

## APPENDIX

## A OUTLINE OF PROOF FOR THEOREM 1, COROLLARY 1, AND PROPOSITION 1

We now provide an outline of our results and proofs.

1. In Appendix A, we introduce Lemmas 1, 2, which will be used to prove Theorem 1.
2. In Appendix B, we provide the proof of Corollary 1 - the simplification of gradient descent under alignment - which relies on Lemma 2.
3. In Appendix C, we provide the proof of Proposition 1 - linear convergence under strong alignment - which relies on Corollary 1.
4. In Appendix D, we introduce Theorem 4, which is a generalization of Theorem 1 to fully connected networks with rectangular layers. We use Lemma 2 and Proposition 1 to prove Theorem 4.
5. In Appendix E, we finally prove Theorem 1, which follows from Theorem 4.

Here, we present two lemmas that will be used extensively in our proofs.

Clearly strong alignment being an invariant implies that alignment is an invariant. Now we show that alignment implies strong alignment in the case of networks with square matrix layers.

**Lemma 1.** *Let  $\{W_i\}_{i=1}^d \subset \mathbb{R}^{k \times k}$ , where  $d \geq 3$ . If alignment is an invariant of training under the squared loss for network  $f = W_d W_{d-1} \dots W_1$  on data  $(X, Y) \in \mathbb{R}^{k \times n} \times \mathbb{R}^{k \times n}$ , then strong alignment is also invariant.*

*Proof.* Assume that alignment is an invariant of training. Gradient descent on the objective

$$\arg \min_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n \|y^{(i)} - f(x^{(i)})\|_2^2 \quad (7)$$

proceeds via the following update rule:

$$W_i^{(t+1)} = W_i^{(t)} + \frac{\gamma}{n} (W_d^{(t)} \dots W_{i+1}^{(t)})^T \sum_{l=1}^n (y^{(l)} - f(x^{(l)})) (W_{i-1}^{(t)} \dots W_1^{(t)} x^{(l)})^T, \quad \forall i \in [d]. \quad (8)$$

Since alignment is an invariant, the initialization satisfies  $W_i^{(t)} = U_i \Sigma_i^{(t)} V_i^T$  for  $2 \leq i \leq d-1$ ,  $W_1^{(t)} = U_1 \Sigma_1^{(t)} V_1^T$ , and  $W_d^{(t)} = U_d^{(t)} \Sigma_d^{(t)} V_d^T$ , where  $U_i = V_{i+1}$  for  $i \in [d-1]$ . For  $2 \leq i \leq d-1$ , substituting into Equation (8) yields

$$\begin{aligned} W_i^{(t+1)} &= U_i \Sigma_i^{(t)} V_i^T + \frac{\gamma}{n} (U_d^{(t)} \Sigma_d^{(t)} \dots \Sigma_{i+1}^{(t)} V_{i+1}^T)^T \sum_{l=1}^n (y^{(l)} - f(x^{(l)})) (U_{i-1} \Sigma_{i-1}^{(t)} \dots \Sigma_1^{(t)} V_1^T x^{(l)})^T \\ &= U_i \left( \Sigma_i^{(t)} + \frac{\gamma}{n} \prod_{j=i+1}^d \Sigma_j^{(t)T} U_d^{(t)T} \sum_{l=1}^n (y^{(l)} - f(x^{(l)})) x^{(l)T} V_1^{(t)} \prod_{j=1}^{i-1} \Sigma_j^{(t)T} \right) V_i^T \\ &= U_i \left( \Sigma_i^{(t)} + \frac{\gamma}{n} \prod_{j=i+1}^d \Sigma_j^{(t)T} (U_d^{(t)T} Y X^T V_1^{(t)} - \Sigma_d^{(t)} \dots \Sigma_1^{(t)} V_1^{(t)T} X X^T V_1^{(t)}) \prod_{j=1}^{i-1} \Sigma_j^{(t)T} \right) V_i^T. \end{aligned}$$

Since alignment is an invariant, the quantity

$$\prod_{j=i+1}^d \Sigma_j^{(t)T} (U_d^{(t)T} Y X^T V_1^{(t)} - \Sigma_d^{(t)} \dots \Sigma_1^{(t)} V_1^{(t)T} X X^T V_1^{(t)}) \prod_{j=1}^{i-1} \Sigma_j^{(t)T} \quad (9)$$

is a diagonal matrix for all  $t$ . Since each of the  $\Sigma_j$  are square, full rank matrices, the quantity

$$U_d^{(t)T} Y X^T V_1^{(t)} - \Sigma_d^{(t)} \dots \Sigma_1^{(t)} V_1^{(t)T} X X^T V_1^{(t)}$$

must be diagonal for all  $t$ .

The update rule for  $W_1$  is given by

$$\begin{aligned} W_1^{(t+1)} &= W_1^{(t)} + \frac{\gamma}{n} (W_d^{(t)} \cdots W_2^{(t)})^T \sum_{l=1}^n (y^{(l)} - f(x^{(l)})) x^{(l)T} \\ U_1 \Sigma_1^{(t+1)} V_1^{(t+1)T} &= U_1 \Sigma_1^{(t)} V_1^{(t)T} + V_2 \prod_{j=2}^d \Sigma_j^{(t)T} U_d^{(t)T} (Y X^T - U_d \Sigma_d^{(t)} \cdots \Sigma_1^{(t)} V_1^{(t)T} X X^T) \\ \implies \Sigma_1^{(t+1)} V_1^{(t+1)T} V_1^{(t)} &= \Sigma_1^{(t)} + \prod_{j=2}^d \Sigma_j^{(t)T} (U_d^{(t)T} Y X^T V_1^{(t)} - \Sigma_d^{(t)} \cdots \Sigma_1^{(t)} V_1^{(t)T} X X^T V_1^{(t)}), \end{aligned}$$

which is diagonal. Therefore  $V_1^{(t+1)T} V_1^{(t)}$  is diagonal, and since this is also an orthogonal matrix we must have that  $V_1^{(t+1)} = V_1^{(t)}$ .

