Appendix

We first briefly outline the different sections in the appendix.

- In appendix A, we provide details of our experimental setup, and provide additional empirical results on fully connected networks, convolutional networks and residual networks with the MNIST, CIFAR-10 and CIFAR-100 datasets.
- In appendix B, we state and prove a lemma on the near orthogonality of random vectors, which we refer to in the main text. This result is often attributed to Milman & Schechtman (1986).
- In appendix C, we provide some intuition on why many standard over-parameterized neural networks with low-rank Hessians might have low gradient confusion for a large set of weights near the minimizer.
- In appendix D, we provide the proofs of the theorems presented in the main section. In appendix D.1, we provide proofs of theorems 3.1 and 3.2. In appendix D.2, we provide the proof of lemma D.1, which we refer to in the main text. In appendix D.3, we provide proofs of theorem 5.1 and corollary 5.1. In appendix D.4, we provide the proof of theorem 4.1. In appendix D.5, we provide the proof of theorem 6.1.
- In appendix E, we briefly describe a few lemmas that we require in our analysis.
- In appendix F, we discuss the small weights assumption (assumption 1), which is required for theorem 5.1, corollary 5.1 and theorem 6.1 in the main text.

A. Additional experimental results

In this section, we present more details about our experimental setup, as well as, additional experimental results on a range of models (MLPs, CNNs and Wide ResNets) and a range of datasets (MNIST, CIFAR-10, CIFAR-100).

A.1. MLPs on MNIST

To further test the main claims in the paper, we performed additional experiments on an image classification problem on the MNIST dataset using fully connected neural networks. We iterated over neural networks of varying depth and width, and considered both the identity activation function (i.e., linear neural networks) and the tanh activation function. We also considered two different weight initializations that are popularly used and appropriate for these activation functions:

- The Glorot normal initializer (Glorot & Bengio, 2010) with weights initialized by sampling from the distribution $\mathcal{N}\big(0, 2/(\text{fan-in} + \text{fan-out})\big)$, where fan-in denotes the number of input units in the weight matrix, and fan-out denotes the number of output units in the weight matrix.
- The LeCun normal initializer (LeCun et al., 2012) with weights initialized by sampling from the distribution $\mathcal{N}\big(0, 1/\text{fan-in}\big)$.

We considered the simplified case where all hidden layers have the same width ℓ. Thus, the first weight matrix $\mathbf{W}_0 \in \mathbb{R}^{\ell \times d}$, where $d = 784$ for the 28×28-sized images of MNIST; all intermediate weight matrices $\{\mathbf{W}_p\}_{p\in[\beta-1]} \in \mathbb{R}^{\ell\times\ell}$; and the final layer $\mathbf{W}_\beta \in \mathbb{R}^{10\times\ell}$ for the 10 image classes in MNIST. We added biases to each layer, which we initialized to 0. We used softmax cross entropy as the loss function. We use MLP-β-ℓ to denote this fully connected network of depth β and width ℓ. We used the standard train-valid-test splits of 40000-10000-10000 for MNIST.

This relatively simple model gave us the ability to iterate over a large number of combinations of network architectures of varying width and depth, and different activation functions and weight initializations. Linear neural networks are an efficient way to directly understand the effect of changing depth and width without increasing model complexity over linear regression. Thus, we considered both linear and non-linear neural networks in our experiments.

We used SGD with constant learning rates for training with a mini-batch size of 128 and trained each model for 40000 iterations (more than 100 epochs). The constant learning rate α was tuned over a logarithmically-spaced grid:

$$\alpha \in \{10^0, 10^{-1}, 10^{-2}, 10^{-3}, 10^{-4}, 10^{-5}, 10^{-6}\}.$$

We ran each experiment 10 times (making sure at least 8 of them ran till completion), and picked the learning rate that achieved the lowest training loss value on average at the end of training. Our grid search was such that the optimal learning rate never occurred at one of the extreme values tested.

To measure gradient confusion at the end training, we sampled 1000 pairs of mini-batches each of size 128 (the same size as the training batch size). We calculated gradients on each of these pairs of mini-batches, and then calculated the cosine similarity between them. To measure the worse-case gradient confusion, we computed the lowest gradient cosine similarity among all pairs. We explored the effect of changing depth and changing width on the different activation functions and weight initializations. We plot the final training loss achieved for each model and the minimum gradient cosine similarities calculated over the 1000 pairs of gradients at the end of training. For each point, we plot both the mean and the standard deviation over the 10 independent runs.

The effect of depth. We first present results showing the effect of network depth. We considered a fixed width of $\ell = 100$, and varied the depth of the neural network, on the log scale, as:

$$\beta \in \{3, 10, 30, 100, 300, 1000\}.$$

Figure 5 shows results on neural networks with identity and tanh activation functions for the two weight initializations considered (Glorot normal and LeCun normal). Similar to the experimental results in section 7, and matching our theoretical results in sections 4 and 5, we notice the consistent trend of gradient confusion increasing with increasing depth. This makes the networks harder to train with increasing depth, and this is evidenced by an increase in the final training loss value. By depth $\beta = 1000$, the increased gradient confusion effectively makes the network untrainable when using tanh non-linearities.

The effect of width. We explored the effect of width by varying the width of the neural network while keeping the depth fixed at $\beta = 300$. We chose a very deep model, which is essentially untrainable for small widths (with standard initialization techniques) and helps better illustrate the effects of increasing width. We varied the width of the network, again on the log scale, as:

$$\ell \in \{10, 30, 100, 300, 1000\}.$$

Crucially, note that the smallest network considered here, MLP-300-10, still has more than 50000 parameters (i.e., more than the number of training samples), and the network with width $\ell = 30$ has almost three times the number of parameters as the high-performing MLP-3-100 network considered in the previous section. Figure 6 show results on linear neural networks and neural networks with tanh activations for both the Glorot normal and LeCun normal initializations. As in the experimental results of section 7, we see the consistent trend of gradient confusion decreasing with increasing width. Thus, wider networks become easier to train and improve the final training loss value. We further see that when the width is too small ($\ell = 30$), the gradient confusion becomes drastically high and the network becomes completely untrainable.

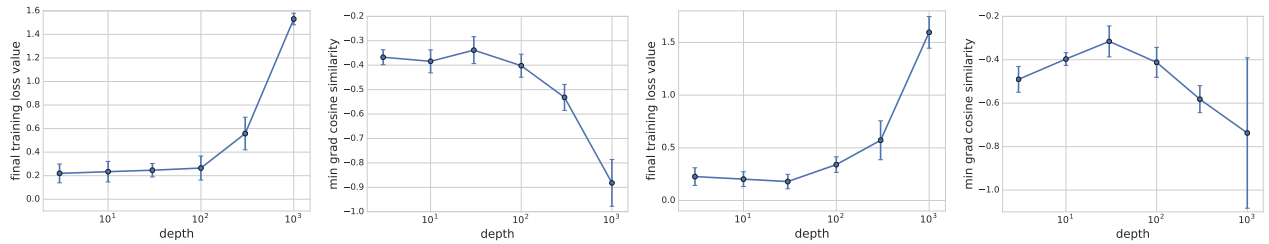
A.2. Additional experimental details for CNNs and WRNs

In this section, we review the details of our setup for the image classification experiments on CNNs and WRNs on the CIFAR-10 and CIFAR-100 datasets.

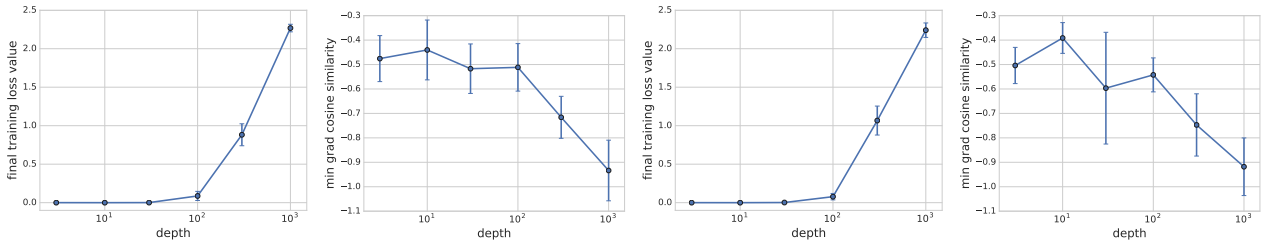
WIDE RESIDUAL NETWORKS

The Wide ResNet (WRN) architecture (Zagoruyko & Komodakis, 2016) for CIFAR datasets is a stack of three groups of residual blocks. There is a downsampling layer between two blocks, and the number of channels (width of a convolutional layer) is doubled after downsampling. In the three groups, the width of convolutional layers is $\{16\ell, 32\ell, 64\ell\}$, respectively. Each group contains β_r residual blocks, and each residual block contains two 3×3 convolutional layers equipped with ReLU activation, batch normalization and dropout. There is a 3×3 convolutional layer with 16 channels before the three groups of residual blocks. And there is a global average pooling, a fully-connected layer and a softmax layer after the three groups. The depth of WRN is $\beta = 6\beta_r + 4$.

