1)^N I_{d_1}, \ldots, (\rho_m)^N I_{d_m}\Big) \cdot V_1^\top$$

It follows that for every $j = 1 \ldots N$:

$$\prod_j^{i=N} W_i(t) \prod_{i=j}^N W_i^\top(t) = \left[W_e(t)W_e^\top(t)\right]^{\frac{N-j+1}{N}} \tag{23}$$

$$\prod_{i=1}^j W_i^\top(t) \prod_1^{i=j} W_i(t) = \left[W_e^\top(t)W_e(t)\right]^{\frac{j}{N}} \tag{24}$$

where $[\cdot]^{\frac{N-j+1}{N}}$ and $[\cdot]^{\frac{j}{N}}$ stand for fractional power operators defined over positive semidefinite matrices.

With Equations 23 and 24 in place, we are finally in a position to complete the proof. Returning to Equation 16, we multiply $\dot{W}_j(t)$ from the left by $\prod_{j+1}^{i=N} W_i(t)$ and from the right by $\prod_1^{i=j-1} W_i(t)$, followed by summation over $j = 1 \ldots N$. This gives:

$$\begin{aligned} &\sum_{j=1}^N \left(\prod_{j+1}^{i=N} W_i(t)\right) \dot{W}_j(t) \left(\prod_1^{i=j-1} W_i(t)\right) = \\ &-\eta\lambda \sum_{j=1}^N \left(\prod_{j+1}^{i=N} W_i(t)\right) W_j(t) \left(\prod_1^{i=j-1} W_i(t)\right) \\ &-\eta \sum_{j=1}^N \left(\prod_{j+1}^{i=N} W_i(t) \prod_{i=j+1}^N W_i^\top(t)\right) \cdot \\ &\qquad \frac{dL^1}{dW}(W_e(t)) \cdot \left(\prod_{i=1}^{j-1} W_i^\top(t) \prod_1^{i=j-1} W_i(t)\right) \end{aligned}$$

By definition $W_e(t) = \prod_1^{i=N} W_j(t)$, so we can substitute the first two lines above:

$$\begin{aligned} &\dot{W}_e(t) = -\eta\lambda N \cdot W_e(t) \\ &-\eta \sum_{j=1}^N \left(\prod_{j+1}^{i=N} W_i(t) \prod_{i=j+1}^N W_i^\top(t)\right) \cdot \\ &\qquad \frac{dL^1}{dW}(W_e(t)) \cdot \left(\prod_{i=1}^{j-1} W_i^\top(t) \prod_1^{i=j-1} W_i(t)\right) \end{aligned}$$

Finally, plugging in the relations in Equations 23 and 24, the sought-after result is revealed:

$$\begin{aligned} \dot{W}_e(t) &= -\eta\lambda N \cdot W_e(t) \\ &\quad -\eta \sum_{j=1}^N \left[W_e(t)W_e^\top(t)\right]^{\frac{N-j}{N}} \cdot \\ &\qquad \frac{dL^1}{dW}(W_e(t)) \cdot \left[W_e^\top(t)W_e(t)\right]^{\frac{j-1}{N}} \end{aligned}$$

□

**Lemma 1.** *Let $I \subset \mathbb{R}$ be a connected interval, and let $f, g : I \to \mathbb{R}$ be differentiable functions. Suppose that there exists a constant $\alpha \geq 0$ for which:*

$$\dot{f}(t) + \alpha \cdot f(t) = \dot{g}(t) + \alpha \cdot g(t) \quad , \forall t \in I$$

*Then, if $f$ and $g$ assume the same value at some $t_0 \in I$ (interior or boundary), they must coincide along the entire interval,* i.e. *it must hold that $f(t) = g(t)$ for all $t \in I$.*

*Proof.* Define $h := f - g$. $h$ is a differentiable function from $I$ to $\mathbb{R}$, and we have:

$$\dot{h}(t) = -\alpha \cdot h(t) \quad , \forall t \in I \tag{25}$$

We know that $h(t_0) = 0$ for some $t_0 \in I$, and would like to show that $h(t) = 0 \; \forall t \in I$. Assume by contradiction that this is not the case, so there exists $t_2 \in I$ for which $h(t_2) \neq 0$. Without loss of generality, suppose that $h(t_2) > 0$, and that $t_2 > t_0$. Let $S$ be the zero set of $h$, *i.e.* $S := \{t \in I : h(t) = 0\}$. Since $h$ is continuous in $I$, $S$ is topologically closed, therefore its intersection with the interval $[t_0, t_2]$ is compact. Denote by $t_1$ the maximal element in this intersection, and consider the interval $J := [t_1, t_2] \subset I$. By construction, $h$ is positive along $J$, besides on the endpoint $t_1$ where it assumes the value of zero. For $t_1 < t \leq t_2$, we may solve as follows the differential equation of $h$ (Equation 25):

$$\frac{\dot{h}(t)}{h(t)} = -\alpha \quad \implies \quad h(t) = \beta e^{-\alpha t}$$

where $\beta$ is the positive constant defined by $h(t_2) = \beta e^{-\alpha t_2}$. Since in particular $h$ is bounded away from zero on $(t_1, t_2]$, and assumes zero at $t_1$, we obtain a contradiction to its continuity. This completes the proof. □