Similarly, the update rule for  $W_d$  is given by:

$$\begin{aligned} W_d^{(t+1)} &= W_d^{(t)} + \frac{\gamma}{n} \sum_{l=1}^n (y^{(l)} - f(x^{(l)})) x^{(l)T} (W_{d-1}^{(t)} \cdots W_1^{(t)})^T \\ U_d^{(t+1)} \Sigma_d^{(t+1)} V_d^{(t+1)T} &= U_d^{(t)} \Sigma_d^{(t)} V_d^{(t)T} + (Y X^T - U_d^{(t)} \Sigma_d^{(t)} \cdots \Sigma_1^{(t)} V_1^{(t)T} X X^T) V_1^{(t)} \prod_{j=1}^{d-1} \Sigma_j^{(t)T} U_{d-1}^{(t)T} \\ \implies U_d^{(t)T} U_d^{(t+1)} \Sigma_d^{(t+1)} &= \Sigma_d^{(t)} + (U_d^{(t)T} Y X^T V_1^{(t)} - \Sigma_d^{(t)} \cdots \Sigma_1^{(t)} V_1^{(t)T} X X^T V_1^{(t)}) \prod_{j=1}^{d-1} \Sigma_j^{(t)T} U_{d-1}^{(t)T}, \end{aligned}$$

which is diagonal. Therefore  $U_d^{(t)T} U_d^{(t+1)}$  is also diagonal, implying that  $U_d^{(t)} = U_d^{(t+1)}$ . Therefore strong alignment is also an invariant. This means that alignment being an invariant and strong alignment being an invariant are equivalent in the setting where all the  $k_i$  are equal.  $\square$

Now that we have shown the equivalence of alignment being an invariant and strong alignment being an invariant in the setting where all the layers are square, we prove the following lemma for the general case where the  $k_i$  are not necessarily all equal.

**Lemma 2.** *Let  $f : \mathbb{R}^{k_0} \rightarrow \mathbb{R}^{k_d}$  be a linear fully connected network as in Equation equation 1, and let  $r = \min(k_0, \dots, k_n)$ . For training under the squared loss on the dataset  $(X, Y)$ , there exists an aligned initialization  $f(x) = W_d^{(0)} \cdots W_1^{(0)} x$  such that  $W_i^{(t)} = U_i \Sigma_i^{(t)} V_i^T$  for all  $i \in [d]$  (that is,  $U_i, V_i$  are not updated) if and only if there exist orthonormal matrices  $U \in \mathbb{R}^{k_d \times k_d}$ ,  $V \in \mathbb{R}^{k_0 \times k_0}$  such that*

$$U^T Y X^T V = \begin{bmatrix} \Lambda' & \mathbf{0} \\ \mathbf{0} & A_1 \end{bmatrix}, \text{ and } V^T X X^T V = \begin{bmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & A_2 \end{bmatrix}$$

for diagonal  $r \times r$  matrices  $\Lambda, \Lambda'$  and arbitrary  $A_1 \in \mathbb{R}^{(k_0-r) \times (k_d-r)}$ ,  $A_2 \in \mathbb{R}^{(k_0-r) \times (k_0-r)}$ .

*Proof.* Gradient descent on the objective

$$\arg \min_{f \in \mathcal{F}} \frac{1}{2n} \sum_{i=1}^n \|y^{(i)} - f(x^{(i)})\|_2^2$$

proceeds via the following update rule:

$$W_i^{(t+1)} = W_i^{(t)} + \frac{\gamma}{n} (W_d^{(t)} \cdots W_{i+1}^{(t)})^T \sum_{l=1}^n (y^{(l)} - f(x^{(l)})) (W_{i-1}^{(t)} \cdots W_1^{(t)} x^{(l)})^T, \quad \forall i \in [d], \quad (10)$$

where  $\gamma$  is the learning rate and superscript  $(t)$  denotes the gradient descent step. Assume that the network is initialized to be aligned, that is, there exist orthonormal  $U_i, V_i$  and diagonal matrices  $\Sigma_i$  such that  $W_i = U_i \Sigma_i V_i^T$  and  $U_i = V_{i+1}$  for  $i \in [d-1]$ . Substituting into Equation (10) yields

$$\begin{aligned} W_i^{(t+1)} &= U_i \Sigma_i^{(t)} V_i^T + \frac{\gamma}{n} (U_d \Sigma_d^{(t)} \cdots \Sigma_{i+1}^{(t)} V_{i+1}^T)^T \sum_{l=1}^n (y^{(l)} - f(x^{(l)})) (U_{i-1} \Sigma_{i-1}^{(t)} \cdots \Sigma_1^{(t)} V_1^T x^{(l)})^T \\ &= U_i \left( \Sigma_i^{(t)} + \frac{\gamma}{n} \prod_{j=i+1}^d \Sigma_j^{(t)T} U_d^T \sum_{l=1}^n (y^{(l)} - f(x^{(l)})) x^{(l)T} V_1 \prod_{j=1}^{i-1} \Sigma_j^{(t)T} \right) V_i^T \\ &= U_i \left( \Sigma_i^{(t)} + \frac{\gamma}{n} \prod_{j=i+1}^d \Sigma_j^{(t)T} (U_d^T Y X^T V_1 - \Sigma_d^{(t)} \cdots \Sigma_1^{(t)} V_1^T X X^T V_1) \prod_{j=1}^{i-1} \Sigma_j^{(t)T} \right) V_i^T. \end{aligned}$$

Thus strong alignment is an invariant if and only if for all  $i$ , the quantity

$$\prod_{j=i+1}^d \Sigma_j^{(t)T} (U_d^T Y X^T V_1 - \Sigma_d^{(t)} \cdots \Sigma_1^{(t)} V_1^T X X^T V_1) \prod_{j=1}^{i-1} \Sigma_j^{(t)T}$$

is an  $k_i \times k_{i-1}$  diagonal matrix for all  $t$ . At initialization each of the  $\Sigma_j$  have rank at least  $r$ . Considering  $i=1$  and  $i=d$ , the above quantity is diagonal if and only if the matrix

$$U_d^T Y X^T V_1 - \Sigma_d^{(t)} \cdots \Sigma_1^{(t)} V_1^T X X^T V_1 \quad (11)$$

has its top  $r$  rows and top  $r$  columns all diagonal; i.e. we can write this expression as

$$\begin{bmatrix} D & \mathbf{0} \\ \mathbf{0} & A \end{bmatrix} \quad (12)$$

for an  $r \times r$  diagonal matrix  $D$  and an arbitrary  $(k_d - r) \times (k_0 - r)$  matrix  $A$ .