For our experiments, we turned off dropout. Unless otherwise specified, we also turned off batch normalization. We added biases to the convolutional layers when not using batch normalization to maintain model expressivity. We used the MSRA initializer (He et al., 2015) for the weights as is standard for this model, and used the same preprocessing steps for the CIFAR images as described in Zagoruyko & Komodakis (2016). This preprocessing step involves normalizing the images and doing data augmentation (Zagoruyko & Komodakis, 2016). We denote this network as WRN-β-ℓ, where β represents the depth and ℓ represents the width factor of the network.

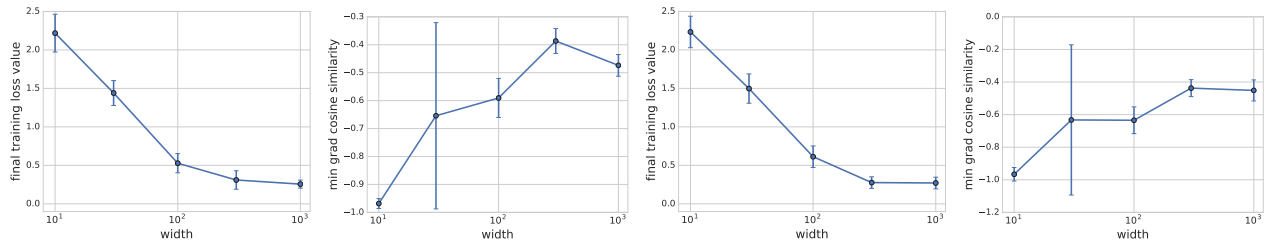


(a) Linear NN, Glorot init (b) Linear NN, Glorot init (c) Linear NN, LeCun init (d) Linear NN, LeCun init

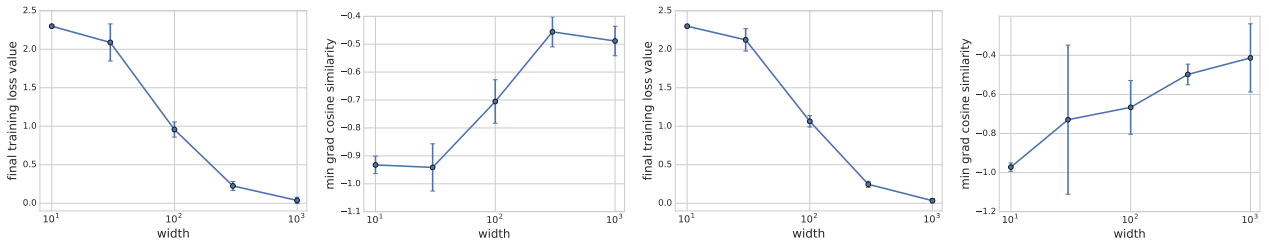


(e) Tanh NN, Glorot init (f) Tanh NN, Glorot init (g) Tanh NN, LeCun init (h) Tanh NN, LeCun init

Figure 5. Effect of varying depth on MLP-β-100.



(a) Linear NN, Glorot init (b) Linear NN, Glorot init (c) Linear NN, LeCun init (d) Linear NN, LeCun init



(e) Tanh NN, Glorot init (f) Tanh NN, Glorot init (g) Tanh NN, LeCun init (h) Tanh NN, LeCun init

Figure 6. Effect of varying width on MLP-300-ℓ.

To study the effect of depth, we considered WRNs with width factor $\ell = 2$ and depth varying as:

$$\beta \in \{16, 22, 28, 34, 40, 52, 76, 100\}.$$

For cleaner figures, we sometimes plot a subset of these results: $\beta \in \{16, 28, 40, 52, 76, 100\}$. To study the effect of width, we considered WRNs with depth $\beta = 16$ and width factor varying as:

$$\ell \in \{2, 3, 4, 5, 6\}.$$

CONVOLUTIONAL NEURAL NETWORKS

The WRN architecture contains skip connections that, as we show, help in training deep networks. To consider VGG-like convolutional networks, we consider a family of networks where we remove the skip connections from WRNs. Following the WRN convention, we denote these networks as CNN-β-ℓ, where β denotes the depth and ℓ denotes the width factor.

To study the effect of depth, we considered CNNs with width factor $\ell = 2$ and depth varying as:

$$\beta \in \{16, 22, 28, 34, 40\}.$$

To study the effect of width, we considered CNNs with depth $\beta = 16$ and width factor varying as:

$$\ell \in \{2, 3, 4, 5, 6\}.$$

HYPERPARAMETER TUNING AND OTHER DETAILS

We used SGD as the optimizer without any momentum. Following Zagoruyko & Komodakis (2016), we ran all experiments for 200 epochs with minibatches of size 128, and reduced the initial learning rate by a factor of 10 at epochs 80 and 160. We turned off weight decay for all our experiments.

We ran each individual experiment 5 times. We ignored any runs that were unable to decrease the loss from its initial value. We also made sure at least 4 out of the 5 independent runs ran till completion. When the learning rate is close to the threshold at which training is still possible, some runs may converge, while others may fail to converge. Thus, these checks ensure that we pick a learning rate that converges reliably in most cases on each problem. We show the standard deviation across runs in our plots.

We tuned the optimal initial learning rate for each model over a logarithmically-spaced grid:

$$\alpha \in \{10^1, 3 \times 10^0, 10^0, 3 \times 10^{-1}, 10^{-1}, 3 \times 10^{-2}, 10^{-2}, 3 \times 10^{-3}, 10^{-3}, 3 \times 10^{-4}, 10^{-4}, 3 \times 10^{-5}\},$$

and selected the run that achieved the lowest final training loss value (averaged over the independent runs). Our grid search was such that the optimal learning rate never occurred at one of the extreme values tested. We used the standard train-valid-test splits of 40000-10000-10000 for CIFAR-10 and CIFAR-100.

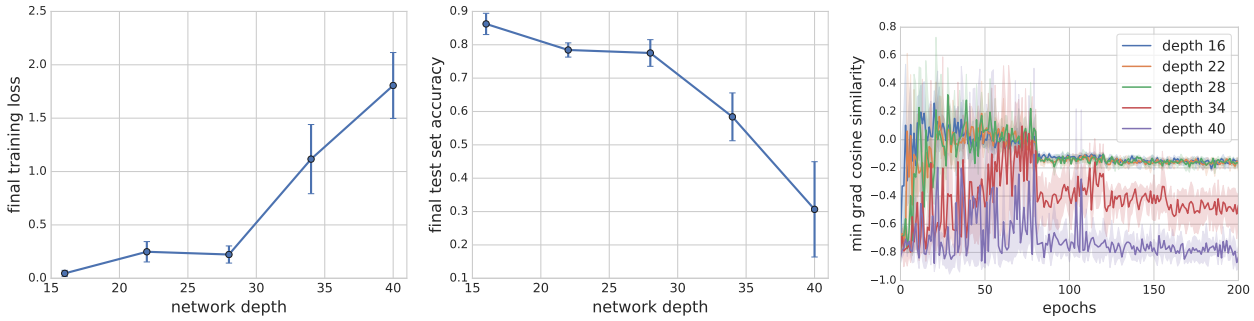
To measure gradient confusion, at the end of every training epoch, we sampled 100 pairs of mini-batches each of size 128 (the same size as the training batch size). We calculated gradients on each mini-batch, and then computed pairwise cosine similarities. To measure the worse-case gradient confusion, we computed the lowest gradient cosine similarity among all pairs. We also show the kernel density estimation of the pairwise gradient cosine similarities of the 100 minibatches sampled at the end of training (after 200 epochs), to see the concentration of the distribution. To do this, we combine together the 100 samples for each independent run and then perform kernel density estimation with a gaussian kernel on this data.