### A.2. Proof of Claim 1

Our proof relies on the *Kronecker product* operation for matrices. For arbitrary matrices $A$ and $B$ of sizes $m_a \times n_a$ and $m_b \times n_b$ respectively, the Kronecker product $A \odot B$ is defined to be the following block matrix:

$$A \odot B := \begin{bmatrix} a_{11} \cdot B & \cdots & a_{1n_a} \cdot B \\ \vdots & \ddots & \vdots \\ a_{m_a 1} \cdot B & \cdots & a_{m_a n_a} \cdot B \end{bmatrix} \in \mathbb{R}^{m_a m_b, n_a n_b} \tag{26}$$

where $a_{ij}$ stands for the element in row $i$ and column $j$ of $A$. The Kronecker product admits numerous useful properties. We will employ the following:

- If $A$ and $B$ are matrices such that the matrix product $AB$ is defined, then:

$$\begin{aligned} vec(AB) &= (B^\top \odot I_{r_A}) \cdot vec(A) \\ &= (I_{c_B} \odot A) \cdot vec(B) \end{aligned} \tag{27}$$

  where $I_{r_A}$ and $I_{c_B}$ are the identity matrices whose sizes correspond, respectively, to the number of rows in $A$ and the number of columns in $B$. $vec(\cdot)$ here, as in claim statement, stands for matrix vectorization in column-first order.

- If $A_1$, $A_2$, $B_1$ and $B_2$ are matrices such that the matrix products $A_1 B_1$ and $A_2 B_2$ are defined, then:

$$(A_1 \odot A_2)(B_1 \odot B_2) = (A_1 B_1) \odot (A_2 B_2) \tag{28}$$

- For any matrices $A$ and $B$:

$$(A \odot B)^\top = A^\top \odot B^\top \tag{29}$$

- Equation 28 and 29 imply, that if $A$ and $B$ are some orthogonal matrices, so is $A \odot B$:

$$\begin{aligned} A^\top = A^{-1} \;\; \wedge \;\; B^\top = B^{-1} \\ \implies (A \odot B)^\top = (A \odot B)^{-1} \end{aligned} \tag{30}$$

With the Kronecker product in place, we proceed to the actual proof. It suffices to show that vectorizing:

$$\sum_{j=1}^{N} \left[ W_e^{(t)} (W_e^{(t)})^\top \right]^{\frac{j-1}{N}} \cdot \frac{dL^1}{dW}(W_e^{(t)}) \cdot \left[ (W_e^{(t)})^\top W_e^{(t)} \right]^{\frac{N-j}{N}}$$

yields:

$$P_{W_e^{(t)}} \cdot vec\left( \frac{dL^1}{dW}(W_e^{(t)}) \right)$$

where $P_{W_e^{(t)}}$ is the preconditioning matrix defined in claim statement. For notational conciseness, we hereinafter omit the iteration index $t$, and simply write $W_e$ instead of $W_e^{(t)}$.

Let $I_d$ and $I_k$ be the identity matrices of sizes $d \times d$ and $k \times k$ respectively. Utilizing the properties of the Kronecker product, we have:

$$\begin{aligned} &vec\left( \sum_{j=1}^{N} \left[ W_e W_e^\top \right]^{\frac{j-1}{N}} \frac{dL^1}{dW}(W_e) \left[ W_e^\top W_e \right]^{\frac{N-j}{N}} \right) \\ &= \sum_{j=1}^{N} \left( I_d \odot \left[ W_e W_e^\top \right]^{\frac{j-1}{N}} \right) \cdot \\ &\qquad \left( \left[ W_e^\top W_e \right]^{\frac{N-j}{N}} \odot I_k \right) \cdot vec\left( \frac{dL^1}{dW}(W_e) \right) \\ &= \sum_{j=1}^{N} \left( \left[ W_e^\top W_e \right]^{\frac{N-j}{N}} \odot \left[ W_e W_e^\top \right]^{\frac{j-1}{N}} \right) vec\left( \frac{dL^1}{dW}(W_e) \right) \end{aligned}$$

The first equality here makes use of Equation 27, and the second of Equation 28. We will show that the matrix:

$$Q := \sum_{j=1}^{N} \left[ W_e^\top W_e \right]^{\frac{N-j}{N}} \odot \left[ W_e W_e^\top \right]^{\frac{j-1}{N}} \tag{31}$$

meets the characterization of $P_{W_e}$, thereby completing the proof. Let:

$$W_e = UDV^\top$$

be a singular value decomposition, *i.e.* $U \in \mathbb{R}^{k,k}$ and $V \in \mathbb{R}^{d,d}$ are orthogonal matrices, and $D$ is a rectangular-diagonal matrix holding (non-negative) singular values on its diagonal. Plug this into the definition of $Q$ (Equation 31):

$$\begin{aligned}
Q &= \sum_{j=1}^{N} \left[VD^\top DV^\top\right]^{\frac{N-j}{N}} \odot \left[UDD^\top U^\top\right]^{\frac{j-1}{N}} \\
&= \sum_{j=1}^{N} \left(V \left[D^\top D\right]^{\frac{N-j}{N}} V^\top\right) \odot \left(U \left[DD^\top\right]^{\frac{j-1}{N}} U^\top\right) \\
&= \sum_{j=1}^{N} (V \odot U) \left(\left[D^\top D\right]^{\frac{N-j}{N}} \odot \left[DD^\top\right]^{\frac{j-1}{N}}\right) (V^\top \odot U^\top) \\
&= (V \odot U) \left(\sum_{j=1}^{N} \left[D^\top D\right]^{\frac{N-j}{N}} \odot \left[DD^\top\right]^{\frac{j-1}{N}}\right) (V \odot U)^\top
\end{aligned}$$