For the first direction, assume that strong alignment is an invariant, i.e. that Equation (11) can be written in the above block diagonal form. Define  $\Sigma_{tot}^{(t)} = \Sigma_d^{(t)} \cdots \Sigma_1^{(t)}$  – this is a diagonal matrix whose only nonzero entries are the first  $r$  on the diagonal. We know that

$$U_d^T Y X^T V_1 - \Sigma_{tot}^{(t)} V_1^T X X^T V_1$$

is of the form of Equation (12) for all gradient descent steps  $t$ , and thus the quantity

$$\left( \Sigma_{tot}^{(t)} - \Sigma_{tot}^{(0)} \right) V_1^T X X^T V_1$$

is of this form as well. Assuming that we've not initialized any of the singular values to be their optimal value (which is satisfied with probability 1), the top  $r$  diagonal entries of  $\Sigma_{tot}^{(t)} - \Sigma_{tot}^{(0)}$  are nonzero, which means that the top left  $r \times r$  submatrix of  $V_1^T X X^T V_1$  is diagonal, and that the top right submatrix consists of all zeros. But since  $V_1^T X X^T V_1$  is symmetric, the bottom left submatrix must also consist of all zeros, and thus we have

$$V_1^T X X^T V_1 = \begin{bmatrix} D_2 & \mathbf{0} \\ \mathbf{0} & A_2 \end{bmatrix}$$

for an  $r \times r$  diagonal matrix  $D_2$  and arbitrary  $(k_0 - r) \times (k_0 - r)$  matrix  $A_2$ . Plugging this into Equation (11) implies that  $U_d^T Y X^T V_1$  must be of this form as well.

We next show the other direction. Assume that for some orthonormal matrices  $U$  and  $V$ , it holds that  $V^T X X^T V$  is diagonal and  $U^T Y X^T V$  can be written in the block matrix form given by Equation (12). Initializing the layers such that  $U_d = U, V_1 = V$ , and  $U_i = V_{i+1}$  for  $i \in [d-1]$  implies that Equation (11) is also of this block diagonal form, as desired.  $\square$

## B PROOF OF COROLLARY 1

*Proof.* The conditions of strong alignment imply the conditions of Lemma 2, which in turn implies that there exist orthonormal matrices  $U, V$  such that

$$\begin{aligned} U^T Y X^T V &= \begin{bmatrix} \Lambda' & \mathbf{0} \\ \mathbf{0} & A_1 \end{bmatrix}, \text{ and} \\ V^T X X^T V &= \begin{bmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & A_2 \end{bmatrix}, \end{aligned}$$

where  $\Lambda, \Lambda'$  are  $r \times r$  diagonal matrices. Furthermore, from the proof of Theorem 1, if the layers are initialized to be aligned, with  $U_d = U$  and  $V_1 = V$ , then the gradient descent updates are as follows:

$$W_i^{(t+1)} = U_i \left( \Sigma_i^{(t)} + \frac{\gamma}{n} \prod_{j=i+1}^d \Sigma_j^{(t)T} (U^T Y X^T V_1 - \Sigma_d^{(t)} \dots \Sigma_1^{(t)} V^T X X^T V) \prod_{j=1}^{i-1} \Sigma_j^{(t)T} \right) V_i^T.$$

Since the minimum of the ranks of the  $\Sigma_i^{(t)}$  is  $r$ , only the top  $r$  singular values of  $W_i$  are updated. Plugging in the expressions for  $U^T Y X^T V$  and  $V^T X X^T V$  and restricting to the top  $r$  singular values (which we denote by  $\Sigma'_i$ ), we obtain the statement of Corollary 1, with the singular values of each layer being updated as:

$$\Sigma'_i{}^{(t+1)} = \Sigma'_i{}^{(t)} + \frac{\gamma}{n} \prod_{j \neq i} \Sigma'_j{}^{(t)} (\Lambda' - \prod_{j=1}^d \Sigma'_j{}^{(t)} \Lambda).$$

This completes the proof.  $\square$

### C PROOF OF PROPOSITION 1

*Proof.* By Corollary 1, under strong alignment, each singular value is updated independently of each other. Thus we can focus on how the  $k$ th singular value for each layer is updated. Recall that  $\sigma_k(W_i^{(t)})$  denotes the  $k$ th diagonal entry of  $\Sigma_i^{(t)}$ . Since we're focusing on a fixed  $k$ , we drop the subscript  $k$  for convenience and let  $\sigma_i^{(t)}$  equal  $\sigma_k(W_i^{(t)})$ . The  $\sigma$  are updated by the following update rule:

$$\sigma_i^{(t+1)} = \sigma_i^{(t)} + \frac{\gamma}{n} \prod_{j \neq i} \sigma_j^{(t)} (\lambda'_k - \lambda_k \prod_{j=1}^d \sigma_j^{(t)}),$$

where  $\lambda'_k, \lambda_k$  are the  $k$ th diagonal elements of  $\Lambda', \Lambda$ . We assume that  $\Lambda'$  and  $\Lambda$  have the same zero pattern. Therefore  $\lambda_k = 0$  if and only if  $\lambda'_k = 0$ . If both of these values are zero, then  $\sigma_i$  is not updated.

Otherwise, assume  $\lambda_k, \lambda'_k \neq 0$ . Note that  $\lambda_k > 0$ , since  $XX^T$  is positive semidefinite. We can also negate columns of  $U$  to ensure that  $\lambda'_k > 0$  as well. Let  $\eta = \frac{\gamma \lambda_k}{n}$ , and define  $S^{(t)} = \prod_{j=1}^d \sigma_j^{(t)}$ . This yields

$$\sigma_i^{(t+1)} = \sigma_i^{(t)} + \eta \frac{S^{(t)}}{\sigma_i^{(t)}} \left( \frac{\lambda'_k}{\lambda_k} - S^{(t)} \right). \quad (13)$$

Therefore (dropping the superscript to let  $S = S^{(t)}$ ),

$$\begin{aligned} S^{(t+1)} &= \prod_{i=1}^d \sigma_i^{(t+1)} = \prod_{i=1}^d \left( \sigma_i^{(t)} + \eta S \frac{1}{\sigma_i^{(t)}} \left( \frac{\lambda'_k}{\lambda_k} - S \right) \right) \\ &= S + \sum_{T \subset [d]: |T| \geq 1} \eta^{|T|} S^{|T|} \left( \frac{\lambda'_k}{\lambda_k} - S \right)^{|T|} \prod_{i \in T} \frac{1}{\sigma_i^{(t)}} \prod_{i \notin T} \sigma_i^{(t)} \\ &= S + \sum_{T \subset [d]: |T| \geq 1} \eta^{|T|} S^{|T|+1} \left( \frac{\lambda'_k}{\lambda_k} - S \right)^{|T|} \prod_{i \in T} \frac{1}{(\sigma_i^{(t)})^2}, \end{aligned}$$

and hence

$$\frac{\lambda'_k}{\lambda_k} - S^{(t+1)} = \frac{\lambda'_k}{\lambda_k} - S - \sum_{T \subset [d]: |T| \geq 1} \eta^{|T|} S^{|T|+1} \left( \frac{\lambda'_k}{\lambda_k} - S \right)^{|T|} \prod_{i \in T} \frac{1}{(\sigma_i^{(t)})^2} \quad (14)$$

$$= \left( \frac{\lambda'_k}{\lambda_k} - S \right) \left( 1 - \sum_{T \subset [d]: |T| \geq 1} \eta^{|T|} S^{|T|+1} \left( \frac{\lambda'_k}{\lambda_k} - S \right)^{|T|-1} \prod_{i \in T} \frac{1}{(\sigma_i^{(t)})^2} \right). \quad (15)$$