A.3. Additional plots for CIFAR-10 on CNNs

In section 7, we showed results for image classification using CNNs on CIFAR-10. In this section, we show some additional plots for this experiment. Figure 7 shows the effect of changing the depth, while figure 8 shows the effect of changing the width factor of the CNN. We see that the final training loss and test set accuracy values show the same trends as in section 7: deeper networks are harder to train, while wider networks are easier to train. As mentioned previously, theorems 3.1 and 3.2 indicate that we would expect the effect of gradient confusion to be more prominent near the end of training. From the plots we see that deeper networks have higher gradient confusion close to minimum, while wider networks have lower gradient confusion close to the minimum.

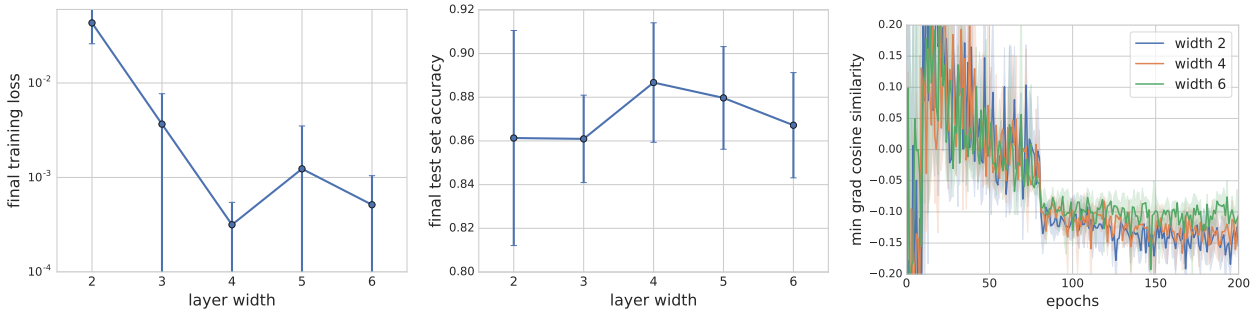
A.4. CIFAR-100 on CNNs

We now consider image classifications tasks with CNNs on the CIFAR-100 dataset. Figure 9 shows the effect of varying depth, while figure 10 shows the effect of varying width. We notice the same trends as in our results with CNNs on CIFAR-10. Interestingly, from the width results in figure 10, we see that while there is no perceptible change to the minimum pairwise gradient cosine similarity, the distribution still sharply concentrates around 0 with increasing width. Thus more gradients become orthogonal to each other with increasing width, implying that SGD on very wide networks becomes closer to decoupling over the data samples.



(a) (b) (c)

Figure 7. The effect of network depth with CNN-β-2 on CIFAR-10. The plots show the (a) final training loss values at the end of training, (b) final test set accuracy values at the end of training, and (c) the minimum of pairwise gradient cosine similarities during training.



(a) (b) (c)

Figure 8. The effect of width with CNN-16-ℓ on CIFAR-10. The plots show the (a) final training loss values at the end of training, (b) final test set accuracy values at the end of training, and the (c) minimum of pairwise gradient cosine similarities during training.

A.5. Image classification with WRNs on CIFAR-10 and CIFAR-100

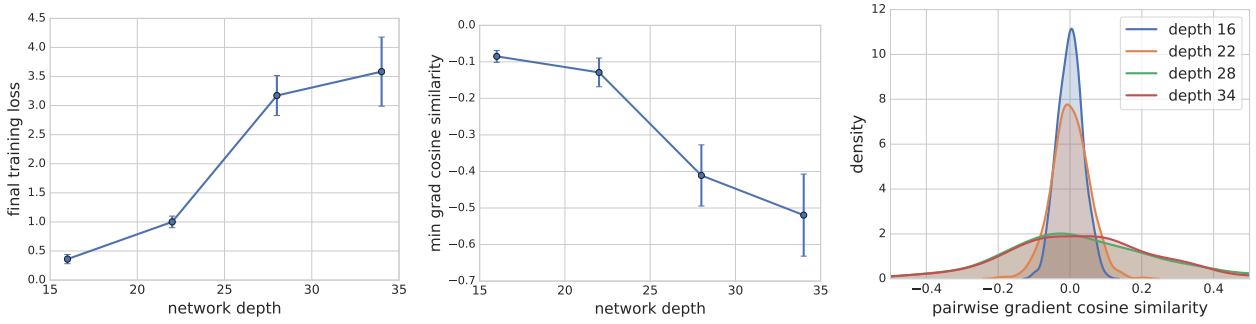
We now show results for image classification problems using wide residual networks (WRNs) on CIFAR-10 and CIFAR-100. The WRNs we consider do not have any batch normalization. Later we show results on the effect of adding batch normalization to these networks.

Figures 11 and 12 show results on the effect of depth using WRNs on CIFAR-10 and CIFAR-100 respectively. We again see the consistent trend of deeper networks having higher gradient confusion, making them harder to train. We further see that increasing depth results in the pairwise gradient cosine similarities concentrating less around 0.

Figures 13 and 14 show results on the effect of width using WRNs on CIFAR-10 and CIFAR-100 respectively. We see that increasing width typically lowers gradient confusion and helps the network achieve lower loss values. The pairwise gradient cosine similarities also typically concentrate around 0 with higher width. We also notice from these figures that in some cases, increasing width might lead to diminishing returns, i.e., the benefits of increased width diminish after a certain point, as one would expect.

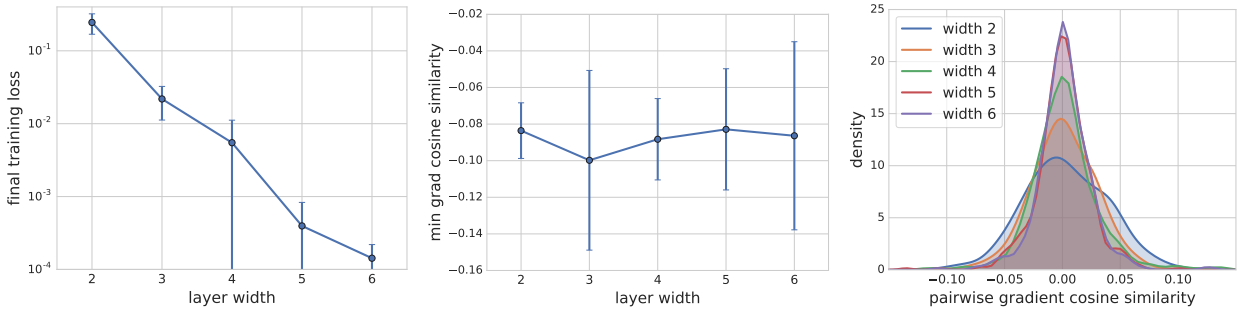
A.6. Effect of batch normalization and skip connections

In section 7 we showed results on the effect of adding batch normalization and skip connections to CNNs and WRNs on an image classification task on CIFAR-10. In this section, we present similar results for image classification on CIFAR-100. Similar to section 7, figure 15 shows that adding skip connections or batch normalization individually help in training deeper models, but these models still suffer from worsening results and increasing gradient confusion as the network gets deeper. Both these techniques together keep the gradient confusion relatively low even for very deep networks, significantly improving trainability of deep models.



(a) (b) (c)

Figure 9. The effect of network depth with CNN-β-2 on CIFAR-100. The plots show the (a) training loss values at the end of training, (b) minimum of pairwise gradient cosine similarities at the end of training, and the (c) kernel density estimate of the pairwise gradient cosine similarities at the end of training.



(a) (b) (c)

Figure 10. The effect of width with CNN-16-ℓ on CIFAR-100. The plots show the (a) training loss values at the end of training, (b) minimum of pairwise gradient cosine similarities at the end of training, and the (c) kernel density estimate of the pairwise gradient cosine similarities at the end of training.

B. Near orthogonality of random vectors

For completeness, we state and prove below a lemma on the near orthogonality of random vectors. This result is often attributed to Milman & Schechtman (1986).