The third equality here is based on the relation in Equation 28, and the last equality is based on Equation 29. Denoting:

$$O := V \odot U \tag{32}$$

$$\Lambda := \sum_{j=1}^{N} \left[D^\top D\right]^{\frac{N-j}{N}} \odot \left[DD^\top\right]^{\frac{j-1}{N}} \tag{33}$$

we have:

$$Q = O\Lambda O^\top \tag{34}$$

Now, since by definition $U$ and $V$ are orthogonal, $O$ is orthogonal as well (follows from the relation in Equation 30). Additionally, the fact that $D$ is rectangular-diagonal implies that the square matrix $\Lambda$ is also diagonal. Equation 34 is thus an orthogonal eigenvalue decomposition of $Q$. Finally, denote the columns of $U$ (left singular vectors of $W_e$) by $\mathbf{u}_1 \ldots \mathbf{u}_k$, those of $V$ (right singular vectors of $W_e$) by $\mathbf{v}_1 \ldots \mathbf{v}_d$, and the diagonal elements of $D$ (singular values of $W_e$) by $\sigma_1 \ldots \sigma_{\max\{k,d\}}$ (by definition $\sigma_r = 0$ if $r > \min\{k,d\}$). The definitions in Equations 32 and 33 imply that the columns of $O$ are:

$$vec(\mathbf{u}_r \mathbf{v}_{r'}^\top) \quad , r = 1 \ldots k \ , \ r' = 1 \ldots d$$

with corresponding diagonal elements in $\Lambda$ being:

$$\sum_{j=1}^{N} \sigma_r^{2\frac{N-j}{N}} \sigma_{r'}^{2\frac{j-1}{N}} \quad , r = 1 \ldots k \ , \ r' = 1 \ldots d$$

We conclude that $Q$ indeed meets the characterization of $P_{W_e}$ in claim statement. This completes the proof.

□

### A.3. Proof of Claim 2

We disregard the trivial case $N = 1$, as well as the scenario $W_e^{(t)} = 0$ (both lead Equations 10 and 12 to equate). Omitting the iteration index $t$ from our notation, it suffices to show that:

$$\sum_{j=1}^{N} \left[W_e W_e^\top\right]^{\frac{j-1}{N}} \cdot \frac{dL^1}{dW}(W_e) \cdot \left[W_e^\top W_e\right]^{\frac{N-j}{N}} = \tag{35}$$
$$\|W_e\|_2^{2-\frac{2}{N}} \left(\tfrac{dL^1}{dW}(W_e) + (N-1) Pr_{W_e}\left\{\tfrac{dL^1}{dW}(W_e)\right\}\right)$$

where $Pr_{W_e}\{\cdot\}$ is the projection operator defined in claim statement (Equation 13), and we recall that by assumption $k = 1$ ($W_e \in \mathbb{R}^{1,d}$). $\left[W_e W_e^\top\right]^{\frac{j-1}{N}}$ is a scalar, equal to $\|W_e\|_2^{2\frac{j-1}{N}}$ for every $j = 1 \ldots N$. $\left[W_e^\top W_e\right]^{\frac{N-j}{N}}$ on the other hand is a $d \times d$ matrix, by definition equal to identity for $j = N$, and otherwise, for $j = 1 \ldots N-1$, it is equal to $\|W_e\|_2^{2\frac{N-j}{N}} \left(W_e/\|W_e\|_2\right)^\top \left(W_e/\|W_e\|_2\right)$. Plugging these equalities into the first line of Equation 35 gives:

$$\sum_{j=1}^{N} \left[W_e W_e^\top\right]^{\frac{j-1}{N}} \frac{dL^1}{dW}(W_e) \left[W_e^\top W_e\right]^{\frac{N-j}{N}} =$$

$$\sum_{j=1}^{N-1} \|W_e\|_2^{2\frac{j-1}{N}} \frac{dL^1}{dW}(W_e) \|W_e\|_2^{2\frac{N-j}{N}} \left(\frac{W_e}{\|W_e\|_2}\right)^\top \left(\frac{W_e}{\|W_e\|_2}\right)$$
$$+ \|W_e\|_2^{2\frac{N-1}{N}} \cdot \frac{dL^1}{dW}(W_e) =$$
$$(N-1) \|W_e\|_2^{2\frac{N-1}{N}} \frac{dL^1}{dW}(W_e) \left(\frac{W_e}{\|W_e\|_2}\right)^\top \left(\frac{W_e}{\|W_e\|_2}\right)$$
$$+ \|W_e\|_2^{2\frac{N-1}{N}} \cdot \frac{dL^1}{dW}(W_e)$$

The latter expression is precisely the second line of Equation 35, thus our proof is complete.

□

### A.4. Proof of Theorem 2

Our proof relies on elementary differential geometry: curves, arc length and line integrals (see Chapters 8 and 9 in Buck (2003)).