Thus we obtain

$$\frac{\lambda'_k}{\lambda_k} - S^{(t+1)} = \left( \frac{\lambda'_k}{\lambda_k} - S^{(t)} \right) \cdot r_k^{(t)}, \quad (16)$$

where

$$r_k^{(t)} = 1 - \sum_{T \subset [d]: |T| \geq 1} \eta^{|T|} S^{|T|+1} \left( \frac{\lambda'_k}{\lambda_k} - S \right)^{|T|-1} \prod_{i \in T} \frac{1}{(\sigma_i^{(t)})^2}. \quad (17)$$

We aim to bound  $r_k^{(t)}$  from both above and below. First, we show that  $r_k^{(t)}$  is nonnegative in order to prove the following lemma:

**Lemma 3.**  $0 < S^{(j)} \leq \frac{\lambda'_k}{\lambda_k}$  for all  $j \geq 0$ .

*Proof.* We proceed by induction. By the original assumptions in Proposition 1,  $0 < S^{(0)} \leq \frac{\lambda'_k}{\lambda_k}$ . Now assume that  $0 < S^{(j)} \leq \frac{\lambda'_k}{\lambda_k}$  for all  $j \leq t$ . By the update rule in Equation (13),  $\sigma_i^{(j+1)} \geq \sigma_i^{(j)}$ . Since  $\sigma_i^{(0)} > 0$ ,  $\sigma_i^{(j)} > 0$ , so  $S^{(j)} > 0$ . We also have that

$$\prod_{i \in T} \frac{1}{(\sigma_i^{(t)})^2} \leq \prod_{i \in T} \frac{1}{(\sigma_i^{(0)})^2} \leq \frac{1}{(\min_i \sigma_i^{(0)})^{2|T|}}.$$

Next, note that we can bound

$$S^{|T|+1} \left( \frac{\lambda'_k}{\lambda_k} - S \right)^{|T|-1} \leq \left( \frac{\lambda'_k}{\lambda_k} \right)^{2|T|}.$$

This means that we can upper bound the sum in Equation (17) as

$$\begin{aligned} \sum_{T \subset [d]: |T| \geq 1} \eta^{|T|} S^{|T|+1} \left( \frac{\lambda'_k}{\lambda_k} - S \right)^{|T|-1} \prod_{i \in T} \frac{1}{(\sigma_i^{(t)})^2} &\leq \sum_{T \subset [d]: |T| \geq 1} \eta^{|T|} (\min_i \sigma_i^{(0)})^{-2|T|} \left( \frac{\lambda'_k}{\lambda_k} \right)^{2|T|} \\ &= \left( 1 + \eta \cdot (\min_i \sigma_i^{(0)})^{-2} \left( \frac{\lambda'_k}{\lambda_k} \right)^2 \right)^d - 1. \end{aligned}$$

Since  $\gamma \leq \frac{n \ln 2}{d} \cdot \frac{\min_i (\sigma_i^{(0)})^2 \lambda_k}{\lambda_k^2}$ , we have that  $\eta \leq \ln 2 \cdot \frac{\min_i (\sigma_i^{(0)})^2}{d} \cdot \frac{\lambda_k^2}{\lambda_k^2}$ , and thus the right-hand side of the above expression can be upper bounded by

$$\left( 1 + \eta \cdot (\min_i \sigma_i^{(0)})^{-2} \right)^d - 1 \leq e^{d\eta(\min_i \sigma_i^{(0)})^{-2}} - 1 \leq e^{\ln 2} - 1 = 1.$$

Therefore  $r_k^{(t)} \geq 0$ . Plugging into Equation (16), since  $S^{(t)} = S \leq \frac{\lambda'_k}{\lambda_k}$ , we get that  $S^{(t+1)} \leq \frac{\lambda'_k}{\lambda_k}$ , which completes the inductive step.  $\square$

Next, we would like to upper bound  $r_k^{(t)}$  by a term independent of  $t$  in order to obtain linear convergence. We can lower bound the sum in Equation (17) by the sets with size 1, so

$$\sum_{T \subset [d]: |T| \geq 1} \eta^{|T|} S^{|T|+1} \left( \frac{\lambda'_k}{\lambda_k} - S \right)^{|T|-1} \prod_{i \in T} \frac{1}{(\sigma_i^{(t)})^2} \geq \sum_{i=1}^d \eta S^2 \frac{1}{(\sigma_i^{(t)})^2} \geq \eta S^2 \cdot d S^{-2/d},$$

where the last inequality is due to AM-GM. Lemma 3 implies that  $S^{(j+1)} \geq S^{(j)}$ , which means that the above sum is at least  $\eta d (S^{(0)})^{2-2/d}$ , which means that we can upper bound  $r_k^{(t)}$  by

$$r_k^{(t)} \leq 1 - \eta d (S^{(0)})^{2-2/d}.$$

This implies that  $S^{(t+1)}$  is closer to  $\frac{\lambda'_k}{\lambda_k}$  than  $S$  is, and in particular

$$\frac{\lambda'_k}{\lambda_k} - S^{(t+1)} \leq \left( \frac{\lambda'_k}{\lambda_k} - S \right) (1 - d\eta (S^{(0)})^{2-2/d});$$

hence

$$\frac{\lambda'_k}{\lambda_k} - S^{(t)} \leq \left( \frac{\lambda'_k}{\lambda_k} - S^{(0)} \right) (1 - d\eta(S^{(0)})^{2-2/d})^t.$$

Since the initialization is fixed, the quantity  $1 - d\eta(S^{(0)})^{2-2/d}$  is fixed, and thus  $S^{(t)}$  converges linearly to  $\frac{\lambda'_k}{\lambda_k}$ . Therefore each of the top  $k$  singular values converge linearly to their optimal value  $\frac{\lambda'_k}{\lambda_k}$ , which means that the loss converges linearly as well.