Lemma B.1 (Near orthogonality of random vectors)**.** For vectors $\{\mathbf{x_i}\}_{i\in[N]}$ drawn uniformly from a unit sphere in d dimensions, and $\nu > 0$,

$$\Pr\left[\exists i, j \;\; |\mathbf{x}_i^\top \mathbf{x}_j| > \nu\right] \leq N^2 \sqrt{\tfrac{\pi}{8}} \exp\left(-\tfrac{d-1}{2}\nu^2\right).$$

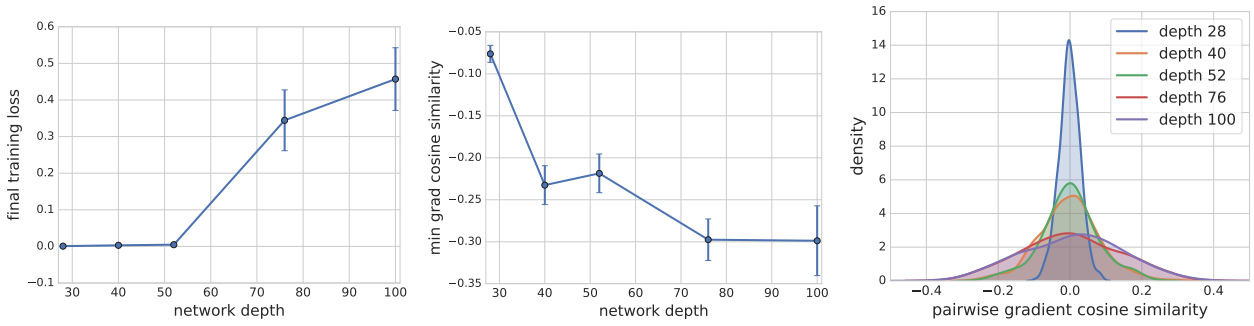
Proof. Given a fixed vector $\mathbf{x}$, a uniform random vector $\mathbf{y}$ satisfies $|\mathbf{x}^\top \mathbf{y}| \geq \nu$ only if $\mathbf{y}$ lies in one of two spherical caps: one centered at $\mathbf{x}$ and the other at $-\mathbf{x}$, and both with angular radius $\cos^{-1}(\nu) \leq \frac{\pi}{2} - \nu$. A simple result often attributed to Milman & Schechtman (1986) bounds the probability of lying in either of these caps as

$$\Pr[|\mathbf{x}^\top \mathbf{y}| \geq \nu] \leq \sqrt{\frac{\pi}{2}} \exp\left(-\frac{d-1}{2}\nu^2\right). \tag{5}$$

Because of rotational symmetry, the bound (5) holds if both $\mathbf{x}$ and $\mathbf{y}$ are chosen uniformly at random.

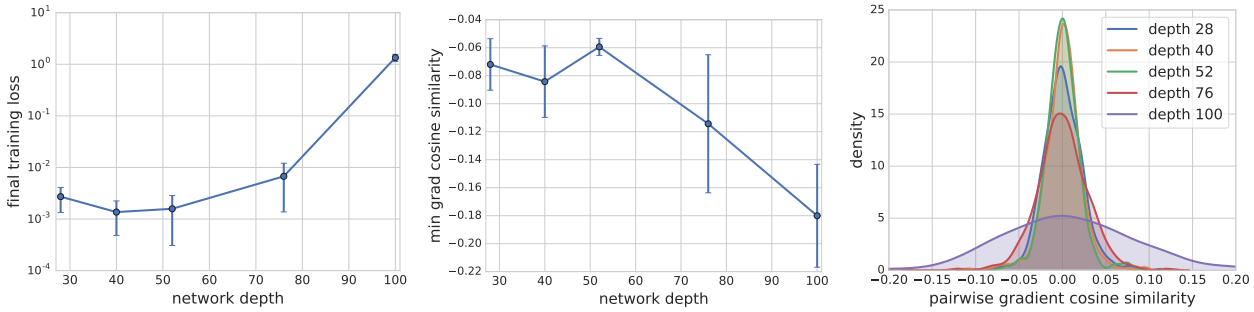
We next apply a union bound to control the probability that $|\mathbf{x}_i^\top \mathbf{x}_j| \geq \nu$ for some pair (i, j). There are fewer than $N^2/2$ such pairs, and so the probability of this condition is

$$\Pr[|\mathbf{x}_i^\top \mathbf{x}_j| \geq \nu, \text{for some } i, j] \leq \frac{N^2}{2} \sqrt{\frac{\pi}{2}} \exp\left(-\frac{d-1}{2}\nu^2\right). \qquad \square$$



(a) (b) (c)

Figure 11. The effect of depth with WRN-β-2 (no batch normalization) on CIFAR-10. The plots show the (a) training loss values at the end of training, (b) minimum of pairwise gradient cosine similarities at the end of training, and the (c) kernel density estimate of the pairwise gradient cosine similarities at the end of training.



(a) (b) (c)

Figure 12. The effect of depth with WRN-β-2 (no batch normalization) on CIFAR-100. The plots show the (a) training loss values at the end of training, (b) minimum of pairwise gradient cosine similarities at the end of training, and the (c) kernel density estimate of the pairwise gradient cosine similarities at the end of training.

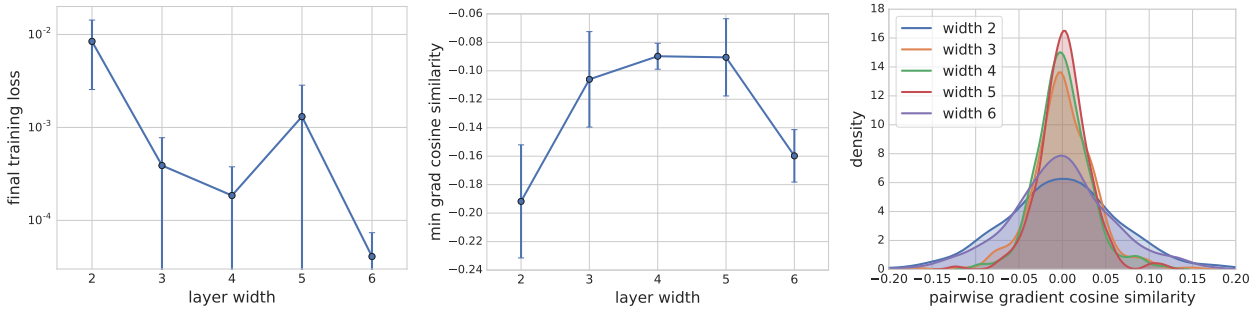
C. Low-rank Hessians lead to low gradient confusion

In this section, we show that low-rank random Hessians result in low gradient confusion. For clarity in presentation, suppose each f_i has a minimizer at the origin (the same argument can be easily extended to the more general case). Suppose also that there is a Lipschitz constant for the Hessian of each function f_i that satisfies $\|\mathbf{H}_i(\mathbf{w}) - \mathbf{H}_i(\mathbf{w}')\| \leq L_H\|\mathbf{w} - \mathbf{w}'\|$ (note that this is a standard optimization assumption (Nesterov, 2018), with evidence that it is applicable for neural networks (Martens, 2016)). Then $\nabla f_i(\mathbf{w}) = \mathbf{H}_i\mathbf{w} + \mathbf{e}$, where $\mathbf{e}$ is an error term bounded as: $\|\mathbf{e}\| \leq \frac{1}{2}L_H\|\mathbf{w}\|^2$, and we use the shorthand $\mathbf{H}_i$ to denote $\mathbf{H}_i(\mathbf{0})$. Then we have:

$$\begin{aligned}
|\langle \nabla f_i(\mathbf{w}), \nabla f_j(\mathbf{w})\rangle| &= |\langle \mathbf{H}_i\mathbf{w}, \mathbf{H}_j\mathbf{w}\rangle| + \langle \mathbf{e}, \mathbf{H}_i\mathbf{w} + \mathbf{H}_j\mathbf{w}\rangle + \|\mathbf{e}\|^2 \\
&\leq \|\mathbf{w}\|^2\|\mathbf{H}_i\|\|\mathbf{H}_j\| + \|\mathbf{e}\|\|\mathbf{w}\|(\|\mathbf{H}_i\| + \|\mathbf{H}_j\|) + \|\mathbf{e}\|^2 \\
&\leq \|\mathbf{w}\|^2\|\mathbf{H}_i\|\|\mathbf{H}_j\| + \frac{1}{2}L_H\|\mathbf{w}\|^3(\|\mathbf{H}_i\| + \|\mathbf{H}_j\|) + \frac{1}{4}L_H^2\|\mathbf{w}\|^4.
\end{aligned}$$