Let $\mathcal{U} \subset \mathbb{R}^{1,d}$ be a neighborhood of $W = 0$ (*i.e.* an open set that includes this point) on which $\frac{dL^1}{dW}$ is continuous ($\mathcal{U}$ exists by assumption). It is not difficult to see that $F(\cdot)$ (Equation 14) is continuous on $\mathcal{U}$ as well. The strategy of our proof will be to show that $F(\cdot)$ does not admit the *gradient theorem* (also known as the *fundamental theorem for line integrals*). According to the theorem, if $h : \mathcal{U} \to \mathbb{R}$ is a continuously differentiable function, and $\Gamma$ is a piecewise smooth curve lying in $\mathcal{U}$ with start-point $\gamma_s$ and end-point $\gamma_e$,

then:

$$\int_{\Gamma} \frac{dh}{dW} = h(\gamma_e) - h(\gamma_s)$$

In words, the line integral of the gradient of $h$ over $\Gamma$, is equal to the difference between the value taken by $h$ at the end-point of $\Gamma$, and that taken at the start-point. A direct implication of the theorem is that if $\Gamma$ is closed ($\gamma_e = \gamma_s$), the line integral vanishes:

$$\oint_{\Gamma} \frac{dh}{dW} = 0$$

We conclude that if $F(\cdot)$ is the gradient field of some function, its line integral over any closed (piecewise smooth) curve lying in $\mathcal{U}$ must vanish. We will show that this is not the case.

For notational conciseness we hereinafter identify $\mathbb{R}^{1,d}$ and $\mathbb{R}^d$, so in particular $\mathcal{U}$ is now a subset of $\mathbb{R}^d$. To further simplify, we omit the subindex from the Euclidean norm, writing $\|\cdot\|$ instead of $\|\cdot\|_2$. Given an arbitrary continuous vector field $\phi : \mathcal{U} \to \mathbb{R}^d$, we define a respective (continuous) vector field as follows:

$$F_\phi : \mathcal{U} \to \mathbb{R}^d$$

$$F_\phi(\mathbf{w}) = \tag{36}$$
$$\begin{cases} \|\mathbf{w}\|^{2-\frac{2}{N}} \left( \phi(\mathbf{w}) + (N-1) \left\langle \phi(\mathbf{w}), \frac{\mathbf{w}}{\|\mathbf{w}\|} \right\rangle \frac{\mathbf{w}}{\|\mathbf{w}\|} \right) & , \mathbf{w} \neq \mathbf{0} \\ \mathbf{0} & , \mathbf{w} = \mathbf{0} \end{cases}$$

Notice that for $\phi = \frac{dL^1}{dW}$, we get exactly the vector field $F(\cdot)$ defined in theorem statement (Equation 14) – the subject of our inquiry. As an operator on (continuous) vector fields, the mapping $\phi \mapsto F_\phi$ is linear.[5] This, along with the linearity of line integrals, imply that for any piecewise smooth curve $\Gamma$ lying in $\mathcal{U}$, the functional $\phi \mapsto \int_\Gamma F_\phi$, a mapping of (continuous) vector fields to scalars, is linear. Lemma 2 below provides an upper bound on this linear functional in terms of the length of $\Gamma$, its maximal distance from origin, and the maximal norm $\phi$ takes on it.

In light of the above, to show that $F(\cdot)$ contradicts the gradient theorem, thereby completing the proof, it suffices to find a closed (piecewise smooth) curve $\Gamma$ for which the linear functional $\phi \mapsto \oint_\Gamma F_\phi$ does not vanish at $\phi = \frac{dL^1}{dW}$. By assumption $\frac{dL^1}{dW}(W{=}0) \neq 0$, and so we may define the unit vector in the direction of $\frac{dL^1}{dW}(W{=}0)$:

$$\mathbf{e} := \frac{\frac{dL^1}{dW}(W{=}0)}{\left\| \frac{dL^1}{dW}(W{=}0) \right\|} \in \mathbb{R}^d \tag{37}$$

[5] For any $\phi_1, \phi_2 : \mathcal{U} \to \mathbb{R}^d$ and $c \in \mathbb{R}$, it holds that $F_{\phi_1+\phi_2} = F_{\phi_1} + F_{\phi_2}$ and $F_{c\cdot\phi_1} = c \cdot F_{\phi_1}$.

Let $R$ be a positive constant small enough such that the Euclidean ball of radius $R$ around the origin is contained in $\mathcal{U}$. Let $r$ be a positive constant smaller than $R$. Define $\Gamma_{r,R}$ to be a curve as follows (see illustration in Figure 1):[6]

$$\Gamma_{r,R} := \Gamma^1_{r,R} \;\to\; \Gamma^2_{r,R} \;\to\; \Gamma^3_{r,R} \;\to\; \Gamma^4_{r,R} \tag{38}$$

where:

- $\Gamma^1_{r,R}$ is the line segment from $-R \cdot \mathbf{e}$ to $-r \cdot \mathbf{e}$.
- $\Gamma^2_{r,R}$ is a geodesic on the sphere of radius $r$, starting from $-r \cdot \mathbf{e}$ and ending at $r \cdot \mathbf{e}$.
- $\Gamma^3_{r,R}$ is the line segment from $r \cdot \mathbf{e}$ to $R \cdot \mathbf{e}$.
- $\Gamma^4_{r,R}$ is a geodesic on the sphere of radius $R$, starting from $R \cdot \mathbf{e}$ and ending at $-R \cdot \mathbf{e}$.

$\Gamma_{r,R}$ is a piecewise smooth, closed curve that fully lies within $\mathcal{U}$. Consider the linear functional it induces: $\phi \mapsto \oint_{\Gamma_{r,R}} F_\phi$. We will evaluate this functional on $\phi = \frac{dL^1}{dW}$. To do so, we decompose the latter as follows:

$$\tfrac{dL^1}{dW}(\cdot) = c \cdot \mathbf{e}(\cdot) + \xi(\cdot) \tag{39}$$

where:

- $c$ is a scalar equal to $\|\frac{dL^1}{dW}(W{=}0)\|$.
- $\mathbf{e}(\cdot)$ is a vector field returning the constant $\mathbf{e}$ (Equation 37).
- $\xi(\cdot)$ is a vector field returning the values of $\frac{dL^1}{dW}(\cdot)$ shifted by the constant $-\frac{dL^1}{dW}(W{=}0)$. It is continuous on $\mathcal{U}$ and vanishes at the origin.