To complete the proof, it suffices to show that this limit solution achieves a training loss of zero. This is proven in a more general setting at the end of Appendix E.  $\square$

#### D PROOF OF THEOREM 4

We can finally state the generalization of Theorem 1 to the non-square setting:

**Theorem 4.** *Let  $f : \mathbb{R}^{k_0} \rightarrow \mathbb{R}^{k_d}$  be a linear fully connected network as in Equation equation 1, and let  $r = \min(k_0, \dots, k_n)$ . Strong alignment is an invariant of training under the squared loss on the dataset  $(X, Y)$  if and only if there exist orthonormal matrices  $U \in \mathbb{R}^{k_d \times k_d}$ ,  $V \in \mathbb{R}^{k_0 \times k_0}$  such that*

$$U^T Y X^T V = \begin{bmatrix} \Lambda' & \mathbf{0} \\ \mathbf{0} & A_1 \end{bmatrix}, \text{ and } V^T X X^T V = \begin{bmatrix} \Lambda & \mathbf{0} \\ \mathbf{0} & A_2 \end{bmatrix}$$

for diagonal  $r \times r$  matrices  $\Lambda, \Lambda'$  and arbitrary  $A_1 \in \mathbb{R}^{(k_0-r) \times (k_d-r)}$ ,  $A_2 \in \mathbb{R}^{(k_0-r) \times (k_0-r)}$ .

*Proof.* By Lemma 2 we know that under strong alignment there exist  $U$  and  $V$  satisfying the above conditions. In the other direction, Lemma 2 also tells us that given  $U$  and  $V$  satisfying the data conditions, all the conditions of strong alignment hold except for convergence to a global minimum.

To conclude, we must show that regardless of the zero pattern of  $\Lambda$  or  $\Lambda'$ , under a strongly aligned initialization the network converges to a solution with a loss of zero.

Using the convenient notation that  $\sigma_i^{(t)} = \sigma_k(W_i^{(t)})$ , we again focus on how the  $k$ th singular values of each layer are updated, for some  $k \in [r]$ . Recall that the  $\sigma$ 's are updated as

$$\sigma_i^{(t+1)} = \sigma_i^{(t)} + \frac{\gamma}{n} \prod_{j \neq i} \sigma_j^{(t)} (\lambda'_k - \lambda_k \prod_{j=1}^d \sigma_j^{(t)}).$$

The rank of  $X$  must be at least the rank of  $Y$  in order for the data to be linearly interpolated. Therefore we can choose  $U, V$  (via permuting columns) to ensure that whenever  $\lambda_k = 0$ ,  $\lambda'_k = 0$  as well. This ensures that  $\sigma_k(W_i^{(t)})$  is never updated. If  $\lambda_k, \lambda'_k \neq 0$ , then we showed in Proposition 1 that  $S^{(t)}$  converges to  $\lambda'_k / \lambda_k$  in the limit.

Finally, we consider the case where  $\lambda'_k = 0, \lambda_k \neq 0$ . Assume that  $\sigma_i^{(t)} < 1$  and  $\gamma < \frac{n}{\lambda_k}$ . Then, the  $\sigma_i$ 's update as

$$\sigma_i^{(t+1)} = \sigma_i^{(t)} + \frac{\gamma}{n} \prod_{j \neq i} \sigma_j^{(t)} \left( -\lambda_k \prod_{j=1}^d \sigma_j^{(t)} \right) = \sigma_i^{(t)} \left( 1 - \eta \prod_{j \neq i} (\sigma_j^{(t)})^2 \right),$$

where  $\eta = \frac{\gamma \lambda_k}{n}$ . We observe that  $0 \leq \sigma_i^{(t+1)} \leq \sigma_i^{(t)}$ . Therefore

$$0 \leq S^{(t+1)} = S^{(t)} \prod_{i=1}^d \left( 1 - \eta \prod_{j \neq i} (\sigma_j^{(t)})^2 \right) \leq S^{(t)} \exp \left( -\eta \sum_{i=1}^d \prod_{j \neq i} (\sigma_j^{(t)})^2 \right) \leq S^{(t)} \exp \left( -\eta d S^{(t)2-2/d} \right).$$

Since  $S^{(0)}$  is positive, we see that  $0 \leq S^{(t+1)} \leq S^{(t)}$ , and therefore  $S^{(t)}$  must converge to some constant  $c$ . Assume that  $c \neq 0$ . For all  $\epsilon > 0$ , there exists some  $t$  such that  $S^{(T)} < c + \epsilon$ . Then,

$$S^{(T+1)} \leq S^{(T)} \exp \left( -\eta d S^{(T)2-2/d} \right) < (c + \epsilon) \exp \left( -\eta c^{2-2/d} \right),$$

where  $\exp(-\eta c^{2-2/d})$  is a constant which is less than 1. Hence if we choose  $\epsilon$  such that  $\exp(-\eta c^{2-2/d}) < \frac{c \pm \epsilon}{c}$ , then  $S^{(T+1)} < c$ , a contradiction. Therefore  $c = 0$ , and hence  $S^{(t)} \rightarrow 0 = \lambda'_k / \lambda_k$ .

In general, we have shown that if  $\lambda_k \neq 0$ , then  $\sigma_k(W_1(t)) \cdots \sigma_k(W_d(t)) \rightarrow \lambda'_k / \lambda_k$ . This solution is given by  $f(x) = U_d \Lambda' \Lambda^{-1} V_1^T x$ , which is the solution given by the pseudoinverse which obviously has a loss of zero.  $\square$

## E COMPLETING THE PROOF OF THEOREM 1

*Proof.* In Lemma 1, we showed that in the setting where all layers are square, alignment is equivalent to strong alignment. Theorem 4 states that in general, strong alignment is an invariant if and only if there exist  $U, V$  satisfying particular data conditions. Since in the square setting  $r = k$ , by Theorem 4 we have that strong alignment is an invariant if and only if there exist  $U, V$  such that  $U^T Y X^T V$  and  $V^T X X^T V$  are diagonal, as desired.  $\square$