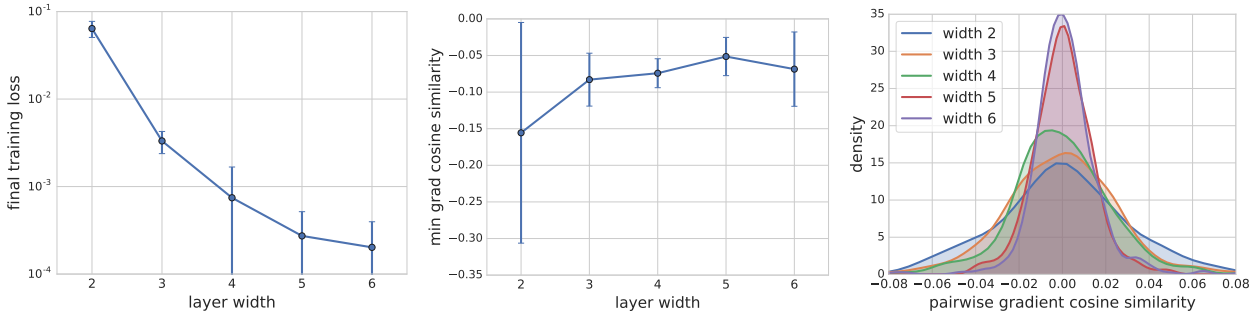
If the Hessians are sufficiently random and low-rank (e.g., of the form $\mathbf{H}_i = \mathbf{a}_i\mathbf{a}_i^\top$ where $\mathbf{a}_i \in \mathbb{R}^{N\times r}$ are randomly sampled from a unit sphere), then one would expect the terms in this expression to be small for all $\mathbf{w}$ within a neighborhood of the minimizer.

There is evidence that the Hessian at the minimizer is very low rank for many standard over-parameterized neural network models (Sagun et al., 2017; Cooper, 2018; Chaudhari et al., 2016; Wu et al., 2017; Ghorbani et al., 2019). While a bit non-rigorous, the above result nonetheless suggests that for many standard neural network models, the gradient confusion might be small for a large class of weights near the minimizer.



(a) (b) (c)

Figure 13. The effect of width with WRN-16-ℓ (no batch normalization) on CIFAR-10. The plots show the (a) training loss values at the end of training, (b) minimum of pairwise gradient cosine similarities at the end of training, and the (c) kernel density estimate of the pairwise gradient cosine similarities at the end of training.



(a) (b) (c)

Figure 14. The effect of width with WRN-16-ℓ (no batch normalization) on CIFAR-100. The plots show the (a) training loss values at the end of training, (b) minimum of pairwise gradient cosine similarities at the end of training, and the (c) kernel density estimate of the pairwise gradient cosine similarities at the end of training.

D. Missing proofs

D.1. Proofs of theorems 3.1 and 3.2

This section presents proofs for the convergence theorems of SGD presented in section 3, under the assumption of low gradient confusion. For clarity of presentation, we re-state each theorem before its proof.

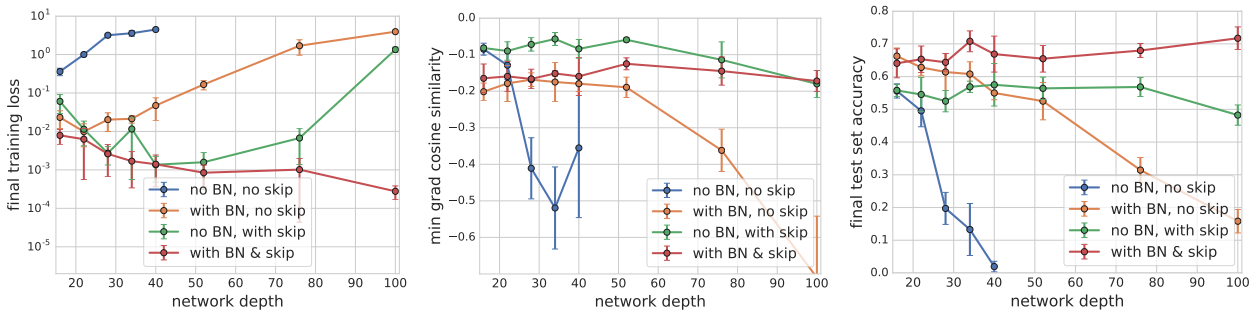
Theorem 3.1. *If the objective function satisfies (A1) and (A2), and has gradient confusion η, SGD converges linearly to a neighborhood of the minima of problem (1) as:*

$$\mathbb{E}[F(\mathbf{w}_T) - F^\star] \leq \rho^T (F(\mathbf{w}_0) - F^\star) + \tfrac{\alpha\eta}{1-\rho},$$

where $\alpha < \frac{2}{NL}$, $\rho = 1 - \frac{2\mu}{N}\left(\alpha - \frac{NL\alpha^2}{2}\right)$, $F^\star = \min_{\mathbf{w}} F(\mathbf{w})$ and $\mathbf{w}_0$ is the initialized weights.

Proof. Let $\tilde{i} \in [N]$ denote the index of the realized function $\tilde{f}_k$ in the uniform sampling from $\{f_i\}_{i\in[N]}$ at step k. From assumption (A1), we have

$$\begin{aligned} F(\mathbf{w}_{k+1}) &\leq F(\mathbf{w}_k) + \langle \nabla F(\mathbf{w}_k),\, \mathbf{w}_{k+1} - \mathbf{w}_k \rangle + \frac{L}{2}\|\mathbf{w}_{k+1} - \mathbf{w}_k\|^2 \\ &= F(\mathbf{w}_k) - \alpha \langle \nabla F(\mathbf{w}_k),\, \nabla \tilde{f}_k(\mathbf{w}_k) \rangle + \frac{L\alpha^2}{2}\|\nabla \tilde{f}_k(\mathbf{w}_k)\|^2 \\ &= F(\mathbf{w}_k) - \left(\frac{\alpha}{N} - \frac{L\alpha^2}{2}\right)\|\nabla \tilde{f}_k(\mathbf{w}_k)\|^2 - \frac{\alpha}{N} \sum_{\forall i : i \neq \tilde{i}} \langle \nabla f_i(\mathbf{w}_k),\, \nabla \tilde{f}_k(\mathbf{w}_k) \rangle \end{aligned}$$



(a) (b) (c)

Figure 15. The effect of adding skip connections and batch normalization to CNN-β-2 on CIFAR-100. Plots show the (a) training loss, (b) minimum pairwise gradient cosine similarities, and the (c) test accuracies at the end of training.

$$\leq F(\mathbf{w}_k) - \Big(\frac{\alpha}{N} - \frac{L\alpha^2}{2}\Big)\|\nabla \tilde{f}_k(\mathbf{w}_k)\|^2 + \frac{\alpha(N-1)\eta}{N},$$
$$\leq F(\mathbf{w}_k) - \Big(\frac{\alpha}{N} - \frac{L\alpha^2}{2}\Big)\|\nabla \tilde{f}_k(\mathbf{w}_k)\|^2 + \alpha\eta,$$

where the second-last inequality follows from definition 2.1. Let the learning rate $\alpha < 2/NL$. Then, using assumption (A2) and subtracting by $F^\star = \min_{\mathbf{w}} F(\mathbf{w})$ on both sides, we get

$$F(\mathbf{w}_{k+1}) - F^\star \leq F(\mathbf{w}_k) - F^\star - 2\mu\Big(\frac{\alpha}{N} - \frac{L\alpha^2}{2}\Big)(\tilde{f}_k(\mathbf{w}_k) - \tilde{f}_k^\star) + \alpha\eta,$$

where $\tilde{f}_k^\star = \min_{\mathbf{w}} \tilde{f}_k(\mathbf{w})$. It is easy to see that by definition we have, $\mathbb{E}_i[f_i^\star] \leq F^\star$. Moreover, from assumption that $\alpha < \frac{2}{NL}$, it implies that $\left(\frac{\alpha}{N} - \frac{L\alpha^2}{2}\right) > 0$. Therefore, taking expectation on both sides we get,

$$\mathbb{E}[F(\mathbf{w}_{k+1}) - F^\star] \leq \Big(1 - \frac{2\mu\alpha}{N} + \mu L\alpha^2\Big)\mathbb{E}[F(\mathbf{w}_k) - F^\star] + \alpha\eta.$$