Applying Lemma 2 to $\xi$ over $\Gamma_{r,R}$ gives:

$$\begin{aligned} \left| \oint_{\Gamma_{r,R}} F_\xi \right| &\leq N \cdot len(\Gamma_{r,R}) \cdot \max_{\gamma \in \Gamma_{r,R}} \|\gamma\|^{2-\frac{2}{N}} \cdot \max_{\gamma \in \Gamma_{r,R}} \|\xi(\gamma)\| \\ &= N \cdot (\pi r + \pi R + 2(R-r)) \cdot R^{2-\frac{2}{N}} \cdot \max_{\gamma \in \Gamma_{r,R}} \|\xi(\gamma)\| \\ &\leq N \cdot 2\pi \cdot R^{3-\frac{2}{N}} \cdot \max_{\gamma \in \Gamma_{r,R}} \|\xi(\gamma)\| \end{aligned}$$

On the other hand, by Lemma 3:

$$\oint_{\Gamma_{r,R}} F_{\mathbf{e}} = \left( \frac{2N}{3 - 2/N} - 2 \right) \left( R^{3-\frac{2}{N}} - r^{3-\frac{2}{N}} \right)$$

[6] The proof would have been slightly simplified had we used a curve that passes directly through the origin. We avoid this in order to emphasize that the result is not driven by some point-wise singularity (the origin received special treatment in the definition of $F(\cdot)$ – see Equations 14 and 13).

The linearity of the functional $\phi \mapsto \oint_{\Gamma_{r,R}} F_\phi$, along with Equation 39, then imply:

$$\begin{aligned}\oint_{\Gamma_{r,R}} F_{\frac{dL^1}{dW}} &= c \cdot \oint_{\Gamma_{r,R}} F_{\mathbf{e}} + \oint_{\Gamma_{r,R}} F_\xi \\ &\geq c \cdot \left(\frac{2N}{3-2/N} - 2\right)\left(R^{3-\frac{2}{N}} - r^{3-\frac{2}{N}}\right) \\ &\quad - N \cdot 2\pi \cdot R^{3-\frac{2}{N}} \cdot \max_{\gamma \in \Gamma_{r,R}} \|\xi(\gamma)\|\end{aligned}$$

We will show that for proper choices of $R$ and $r$, the lower bound above is positive. $\Gamma_{r,R}$ will then be a piecewise smooth closed curve lying in $\mathcal{U}$, for which the functional $\phi \mapsto \oint_{\Gamma_{r,R}} F_\phi$ does not vanish at $\phi = \frac{dL^1}{dW}$. As stated, this will imply that $F(\cdot)$ violates the gradient theorem, thereby concluding our proof.

All that is left is to affirm that the expression:

$$\begin{aligned}&c \cdot \left(\tfrac{2N}{3-2/N} - 2\right)\left(R^{3-\frac{2}{N}} - r^{3-\frac{2}{N}}\right) \\ &-N \cdot 2\pi \cdot R^{3-\frac{2}{N}} \cdot \max\nolimits_{\gamma \in \Gamma_{r,R}} \|\xi(\gamma)\|\end{aligned}$$

can indeed be made positive with proper choices of $R$ and $r$. Recall that:

- $N > 2$ by assumption; implies $\frac{2N}{3-2/N} - 2 > 0$.
- $R$ is any positive constant small enough such that the ball of radius $R$ around the origin is contained in $\mathcal{U}$.
- $r$ is any positive constant smaller than $R$.
- $\Gamma_{r,R}$ is a curve whose points are all within distance $R$ from the origin.
- $c = \|\frac{dL^1}{dW}(W{=}0)\|$ – positive by assumption.
- $\xi(\cdot)$ is a vector field that is continuous on $\mathcal{U}$ and vanishes at the origin.

The following procedure gives $R$ and $r$ as required:

- Set $r$ to follow $R$ such that: $r^{3-\frac{2}{N}} = 0.5 \cdot R^{3-\frac{2}{N}}$.
- Choose $\epsilon > 0$ for which $0.5c\left(\frac{2N}{3-\frac{2}{N}} - 2\right) - 2\pi N\epsilon > 0$.
- Set $R$ to be small enough such that $\|\xi(\mathbf{w})\| \leq \epsilon$ for any point $\mathbf{w}$ within distance $R$ from the origin.

The proof is complete.

□

**Lemma 2.** *Let $\phi : \mathcal{U} \to \mathbb{R}^d$ be a continuous vector field, and let $\Gamma$ be a piecewise smooth curve lying in $\mathcal{U}$. Consider the (continuous) vector field $F_\phi : \mathcal{U} \to \mathbb{R}^d$ defined in Equation 36. The line integral of the latter over $\Gamma$ is bounded as follows:*

$$\left|\int_\Gamma F_\phi\right| \leq N \cdot len(\Gamma) \cdot \max_{\gamma \in \Gamma} \|\gamma\|^{2-\frac{2}{N}} \cdot \max_{\gamma \in \Gamma} \|\phi(\gamma)\|$$

*where $len(\Gamma)$ is the arc length of $\Gamma$, and $\gamma \in \Gamma$ refers to a point lying on the curve.*

*Proof.* We begin by noting that the use of $\max$ (as opposed to $\sup$) in stated upper bound is appropriate, since under our definition of a curve (adopted from Buck (2003)), points lying on it constitute a compact set. This subtlety is of little importance – one may as well replace $\max$ by $\sup$, and the lemma would still serve its purpose.