## F ALIGNMENT FOR 1-DIMENSIONAL OUTPUTS

**Proposition 5.** *Assuming gradient descent avoids the point where all parameters are zero, alignment is an invariant of training for any linear fully connected network  $f : \mathbb{R}^{k_0} \rightarrow \mathbb{R}$ , any convex, twice continuously differentiable loss function, and data  $(X, Y) \in \mathbb{R}^{k_0 \times n} \times \mathbb{R}^{1 \times n}$  for which the network can achieve zero training error.*

*Proof.* If we initialize the weight matrices to be rank 1 and aligned, then the matrices  $\{\Sigma_i^{(t)}\}_{i=1}^d$  are diagonal with a single non-zero entry. Following the proof of Theorem 1, we obtain that alignment is an invariant if the matrix

$$\prod_{j=i+1}^d \Sigma_j^{(t)T} \left( U_d^T \sum_{k=1}^n \frac{\partial \ell}{\partial f} \Big|_{(x^{(k)}, y^{(k)})} x^{(k)T} V_1^{(t)} \right) \prod_{j=1}^{i-1} \Sigma_j^{(t)T}$$

is diagonal. When  $i \neq 1, d$ , this matrix is clearly of rank 1 and diagonal (and has a single nonzero entry). This implies that  $U_i, V_i$  are invariant for all  $i \neq 1, d$ . If  $i = d$ , then since  $k_d = 1$ , the above quantity is also a rank 1 diagonal matrix, implying that  $U_d$  and  $V_d$  are invariant. Finally, if  $i = 1$ , the above matrix is rank-1 but not necessarily diagonal. However, all but the top row are zeros, which after plugging into the gradient descent update rule implies that  $U_1$  is invariant as well. Importantly, layers  $W_{i+1}, W_i$  for  $i \in [d-1]$  remain aligned regardless of the loss function used, as the expression above is always a diagonal matrix with a single nonzero entry when the layers are initialized to be rank 1. The final step is to show that training leads to zero error according to Definition 3. To do this, we first characterize the stationary points and then under assumptions, we prove that the loss converges to zero.

We now characterize the stationary points of the above update. Let  $v_1^{(t)}$  denote the first column of  $V_1^{(t)}$ , and let  $\sigma_1(W_j^{(t)})$  denote the top singular value in the usSVD of  $W_j^{(t)}$ . Then the stationary points are given by:

1.  $\sigma_1(W_j^{(t)}) = 0$  for  $j \in [d]$ .
2.  $v_1^{(t)} \perp \sum_{k=1}^n \frac{\partial \ell}{\partial f} \Big|_{(x^{(k)}, y^{(k)})} x^{(k)T}$

If we initialize  $\sigma_1(W_1^{(0)}) = 0$ , then we have that:

$$\begin{aligned} \sigma_1(W_1^{(t)}) v_1^{(t)T} &= \sum_{k=1}^n c_k^{(t)} x^{(k)T} \\ c_k^{(t+1)} &= \sum_{k=1}^n \left( c_k^{(t)} + \gamma \prod_{j \neq k} \sigma_1(W_j^{(t)}) \frac{\partial \ell}{\partial f} \Big|_{(x^{(k)}, y^{(k)})} \right) x^{(k)T} \end{aligned}$$

for  $c_k^{(t)} \in \mathbb{R}$  and  $\forall t \in \mathbb{Z}_{\geq 0}$ . Hence, updates to  $v_1^{(t)}$  are in the span of the data, and so assuming that  $\{x^{(k)}\}_{k=1}^n$  are linearly independent,  $v_1^{(t)}$  cannot be orthogonal to  $\sum_{k=1}^n \frac{\partial \ell}{\partial f} \Big|_{(x^{(k)}, y^{(k)})} x^{(k)T}$  unless the  $c_k^{(t)}$  are all 0, i.e.  $\sigma_1(W_1^{(t)}) = 0$  for  $t > 0$ .

Next, if we initialize  $\sigma_1(W_i^{(0)}) = \sigma_1(W_j^{(0)})$ , then  $\sigma_1(W_i^{(t)}) = \sigma_1(W_j^{(t)})$  for all  $i, j \in \{2, \dots, d\}$ ,  $t \geq 0$  since for all  $i \in \{2, \dots, d\}$ :

$$\sigma_1(W_i^{(t+1)}) = \sigma_1(W_i^{(t)}) + \prod_{j \neq i} \sigma_1(W_j^{(t)}) \left( \sum_{k=1}^n \frac{\partial \ell}{\partial f} \Big|_{(x^{(k)}, y^{(k)})} x^{(k)T} v_1^{(t)} \right)$$

This initialization corresponds to layers  $W_{i+1}, W_i$  being balanced for  $i \in \{2, \dots, d\}$ . Thus, under this initialization, the only other stationary point is given by  $\sigma_1(W_i^{(t)}) = 0$  for all  $i \in \{2, \dots, d\}$ .

Hence, if gradient descent avoids the non-strict saddle points given by  $\sigma_1(W_i^{(t)}) = 0$  for all  $i \in \{2, \dots, d\}$  and  $\sigma_1(W_i^{(t)}) = 0$  for all  $i \in [d]$ , then gradient descent converges to a local (and thus global) minimum of the convex loss. The former stationary point can be avoided by re-parameterizing the network such that  $\sigma_1(W_i^{(t)}) = \sigma_1$  for all  $i \in \{2, \dots, d\}$  (i.e.  $\sigma_1 = 0$  now corresponds to a strict saddle as defined in Lee et al. (2016)), and then taking a random initialization for  $\sigma_1$ . This would correspond to gradient descent on the original parameterization with a scaling factor on the learning rate for parameters  $\sigma_1(W_i^{(t)})$  for  $i \in \{2, \dots, d\}$ . The latter stationary point is avoided by the assumption in the proposition.  $\square$

## G PROOF OF PROPOSITION 2

*Proof.* For any matrices  $A, B \in \mathbb{C}^{m \times n}$ , we have that  $2\sigma_i(AB^*) \leq \sigma_i(A^*A + B^*B)$  (Bhatia, 1997). Thus letting  $A = W_2, B = W_1^T$ , we see that

$$\begin{aligned} 2\sigma_i(W_2W_1) &\leq \sigma_i(W_2^T W_2 + W_1W_1^T) \\ \implies 2 \sum_i \sigma_i(P) &\leq \sum_i \sigma_i(W_2^T W_2 + W_1W_1^T) \\ &= \|W_2^T W_2 + W_1W_1^T\|_1 \\ &\leq \|W_2^T W_2\|_1 + \|W_1W_1^T\|_1 \\ &= \|W_2\|_F^2 + \|W_1\|_F^2 \end{aligned}$$