Writing $\rho = 1 - \frac{2\mu\alpha}{N} + \mu L\alpha^2$, and unrolling the iterations, we get

$$\begin{aligned}\mathbb{E}[F(\mathbf{w}_{k+1}) - F^\star] &\leq \rho^{k+1}(F(\mathbf{w}_0) - F^\star) + \sum_{i=0}^{k}\rho^i\alpha\eta \\ &\leq \rho^{k+1}(F(\mathbf{w}_0) - F^\star) + \sum_{i=0}^{\infty}\rho^i\alpha\eta \\ &= \rho^{k+1}(F(\mathbf{w}_0) - F^\star) + \frac{\alpha\eta}{1-\rho}.\end{aligned}$$

□

Theorem 3.2. *If the objective satisfies (A1) and has gradient confusion η, then SGD converges to a neighborhood of a stationary point of problem (1) as:*

$$\min_{k=1,\ldots,T}\mathbb{E}\|\nabla F(\mathbf{w}_k)\|^2 \leq \frac{\rho(F(\mathbf{w}_1)-F^\star)}{T} + \rho\eta,$$

for $\alpha < \frac{2}{NL}$, $\rho = \frac{2N}{2-NL\alpha}$, and $F^\star = \min_{\mathbf{w}} F(\mathbf{w})$.

Proof. From theorem 3.1, we have:

$$F(\mathbf{w}_{k+1}) \leq F(\mathbf{w}_k) - \Big(\frac{\alpha}{N} - \frac{L\alpha^2}{2}\Big)\|\nabla \tilde{f}_k(\mathbf{w}_k)\|^2 + \alpha\eta. \tag{6}$$

Now we know that:

$$\mathbb{E}\|\nabla \tilde{f}_k(\mathbf{w}_k)\|^2 = \mathbb{E}\|\nabla \tilde{f}_k(\mathbf{w}_k) - \nabla F(\mathbf{w}_k)\|^2 + \mathbb{E}\|\nabla F(\mathbf{w}_k)\|^2 \geq \mathbb{E}\|\nabla F(\mathbf{w}_k)\|^2.$$

Thus, taking expectation and assuming the step size $\alpha < 2/(NL)$, we can rewrite equation 6 as:

$$\mathbb{E}\|\nabla F(\mathbf{w}_k)\|^2 \leq \frac{2N}{2\alpha - NL\alpha^2}\mathbb{E}[F(\mathbf{w}_k) - F(\mathbf{w}_{k+1})] + \frac{2N\eta}{2 - NL\alpha}.$$

Taking an average over T iterations, and using $F^\star = \min_{\mathbf{w}} F(\mathbf{w})$, we get:

$$\min_{k=1,\ldots,T} \mathbb{E}\|\nabla F(\mathbf{w}_k)\|^2 \leq \frac{1}{T}\sum_{k=1}^{\top} \mathbb{E}\|\nabla F(\mathbf{w}_k)\|^2 \leq \frac{2N}{2\alpha - NL\alpha^2}\frac{F(\mathbf{w}_1) - F^\star}{T} + \frac{2N\eta}{2 - NL\alpha}. \qquad \square$$

D.2. Proof of lemma D.1

Lemma D.1. *Consider the set of loss-functions $\{f_i(\mathbf{W})\}_{i\in[N]}$ where all f_i are either the square-loss function or the logistic-loss function. Recall that $f_i(\mathbf{W}) := f(\mathbf{W}, \mathbf{x}_i)$. Consider a feed-forward neural network as defined in equation 4 whose weights $\mathbf{W}$ satisfy assumption 1. Consider the gradient $\nabla_{\mathbf{W}} f_i(\mathbf{W})$ of each function f_i. From definition we have that $\nabla_{\mathbf{W}} f_i(\mathbf{W}) = \zeta_{\mathbf{x}_i}(\mathbf{W})\nabla_{\mathbf{W}} g_{\mathbf{W}}(\mathbf{x}_i)$, where we define $\zeta_{\mathbf{x}_i}(\mathbf{W}) = \partial f_i(\mathbf{W})/\partial g_{\mathbf{W}}$. Then we have the following properties.*

1. *When $\|\mathbf{x}\| \leq 1$ for every $p \in [\beta]$ we have $\|\nabla_{\mathbf{W}_p} g_{\mathbf{W}}(\mathbf{x}_i)\| \leq 1$.*
2. *There exists $0 < \zeta_0 \leq 2\sqrt{\beta}$, such that $|\zeta_{\mathbf{x}_i}(\mathbf{W})| \leq 2$, $\|\nabla_{\mathbf{x}_i}\zeta_{\mathbf{x}_i}(\mathbf{W})\|_2 \leq \zeta_0$, $\|\nabla_{\mathbf{W}}\zeta_{\mathbf{x}_i}(\mathbf{W})\|_2 \leq \zeta_0$.*

Proof. The first property is a direct consequence of assumption 1 and property (P2) of the activation function.

Let $\mathbf{W}$ denote the tuple $(\mathbf{W}_p)_{p\in[\beta]_0}$. Consider $|\zeta_{\mathbf{x}_i}(\mathbf{W})| = |\partial f_i(\mathbf{W})/\partial g_{\mathbf{W}}|$. In the case of square-loss function this evaluates to $|g_{\mathbf{W}}(\mathbf{x}) - \mathcal{C}(\mathbf{x})| \leq 2$. In case of logistic regression, this evaluates to $|\frac{-1}{1+\exp(\mathcal{C}(\mathbf{x}_i)g_{\mathbf{W}}(\mathbf{x}_i))}| \leq 1$. Now we consider $\|\nabla_{\mathbf{x}_i}\zeta_{\mathbf{x}_i}(\mathbf{W})\|$. Consider the squared loss function. We then have the following.

$$\begin{aligned}\|\nabla_{\mathbf{x}_i}\zeta_{\mathbf{x}_i}(\mathbf{W})\| &= \|\nabla_{\mathbf{x}_i} f'(\mathbf{W})\| \\ &= \|\nabla_{\mathbf{x}_i} g_{\mathbf{W}}(\mathbf{x}_i) - \mathcal{C}(\mathbf{x}_i)\| \\ &\leq \|\nabla_{\mathbf{x}_i} g_{\mathbf{W}}(\mathbf{x}_i)\| + 1.\end{aligned}$$

Likewise, consider the logistic-loss function. We then have the following.

$$\begin{aligned}\|\nabla_{\mathbf{x}_i}\zeta_{\mathbf{x}_i}(\mathbf{W})\| &\leq \left\| \frac{\mathcal{C}(\mathbf{x}_i)^2}{(1+\exp(\mathcal{C}(\mathbf{x}_i)g_{\mathbf{W}}(\mathbf{x}_i)))^2}\exp(\mathcal{C}(\mathbf{x}_i)g_{\mathbf{W}}(\mathbf{x}_i)) \right\| \|\nabla_{\mathbf{x}_i} g_{\mathbf{W}}(\mathbf{x}_i)\| \\ &\leq \|\nabla_{\mathbf{x}_i} g_{\mathbf{W}}(\mathbf{x}_i)\|.\end{aligned}$$

Thus, it suffices to bound $\|\nabla_{\mathbf{x}_i} g_{\mathbf{W}}(\mathbf{x}_i)\|$. Using assumption 1 and the properties (P1), (P2) of σ, this can be upper-bounded by 1.