It is not difficult to see that for any $\mathbf{w} \in \mathcal{U}$, $\mathbf{w} \neq 0$:

$$\begin{aligned}&\|F_\phi(\mathbf{w})\| = \|\mathbf{w}\|^{2-\frac{2}{N}} \left\|\phi(\mathbf{w}) + (N-1)\left\langle \phi(\mathbf{w}), \frac{\mathbf{w}}{\|\mathbf{w}\|}\right\rangle \frac{\mathbf{w}}{\|\mathbf{w}\|}\right\| \\ &\leq \|\mathbf{w}\|^{2-\frac{2}{N}} \left(\|\phi(\mathbf{w})\| + (N-1)\left|\left\langle \phi(\mathbf{w}), \frac{\mathbf{w}}{\|\mathbf{w}\|}\right\rangle\right| \cdot \left\|\frac{\mathbf{w}}{\|\mathbf{w}\|}\right\|\right) \\ &= \|\mathbf{w}\|^{2-\frac{2}{N}} \left(\|\phi(\mathbf{w})\| + (N-1)\left|\left\langle \phi(\mathbf{w}), \frac{\mathbf{w}}{\|\mathbf{w}\|}\right\rangle\right|\right) \\ &\leq \|\mathbf{w}\|^{2-\frac{2}{N}} (\|\phi(\mathbf{w})\| + (N-1)\|\phi(\mathbf{w})\|) \\ &\leq N \|\mathbf{w}\|^{2-\frac{2}{N}} \|\phi(\mathbf{w})\|\end{aligned}$$

Trivially, $\|F_\phi(\mathbf{w})\| \leq N \|\mathbf{w}\|^{2-\frac{2}{N}} \|\phi(\mathbf{w})\|$ holds for $\mathbf{w}{=}0$ as well. The sought-after result now follows from the properties of line integrals:

$$\begin{aligned}\left|\int_\Gamma F_\phi\right| &\leq \int_\Gamma \|F_\phi\| \leq \int_\Gamma N \|\mathbf{w}\|^{2-\frac{2}{N}} \|\phi(\mathbf{w})\| \\ &\leq N \cdot len(\Gamma) \cdot \max_{\gamma \in \Gamma} \|\gamma\|^{2-\frac{2}{N}} \cdot \max_{\gamma \in \Gamma} \|\phi(\gamma)\|\end{aligned}$$

□

**Lemma 3.** *Let $\mathbf{e}$ be a unit vector, let $\Gamma_{r,R}$ be a piecewise smooth closed curve as specified in Equation 38 and the text thereafter, and let $\phi \mapsto F_\phi$ be the operator on continuous vector fields defined by Equation 36. Overloading notation by regarding $\mathbf{e}(\cdot) \equiv \mathbf{e}$ as a constant vector field, it holds that:*

$$\oint_{\Gamma_{r,R}} F_{\mathbf{e}} = \left(\frac{2N}{3-2/N} - 2\right)\left(R^{3-\frac{2}{N}} - r^{3-\frac{2}{N}}\right)$$

*Proof.* We compute the line integral by decomposing $\Gamma_{r,R}$ into its smooth components $\Gamma^1_{r,R} \ldots \Gamma^4_{r,R}$:

$$\oint_{\Gamma_{r,R}} F_{\mathbf{e}} = \int_{\Gamma^1_{r,R}} F_{\mathbf{e}} + \int_{\Gamma^2_{r,R}} F_{\mathbf{e}} + \int_{\Gamma^3_{r,R}} F_{\mathbf{e}} + \int_{\Gamma^4_{r,R}} F_{\mathbf{e}} \qquad (40)$$

Starting from $\Gamma^1_{r,R}$, notice that for every point $\mathbf{w}$ lying on this curve: $\langle \mathbf{e}, \frac{\mathbf{w}}{\|\mathbf{w}\|} \rangle \frac{\mathbf{w}}{\|\mathbf{w}\|} = \mathbf{e}$. Therefore:

$$\int_{\Gamma^1_{r,R}} F_{\mathbf{e}} = \int_{\Gamma^1_{r,R}} \|\mathbf{w}\|^{2-\frac{2}{N}} (\mathbf{e}+(N-1)\mathbf{e}) = N \int_{\Gamma^1_{r,R}} \|\mathbf{w}\|^{2-\frac{2}{N}} \mathbf{e}$$

The line integral on the right translates into a simple univariate integral:

$$\begin{aligned}\int_{\Gamma^1_{r,R}} \|\mathbf{w}\|^{2-\frac{2}{N}} \mathbf{e} &= \int_{-R}^{-r} |\rho|^{2-\frac{2}{N}} d\rho = \int_{r}^{R} \rho^{2-\frac{2}{N}} d\rho \\ &= \frac{1}{3-2/N} \left( R^{3-\frac{2}{N}} - r^{3-\frac{2}{N}} \right)\end{aligned}$$

We thus have:

$$\int_{\Gamma^1_{r,R}} F_{\mathbf{e}} = \frac{N}{3-2/N} \left( R^{3-\frac{2}{N}} - r^{3-\frac{2}{N}} \right) \tag{41}$$