This lower bound is in fact achieved for an aligned solution. If the SVD of  $P$  is  $P = U\Sigma V^T$ , setting  $W_1 = W\Sigma^{\frac{1}{2}}U^T$  and  $W_2 = U\Sigma^{\frac{1}{2}}V^T$  yields  $\|W_1\|_F^2 = \|W_2\|_F^2 = \text{Tr}(\Sigma)$ , so  $\|W_1\|_F^2 + \|W_2\|_F^2 = 2\text{Tr}(\Sigma)$ .  $\square$

## H PROOF OF THEOREM 2

*Proof.* Given an arbitrary loss function, assume that the  $i$ th layer is restricted to some structure given by a subspace  $\mathcal{S}$  and basis matrices  $A_1, \dots, A_m$ , so that at timestep  $t$  we have that

$$W_i^{(t)} = \sum_{j=1}^m (c_j^i)^{(t)} A_j$$

We take the gradient of the loss with respect to the  $c_j^i$ . The chain rule yields:

$$\frac{\partial l}{\partial c_j^i} = \sum_{p,q=1}^n \frac{\partial l}{\partial (W_i)_{pq}} \cdot \frac{\partial (W_i)_{pq}}{\partial c_j^i} = \sum_{p,q=1}^n \frac{\partial l}{\partial (W_i)_{pq}} \cdot A_{pq}^j$$

The gradient descent update on  $c_j^i$  is thus:

$$(c_j^i)^{(t+1)} = (c_j^i)^{(t)} - \eta \cdot \frac{\partial l}{\partial c_j^i} = (c_j^i)^{(t)} - \eta \sum_{p,q=1}^n \frac{\partial l}{\partial (W_i)_{pq}} \cdot A_{pq}^j$$

The corresponding update on  $W^i$  becomes

$$\begin{aligned} W_i^{(t+1)} &= \sum_{j=1}^m (c_j^i)^{(t+1)} A_j \\ &= \sum_{j=1}^m (c_j^i)^{(t)} A_j - \eta \sum_{j=1}^m \sum_{p,q=1}^n \frac{\partial l}{\partial (W_i)_{pq}} \cdot A_{pq}^j A_j \\ &= W_i^{(t)} - \eta \sum_{j=1}^m \sum_{p,q=1}^n \frac{\partial l}{\partial (W_i)_{pq}} \cdot A_{pq}^j A_j \end{aligned}$$

We calculate the projection operator  $\pi$  of some arbitrary matrix  $M$  onto  $\mathcal{S}$ . We can write

$$\pi(M) = \sum_{j=1}^m \frac{\langle M, A^j \rangle A^j}{\|A_j\|_2^2} = \sum_{j=1}^m \sum_{p,q=1}^m \frac{M_{pq} A_{pq}^j A^j}{\|A_j\|_2^2}.$$

If we define the operator  $\pi_{\mathcal{S}}$  as

$$\pi_{\mathcal{S}}(M) = \sum_{j=1}^m \langle M, A^j \rangle A^j = \sum_{j=1}^m \sum_{p,q=1}^m M_{pq} A_{pq}^j A^j,$$

then gradient descent on the  $c$  gives the following update rule on the  $W^i$ :

$$W_i^{(t+1)} = W_i^{(t)} - \eta \cdot \pi_{\mathcal{S}} \left( \frac{\partial l}{\partial W_i} \right).$$

If the  $A_j$  all have norm 1, then,  $\pi = \pi_{\mathcal{S}}$ , and this is the same update rule given by projected gradient descent with respect to the subspace  $\mathcal{S}$ . Otherwise,  $\pi_{\mathcal{S}}$  is simply the projection  $\pi$  followed by appropriate scaling in each of the basis directions.  $\square$

## I TREATING A CONVOLUTIONAL LAYER AS A LINEAR SUBSPACE

Consider a  $3 \times 3$  image. We map it to a 9-dimensional vector as follows

$$\begin{bmatrix} x_1 & x_2 & x_3 \\ x_4 & x_5 & x_6 \\ x_7 & x_8 & x_9 \end{bmatrix} \implies [x_1 \ x_2 \ x_3 \ x_4 \ x_5 \ x_6 \ x_7 \ x_8 \ x_9]^T.$$

Then, the linear transformation given by applying the  $3 \times 3$  convolutional filter  $\begin{bmatrix} c_1 & c_2 & c_3 \\ c_4 & c_5 & c_6 \\ c_7 & c_8 & c_9 \end{bmatrix}$  is

given by the matrix

$$W = \begin{bmatrix} c_5 & c_4 & 0 & c_2 & c_1 & 0 & 0 & 0 & 0 \\ c_6 & c_5 & c_4 & c_3 & c_2 & c_1 & 0 & 0 & 0 \\ 0 & c_6 & c_5 & 0 & c_3 & c_2 & 0 & 0 & 0 \\ c_8 & c_7 & 0 & c_5 & c_4 & 0 & c_2 & c_1 & 0 \\ c_9 & c_8 & c_7 & c_6 & c_5 & c_4 & c_3 & c_2 & c_1 \\ 0 & c_9 & c_8 & 0 & c_6 & c_5 & 0 & c_3 & c_2 \\ 0 & 0 & 0 & c_8 & c_7 & 0 & c_5 & c_4 & 0 \\ 0 & 0 & 0 & c_9 & c_8 & c_7 & c_6 & c_5 & c_4 \\ 0 & 0 & 0 & 0 & c_9 & c_8 & 0 & c_6 & c_5 \end{bmatrix}.$$

Then  $\mathcal{S}$  consists of all matrices of the form  $W$ .  $\mathcal{S}$  is a 9-dimensional subspace of  $\mathbb{R}^{9 \times 9}$ , with an orthonormal basis with coefficients being the  $c_i$ .

## J PROOF OF PROPOSITION 3

*Proof.* For  $i \in [d]$ , let  $U_i \Sigma_i V_i^T$  be a usSVD of  $W_i$  witnessing alignment of  $f$ . We can then rewrite  $Y = f(X)$  as  $Y = U_d \prod_{i=1}^d \Sigma_i V_i^T X$ , thus proving the desired statement.  $\square$

## K PROOF OF PROPOSITION 4

Before we can prove Proposition 4, we require the following definition from combinatorics.