Consider $\nabla_{\mathbf{W}_p}\zeta_{\mathbf{x}_i}(\mathbf{W})$ for some layer index $p \in [\beta]_0$. We will show that $\|\nabla_{\mathbf{W}_p}\zeta_{\mathbf{x}_i}(\mathbf{W})\|_2 \leq 2$. Then it immediately follows that $\|\nabla_{\mathbf{W}}\zeta_{\mathbf{x}_i}(\mathbf{W})\|_2 \leq 2\sqrt{\beta}$. In the case of a squared loss function. We have the following.

$$\begin{aligned}\|\nabla_{\mathbf{W}_p}\zeta_{\mathbf{x}_i}(\mathbf{W})\| &= \|\nabla_{\mathbf{W}_p} f'(\mathbf{W})\| \\ &= \|\nabla_{\mathbf{W}_p} g_{\mathbf{W}}(\mathbf{x}_i) - \mathcal{C}(\mathbf{x}_i)\| \\ &\leq \|\nabla_{\mathbf{W}_p} g_{\mathbf{W}}(\mathbf{x}_i)\| + 1.\end{aligned}$$

Likewise, consider the logistic-loss function. We then have the following.

$$\begin{aligned}\|\nabla_{\mathbf{W}_p}\zeta_{\mathbf{x}_i}(\mathbf{W})\| &\leq \left\| \frac{\mathcal{C}(\mathbf{x}_i)^2}{(1+\exp(\mathcal{C}(\mathbf{x}_i)g_{\mathbf{W}}(\mathbf{x}_i)))^2}\exp(\mathcal{C}(\mathbf{x}_i)g_{\mathbf{W}}(\mathbf{x}_i)) \right\| \|\nabla_{\mathbf{W}_p} g_{\mathbf{W}}(\mathbf{x}_i)\| \\ &\leq \|\nabla_{\mathbf{W}_p} g_{\mathbf{W}}(\mathbf{x}_i)\|.\end{aligned}$$

Since $\|\nabla_{\mathbf{W}_p} g_{\mathbf{W}}(\mathbf{x}_i)\| \leq 1$, we have that $\|\nabla_{\mathbf{W}_p}\zeta_{\mathbf{x}_i}(\mathbf{W})\| \leq 2$ in both the cases. Thus, $\zeta_0 = 2\sqrt{\beta}$. $\square$

D.3. Proofs of theorem 5.1 and corollary 5.1

In this section, we will present the proofs of theorem 5.1 and corollary 5.1.

Theorem 5.1. *Let $\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_\beta$ satisfy assumption 1. For some fixed constant $c > 0$, the gradient confusion bound (equation 3) holds with probability at least*

$$1 - N^2 \exp\left(\frac{-cd\eta^2}{16\zeta_0^4(\beta+2)^4}\right).$$

Proof. We show two key properties, namely bounded gradient and non negative expectation. We will then use both these properties to complete the proof.

Bounded gradient. For every $i \in [n]$ define $\zeta_{\mathbf{x}_i}(\mathbf{W}) := f'(\mathbf{W})$. For every $p \in [\beta]$ define $\mathbf{H}_p$ as follows.

$$\mathbf{H}_p(\mathbf{x}) := \sigma(\mathbf{W}_p \cdot \sigma(\mathbf{W}_{p-1} \cdot \sigma(\ldots \cdot \sigma(\mathbf{W}_0 \cdot \mathbf{x})\ldots).$$

Fix an $i \in [N]$. Then we have the following recurrence

$$\begin{aligned} &g_\beta(\mathbf{x}_i) := \sigma'(H_\beta(\mathbf{x}_i)) \\ &\mathbf{g}_p(\mathbf{x}_i) := (\mathbf{W}_{p+1}^\top \cdot \mathbf{g}_{p+1}(\mathbf{x}_i)) \cdot \mathrm{Diag}(\sigma'(\mathbf{H}_p(\mathbf{x}_i))) \qquad \forall p \in \{0, 1, \ldots, \beta-1\}. \end{aligned}$$

Then the gradients can be written in terms of the above quantities as follows.

$$\nabla_{\mathbf{W}_p} f_i(\mathbf{W}) = \mathbf{g}_p(\mathbf{x}_i) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)^\top \qquad \forall p \in [\beta]_0.$$

We can write, the gradient confusion denote by $h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)$, as follows.

$$\zeta_{\mathbf{x}_i}(\mathbf{W})\zeta_{\mathbf{x}_j}(\mathbf{W}) \left(\sum_{p \in [\beta]_0} \mathrm{Tr}[\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)^\top] \right). \tag{7}$$

We will now bound $\|\nabla_{(\mathbf{x}_i, \mathbf{x}_j)} h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)\|_2$. Consider $\nabla_{\mathbf{x}_i} h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)$. This can be written as follows.

$$\begin{aligned} (\nabla_{\mathbf{x}_i}\zeta_{\mathbf{x}_i}(\mathbf{W}))\zeta_{\mathbf{x}_j}(\mathbf{W}) \left(\sum_{p \in [\beta]_0} \mathrm{Tr}[\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)^\top] \right) + \\ \zeta_{\mathbf{x}_i}(\mathbf{W})\zeta_{\mathbf{x}_j}(\mathbf{W}) \sum_{p \in [\beta]_0} \left[\nabla_{\mathbf{x}_i} \left(\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)\right)\right]^\top. \end{aligned} \tag{8}$$

Observe that each of the entries in the diagonal matrix $\mathrm{Diag}(\sigma'(\mathbf{H}_p(\mathbf{x}_i)))$ is at most 1. Thus, we have that $\|\mathrm{Diag}(\sigma'(\mathbf{H}_p(\mathbf{x}_i)))\| \leq 1$.

We have the following relationship.

$$\begin{aligned} &\|g_\beta(\mathbf{x}_i)\| \leq 1 \\ &\|\mathbf{g}_p(\mathbf{x}_i)\| \leq \|\mathbf{W}_{p+1}^\top\| \|\mathbf{g}_{p+1}(\mathbf{x}_i))\| \|\mathrm{Diag}(\sigma'(\mathbf{H}_p(\mathbf{x}_i)))\| \leq 1 \qquad \forall p \in \{0, 1, \ldots, \beta-1\}. \end{aligned}$$

Moreover we have,

$$\|\mathrm{Tr}[\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)^\top]\| \leq \|\mathbf{H}_{p-1}(\mathbf{x}_i)\| \|\mathbf{g}_p(\mathbf{x}_i)^\top\| \|\mathbf{g}_p(\mathbf{x}_j)\| \|\mathbf{H}_{p-1}(\mathbf{x}_i)^\top\| \leq 1.$$

Consider $\|\nabla_{\mathbf{x}_i} \left(\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)\right)\|$ for every $p \in [\beta]_0$.

This can be upper-bounded by,

$$\|\nabla_{\mathbf{x}_i} \mathbf{H}_{p-1}(\mathbf{x}_i)\| \|\mathbf{g}_p(\mathbf{x}_i)^\top\| \|\mathbf{g}_p(\mathbf{x}_j)\| \|\mathbf{H}_{p-1}(\mathbf{x}_i)\| + \|\mathbf{H}_{p-1}(\mathbf{x}_i)\| \|\nabla_{\mathbf{x}_i} \mathbf{g}_p(\mathbf{x}_i)^\top\| \|\mathbf{g}_p(\mathbf{x}_j)\| \|\mathbf{H}_{p-1}(\mathbf{x}_i)\|.$$

Note that $\nabla_{\mathbf{x}_i}\mathbf{H}_{p-1}(\mathbf{x}_i) = \mathbf{g}_1(\mathbf{x}_i)\cdot \mathrm{Diag}(\sigma'(\mathbf{W}_0\cdot\mathbf{x}_i))\cdot\mathbf{W}_0^\top\cdot\mathbf{g}_p(\mathbf{x}_i)^\top$. Thus, $\|\nabla_{\mathbf{x}_i}\mathbf{H}_{p-1}(\mathbf{x}_i)\|\leq 1$. We will now show that $\|\nabla_{\mathbf{x}_i}\mathbf{g}_p(\mathbf{x}_i)\|\leq \beta - p + 1$. We prove this inductively. Consider the base case when $p=\beta$.

$$\|\nabla_{\mathbf{x}_i}\mathbf{g}_\beta(\mathbf{x}_i)\| = \|\nabla_{\mathbf{x}_i}\sigma'(\mathbf{H}_\beta(\mathbf{x}_i))\|\leq 1 = \beta-\beta+1.$$

Now, the inductive step.

$$\|\nabla_{\mathbf{x}_i}\mathbf{g}_p(\mathbf{x}_i)\|\leq \|\nabla_{\mathbf{x}_i}\mathbf{g}_{p+1}(\mathbf{x}_i)\| + \|\nabla_{\mathbf{x}_i}\mathrm{Diag}(\sigma'(\mathbf{H}_p(\mathbf{x}_i)))\|\leq \beta - p\leq \beta - p + 1.$$

Thus, using equation 8 and the above arguments, we obtain, $\|\nabla_{\mathbf{x}_i}h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\|_2\leq \zeta_0^2(\beta+1)+\zeta_0^2(\beta+1)(\beta+2)\leq 2\zeta_0^2(\beta+2)^2$ and thus, $\|\nabla_{(\mathbf{x}_i,\mathbf{x}_j)}h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\|_2\leq 4\zeta_0^2(\beta+2)^2$.