Turning to $\Gamma^2_{r,R}$, note that for any point $\mathbf{w}$ along this curve $\|\mathbf{w}\|^{2-\frac{2}{N}} = r^{2-\frac{2}{N}}$, and $\frac{\mathbf{w}}{\|\mathbf{w}\|}$ is perpendicular to the direction of motion. This implies:

$$\int_{\Gamma^2_{r,R}} F_{\mathbf{e}} = r^{2-\frac{2}{N}} \int_{\Gamma^2_{r,R}} \mathbf{e}$$

The line integral $\int_{\Gamma^2_{r,R}} \mathbf{e}$ is simply equal to the progress $\Gamma^2_{r,R}$ makes in the direction of $\mathbf{e}$, which is $2r$. Accordingly:

$$\int_{\Gamma^2_{r,R}} F_{\mathbf{e}} = r^{2-\frac{2}{N}} \cdot 2r = 2r^{3-\frac{2}{N}} \tag{42}$$

As for $\Gamma^3_{r,R}$ and $\Gamma^4_{r,R}$, their line integrals may be computed similarly to those of $\Gamma^1_{r,R}$ and $\Gamma^2_{r,R}$ respectively. Such computations yield:

$$\int_{\Gamma^3_{r,R}} F_{\mathbf{e}} = \frac{N}{3-2/N} \left( R^{3-\frac{2}{N}} - r^{3-\frac{2}{N}} \right) \tag{43}$$

$$\int_{\Gamma^4_{r,R}} F_{\mathbf{e}} = -2R^{3-\frac{2}{N}} \tag{44}$$

Combining Equation 40 with Equations 41, 42, 43 and 44, we obtain the desired result. □

## B. A Concrete Acceleration Bound

In Section 7 we illustrated qualitatively, on a family of very simple hypothetical learning problems, the potential of overparameterization (use of depth-$N$ linear network in place of classic linear model) to accelerate optimization. In this appendix we demonstrate how the illustration can be made formal, by considering a special case and deriving a concrete bound on the acceleration.

In the context of Section 7, we will treat the setting of $p = 4$ ($\ell_4$ loss) and $N = 2$ (depth-2 network). We will also assume, in accordance with the problem being ill-conditioned – $y_1 \gg y_2$, that initialization values are ill-conditioned as well, and in particular $\epsilon_1/\epsilon_2 \approx y_1/y_2$, where $\epsilon_i := |w_i^{(0)}|$. An additional assumption we make is that $y_2$ is on the order of 1, and thus the near-zero initialization of $w_1$ and $w_2$ implies $y_2 \gg \epsilon_1, \epsilon_2$. Finally, we assume that $\epsilon_1 y_1 \gg 1$.

As shown in Section 7, under gradient descent, $w_1$ and $w_2$ move independently, and to prevent divergence, the learning rate must satisfy $\eta < \min\{2/y_1^{p-2}, 2/y_2^{p-2}\}$. In our setting, this translates to (GD below stands for gradient descent):

$$\eta^{GD} < 2/y_1^2 \tag{45}$$

For $w_2$, the optimal learning rate (convergence in a single step) is $1/y_2^2$, and the constraint above will lead to very slow convergence (see Equation 15 and its surrounding text).

Suppose now that we optimize via overparameterization, *i.e.* with the update rule in Equation 12 (single output). In our particular setting (recall, in addition to the above, that we omitted weight decay for simplicity – $\lambda = 0$), this update rule translates to:

$$\begin{aligned}&[w_1^{(t+1)}, w_2^{(t+1)}]^\top \leftarrow [w_1^{(t)}, w_2^{(t)}]^\top \\ &-\eta \Big( (w_1^{(t)})^2 + (w_2^{(t)})^2 \Big)^{1/2} \cdot [(w_1^{(t)} - y_1)^3, (w_2^{(t)} - y_2)^3]^\top \\ &-\eta \Big( (w_1^{(t)})^2 + (w_2^{(t)})^2 \Big)^{-1/2} \\ &\quad \cdot \big( w_1^{(t)} (w_1^{(t)} - y_1)^3 + w_2^{(t)} (w_2^{(t)} - y_2)^3 \big) \cdot [w_1^{(t)}, w_2^{(t)}]^\top\end{aligned} \tag{46}$$

For the first iteration ($t = 0$), replacing $\epsilon_i := |w_i^{(0)}|$, while recalling that $y_1 \gg y_2 \gg \epsilon_1 \gg \epsilon_2$, we obtain:

$$\begin{aligned}[w_1^{(1)}, w_2^{(1)}]^\top &\approx \eta \cdot \epsilon_1 \cdot [y_1^3, y_2^3]^\top + \eta \cdot y_1^3 \cdot [\epsilon_1, \epsilon_2]^\top \\ &= \eta \cdot [2\epsilon_1 y_1^3, \epsilon_1 y_2^3 + \epsilon_2 y_1^3]^\top\end{aligned}$$