**Definition 5.** A partition of an integer  $k$  is a tuple  $\lambda = (\lambda_1, \dots, \lambda_s)$  such that  $\lambda_i \geq \lambda_{i+1}$  for all  $i$  and  $k = \lambda_1 + \dots + \lambda_s$ . Each  $\lambda_i$  is called a part of  $\lambda$ . We let  $s(\lambda)$  denote the number of parts of  $\lambda$  and we write  $\lambda \vdash k$  to indicate that  $\lambda$  is a partition of  $k$ .

*Proof of Proposition 4.* Given a  $k \times k$  matrix  $A$ , let  $\lambda(A)$  denote the partition  $\lambda$  of  $k$  such that  $\lambda_i$  is the multiplicity of the  $i^{\text{th}}$  greatest singular value of  $A$ . Let  $U(A)$  denote the set of matrices  $U$  such that  $U \Sigma V^T$  is a usSVD of  $A$ . The dimension of  $U(A)$  is

$$\sum_{i=1}^{s(\lambda(A))} \binom{\lambda_i}{2}.$$

To see this, note that any orthonormal basis of the eigenspace of  $AA^T$  corresponding to the multiplicity- $\lambda_i$  eigenvalue of  $AA^T$  can be the corresponding columns in an element of  $U(A)$  and that the set of orthonormal bases of an  $m$ -dimensional linear space is  $\binom{m}{2}$ .

For any set  $Q$  of matrices, Define  $U(Q)$  to be the set of all possible sets of left-singular vectors of elements of  $S$ . That is,

$$U(Q) := \bigcup_{A \in Q} U(A).$$

For each partition  $\lambda$  of  $k$ , let  $T_\lambda$  denote the set of matrices  $A$  such that  $\lambda(A) = \lambda$ . The dimension of  $T_\lambda \cap S$  is at most  $r$  and therefore the dimension of  $U(S \cap T_\lambda)$  is at most

$$r + \sum_{i=1}^{s(\lambda)} \binom{\lambda_i}{2}.$$

Let  $\mathcal{O}(k, n)$  denote the set of  $k \times n$  matrices with orthonormal columns. Assume alignment is possible over  $S$  for a non-measure-zero set of matrices with  $n$  columns. Then there exists  $B \subseteq \mathcal{O}(k, n)$  with  $\dim(B) = \dim(\mathcal{O}(k, n))$  such that for every  $U' \in B$ ,  $U(S)$  contains a matrix whose first  $n$  columns are  $U'$ . Therefore  $\dim(U(S)) \geq \dim(\mathcal{O}(k, n))$ . Since  $\dim(\mathcal{O}(k, n)) = \binom{k}{2} - \binom{k-n}{2}$ , the following must be satisfied for some  $\lambda \vdash k$

$$r + \sum_{i=1}^{s(\lambda)} \binom{\lambda_i}{2} \geq \binom{k}{2} - \binom{k-n}{2}. \quad (18)$$

This is attained when  $\lambda = (k)$ , but in this case  $T_\lambda$  is simply the set of scalar multiples of the identity. If we forbid  $\lambda = (k)$ , then we claim that the maximum value of  $r + \sum_{i=1}^{s(\lambda)} \binom{\lambda_i}{2}$  is attained by  $\lambda = (k-1, 1)$ . To see this, note that for all  $p < q$ ,

$$\binom{q-p}{2} + \binom{p}{2} = \binom{q}{2} - p(q-p) < \binom{q}{2}.$$

For  $p > 0$ , this is maximized when  $p = 1$ . This implies that the maximum value of  $\sum_{i=1}^{s(\lambda)} \binom{\lambda_i}{2}$  will be obtained in as few summands as possible (which in our case is two), and in particular when  $\lambda_1 = k-1$  and  $\lambda_2 = 1$ . In this case, equation 18 becomes

$$r + \binom{k-1}{2} \geq \binom{k}{2} - \binom{k-n}{2}.$$

Taking the logical negation of the above inequality and simplifying gives  $r < k-1 - \binom{k-n}{2}$ .  $\square$

## L ADDITIONAL EXPERIMENTS

We provide the following empirical evidence demonstrating that when the conditions of Theorem 1 are satisfied, invariance of alignment can indeed be observed empirically. We use a 2-hidden layer fully connected network with 9 hidden units per layer.

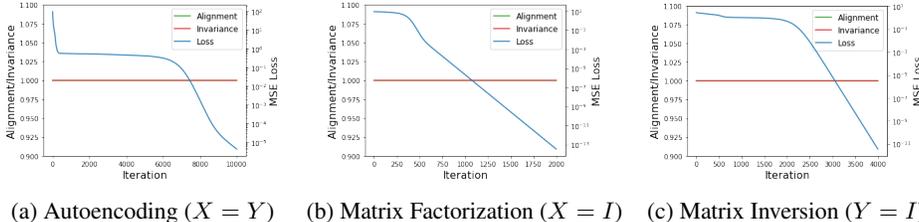


Figure 3: As proven in our work, alignment is an invariant of training when  $X, Y$  satisfy the conditions of Theorem 1.

## M EXPERIMENTAL SETUP

We provide network architectures and hyperparameters used for our experiments below. We trained our networks on an NVIDIA TITAN RTX GPU using the PyTorch library. In all settings, we train using gradient descent with a learning rate of  $10^{-2}$  until the loss was below  $10^{-4}$ .

1. Figure 1a: We use a 2-hidden layer fully connected network with 9 hidden units per layer. Our data is given by matrices  $(X, Y) \in \mathbb{R}^{9 \times 9}$  where each matrix entry is drawn from a standard normal distribution.
2. Figure 1b: We use a 2-hidden layer fully connected network with 1024 hidden units in the first hidden layer and 64 hidden units in the second hidden layer. Our data consists of 256 linearly separable examples from MNIST and is trained using Squared Loss.
3. Figure 1c: We use a 2-hidden layer fully connected network with 1024 hidden units in the first hidden layer and 64 hidden units in the second hidden layer. Our data consists of 256 linearly separable examples from MNIST and is trained using Cross Entropy Loss.
4. Figure 2a: We use a 2-hidden layer network with 4 hidden units per layer, where each layer is constrained to be a Toeplitz matrix. Our input  $X$  is equal to the identity, and our output  $Y$  is a  $4 \times 4$  matrix with each entry sampled from a standard normal distribution.
5. Figure 2b: We use a 2-hidden layer convolutional network with a single  $3 \times 3$  filter in each layer, stride of 1, and padding of 1. Our data consists of a single example from MNIST.

Code for the experiments can be found at the following anonymized github link: <https://anonymous.4open.science/r/33277cc0-6074-46c4-8642-7feadd678278/>.