Set $\eta = 1/2\epsilon_1 y_1^2$. Then $w_1^{(1)} \approx y_1$ and $w_2^{(1)} \approx y_2^3/2y_1^2 + \epsilon_2 y_1/2\epsilon_1$. Our assumptions thus far ($y_1 \gg y_2$ and $\epsilon_1 \gg \epsilon_2$) imply $w_1^{(1)} \gg w_2^{(1)}$. Moreover, since $\epsilon_2/\epsilon_1 \approx y_2/y_1$, it holds that $w_2^{(1)} \in \mathcal{O}(y_2) = \mathcal{O}(1)$. Taking all of this into account, the second iteration ($t = 1$) of the overparameterized update rule (Equation 46) becomes:

$$\begin{aligned}[w_1^{(2)}, w_2^{(2)}]^\top &\approx [y_1, w_2^{(1)}]^\top \\ &\quad - \frac{1}{2\epsilon_1 y_1} [(w_1^{(1)} - y_1)^3, (w_2^{(1)} - y_2)^3]^\top \\ &\quad - \frac{y_1 (w_1^{(1)} - y_1)^3 + w_2^{(1)} (w_2^{(1)} - y_2)^3}{2\epsilon_1 y_1^3} [y_1, w_2^{(1)}]^\top \\ &\approx [y_1, w_2^{(1)} - 1/2\epsilon_1 y_1 \cdot (w_2^{(1)} - y_2)^3]^\top\end{aligned}$$

In words, $w_1$ will stay approximately equal to $y_1$, whereas $w_2$ will take a step that corresponds to gradient descent with learning rate (OP below stands for overparameterization):

$$\eta^{OP} := 1/2\epsilon_1 y_1 \tag{47}$$

By assumption $\epsilon_1 y_1 \gg 1$ and $y_2 \in \mathcal{O}(1)$, thus $\eta^{OP} < 2/y_2^2$, meaning that $w_2$ will remain on the order of $y_2$ (or less). An inductive argument can therefore be applied, and our observation regarding the second iteration ($t = 1$) continues to hold throughout – $w_1$ is (approximately) fixed at $y_1$, and $w_2$ follows steps that correspond to gradient descent with learning rate $\eta^{OP}$.

To summarize our findings, we have shown that while standard gradient descent limits $w_2$ with a learning rate $\eta^{GD}$ that is at most $2/y_1^2$ (Equation 45), overparameterization can be adjusted to induce on $w_2$ an implicit gradient descent scheme with learning rate $\eta^{OP} = 1/2\epsilon_1 y_1$ (Equation 47), all while admitting immediate (single-step) convergence for $w_1$. Since both $\eta^{GD}$ and $\eta^{OP}$ are well below $1/y_2^2$, we obtain acceleration by at least $\eta^{OP}/\eta^{GD} > y_1/4\epsilon_1$ (we remind the reader that $y_1 \gg 1$ is the target value of $w_1$, and $\epsilon_1 \ll 1$ is the magnitude of its initialization).

## C. Implementation Details

Below we provide implementation details omitted from our experimental report (Section 8).

### C.1. Linear Neural Networks

The details hereafter apply to all of our experiments besides that on the convolutional network (Figure 5-right).

In accordance with our theoretical setup (Section 4), evaluated linear networks did not include bias terms, only weight matrices. The latter were initialized to small values, drawn i.i.d. from a Gaussian distribution with mean zero and standard deviation $0.01$. The only exception to this was the setting of identity initialization (Figure 5-left), in which an offset of $1$ was added to the diagonal elements of each weight matrix (including those that are not square).

When applying a grid search over learning rates, the values $\{10^{-5}, 5 \cdot 10^{-5}, \ldots, 10^{-1}, 5 \cdot 10^{-1}\}$ were tried. We note that in the case of depth-$8$ network with standard near-zero initialization (Figure 5-left), all learning rates led either to divergence, or to a failure to converge (vanishing gradients).

For computing optimal $\ell_2$ loss (used as an offset in respective convergence plots), we simply solved, in closed form, the corresponding least squares problem. For the optimal $\ell_4$ loss, we used `scipy.optimize.minimize` – a numerical optimizer built into SciPy (Jones et al., 2001–), with the default method of BFGS (Nocedal, 1980).

### C.2. Convolutional Network

For the experiment on TensorFlow's MNIST convolutional network tutorial, we simply downloaded the code,[7] and introduced two minor changes:

- Hidden dense layer: $3136 \times 512$ weight matrix replaced by multiplication of $3136 \times 512$ and $512 \times 512$ matrices.
- Output layer: $512 \times 10$ weight matrix replaced by multiplication of $512 \times 10$ and $10 \times 10$ matrices.

The newly introduced weight matrices were initialized in the same way as their predecessors (random Gaussian distribution with mean zero and standard deviation $0.1$). Besides the above, no change was made. An addition of roughly $250K$ parameters to a $1.6M$-parameter model gave the speedup presented in Figure 5-right.

To rule out the possibility of the speedup resulting from suboptimal learning rates, we reran the experiment with grid search over the latter. The learning rate hardcoded into the tutorial follows an exponentially decaying schedule, with base value $10^{-2}$. For both the original and overparameterized models, training was run multiple times, with the base value varying in $\{10^{-5}, 5 \cdot 10^{-5}, \ldots, 10^{-1}, 5 \cdot 10^{-1}\}$. We chose, for each model separately, the configuration giving fastest convergence, and then compared the models one against the other. The observed gap in convergence rates was similar to that in Figure 5-right.

An additional point we set out to examine, is the sensitivity of the speedup to initialization of overparameterized layers. For this purpose, we retrained the overparameterized model multiple times, varying in $\{10^{-3}, 5 \cdot 10^{-3}, \ldots, 10^{-1}, 5 \cdot 10^{-1}\}$ the standard deviation of the Gaussian distribution initializing overparameterized layers (as stated above, this standard deviation was originally set to $10^{-1}$). Convergence rates across the different runs were almost identical. In particular, they were all orders of magnitude faster than the convergence rate of the baseline, non-overparameterized model.

[7] https://github.com/tensorflow/models/tree/master/tutorials/image/mnist