Non-negative expectation.

$$\begin{aligned}\mathbb{E}_{\mathbf{x}_i,\mathbf{x}_j}[h(\mathbf{x}_i,\mathbf{x}_j)] =& \mathbb{E}_{\mathbf{x}_i,\mathbf{x}_j}[\langle\nabla f_i(\mathbf{W}),\nabla f_j(\mathbf{W})\rangle]\\ &= \langle\mathbb{E}_{\mathbf{x}_i}[\nabla f_i(\mathbf{W})],\mathbb{E}_{\mathbf{x}_j}[\nabla f_j(\mathbf{W})]\rangle\\ &= \|\mathbb{E}_{\mathbf{x}_i}[\nabla f_i(\mathbf{W})]\|^2\geq 0.\end{aligned}\tag{9}$$

We have used the fact that $\nabla f_i(\mathbf{W})$ and $\nabla f_j(\mathbf{W})$ are identically distributed and independent.

Concentration of Measure. We combine the two properties as follows. From **Non-negative Expectation** property and equation 26, we have that

$$\Pr[h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\leq -\eta]\leq \Pr[h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\leq \mathbb{E}_{(\mathbf{x}_i,\mathbf{x}_j)}[h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)]-\eta]\leq \exp\left(\frac{-cd\eta^2}{16\zeta_0^4(\beta+2)^4}\right).\tag{10}$$

To obtain the probability that *some* value of $h_{\mathbf{w}}(\nabla_{\mathbf{w}}f_i,\nabla_{\mathbf{w}}f_j)$ lies below $-\eta$, we use a union bound. There are $N(N-1)/2 < N^2/2$ possible pairs of data points to consider, and so this probability is bounded above by $N^2\exp\left(\frac{-cd\eta^2}{16\zeta_0^4(\beta+2)^4}\right)$. □

D.3.1. Proof of corollary 5.1

Before we prove corollary 5.1 we first prove the following helper lemma.

Lemma D.2. *Suppose* $\max_{\mathbf{W}}\|\nabla_{\mathbf{W}}f_i(\mathbf{W})\|\leq M$, *and both* $\nabla_{\mathbf{W}}f_i(\mathbf{w})$ *and* $\nabla_{\mathbf{W}}f_j(\mathbf{W})$ *are Lipschitz in* $\mathbf{W}$ *with constant* L. *Then* $h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)$ *is Lipschitz in* $\mathbf{W}$ *with constant* $2LM$.

Proof. We view $\mathbf{W}$ as flattened vector. We now prove the above result for these two vectors. For two vectors $\mathbf{w},\mathbf{w}'$,

$$\begin{aligned}&|h_{\mathbf{w}}(\mathbf{x}_i,\mathbf{x}_j)-h_{\mathbf{w}'}(\mathbf{x}_i,\mathbf{x}_j)|\\ &= |\langle\nabla_{\mathbf{w}}f_i(\mathbf{w}),\nabla_{\mathbf{w}}f_j(\mathbf{w})\rangle - \langle\nabla_{\mathbf{w}'}f_i(\mathbf{w}'),\nabla_{\mathbf{w}'}f_j(\mathbf{w}')\rangle|\\ &= |\langle\nabla_{\mathbf{w}}f_i(\mathbf{w})-\nabla_{\mathbf{w}'}f_i(\mathbf{w}')+\nabla_{\mathbf{w}'}f_i(\mathbf{w}'),\nabla_{\mathbf{w}}f_j(\mathbf{w})\rangle\\ &\quad -\langle\nabla_{\mathbf{w}'}f_i(\mathbf{w}'),\nabla_{\mathbf{w}'}f_j(\mathbf{w}')-\nabla_{\mathbf{w}}f_j(\mathbf{w})+\nabla_{\mathbf{w}}f_j(\mathbf{w})\rangle|\\ &= |\langle\nabla_{\mathbf{w}}f_i(\mathbf{w})-\nabla_{\mathbf{w}'}f_i(\mathbf{w}'),\nabla_{\mathbf{w}}f_j(\mathbf{w})\rangle - \langle\nabla_{\mathbf{w}'}f_i(\mathbf{w}'),\nabla_{\mathbf{w}'}f_j(\mathbf{w}')-\nabla_{\mathbf{w}}f_j(\mathbf{w})\rangle|\\ &\leq |\langle\nabla_{\mathbf{w}}f_i(\mathbf{w})-\nabla_{\mathbf{w}'}f_i(\mathbf{w}'),\nabla_{\mathbf{w}}f_j(\mathbf{w})\rangle| + |\langle\nabla_{\mathbf{w}'}f_i(\mathbf{w}'),\nabla_{\mathbf{w}'}f_j(\mathbf{w}')-\nabla_{\mathbf{w}}f_j(\mathbf{w})\rangle|\\ &\leq \|\nabla_{\mathbf{w}}f_i(\mathbf{w})-\nabla_{\mathbf{w}'}f_i(\mathbf{w}')\|\|\nabla_{\mathbf{w}}f_j(\mathbf{w})\| + \|\nabla_{\mathbf{w}'}f_i(\mathbf{w}')\|\|\nabla_{\mathbf{w}'}f_j(\mathbf{w}')-\nabla_{\mathbf{w}}f_j(\mathbf{w})\|\\ &\leq L\|\mathbf{w}-\mathbf{w}'\|\|\nabla_{\mathbf{w}}f_j(\mathbf{w})\| + \|\nabla_{\mathbf{w}'}f_i(\mathbf{w}')\|L\|\mathbf{w}'-\mathbf{w}\|\\ &\leq 2LM\|\mathbf{w}-\mathbf{w}'\|.\end{aligned}$$

Here the first inequality uses the triangle inequality, the second inequality uses the Cauchy-Schwartz inequality, and the third and fourth inequalities use the assumptions that $\nabla_{\mathbf{w}}f_i(\mathbf{w})$ and $\nabla_{\mathbf{w}}f_j(\mathbf{w})$ are Lipschitz in $\mathbf{w}$ and have bounded norm. □

We are now ready to prove the corollary, which we restate here. The proof uses a standard "epsilon-net" argument; we identify a fine net of points within the ball $\mathcal{B}_r$. If the gradient confusion is small at every point in this discrete set, and the gradient confusion varies slowly enough with $\mathbf{W}$, when we can guarantee small gradient confusion at every point in $\mathcal{B}_r$.

Corollary 5.1. *Select a point $\mathbf{W} = (\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_\beta)$, satisfying assumption 1. Consider a ball $\mathcal{B}_r$ centered at $\mathbf{W}$ of radius $r > 0$. If the data $\{\mathbf{x}_i\}_{i\in[N]}$ are sampled uniformly from a unit sphere, then the gradient confusion bound in equation 3 holds uniformly at all points $\mathbf{W}' \in \mathcal{B}_r$ with probability at least*

$$\begin{aligned} &1 - N^2 \exp\left(-\tfrac{cd\eta^2}{64\zeta_0^4(\beta+2)^4}\right), && \text{if } r \le \eta/4\zeta_0^2, \\ &1 - N^2 \exp\left(-\tfrac{cd\eta^2}{64\zeta_0^4(\beta+2)^4} + \tfrac{8d\zeta_0^2 r}{\eta}\right), && \text{otherwise.} \end{aligned}$$

Proof. Define the function $h^+(\mathbf{W}) = \max_{ij} h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)$. Our goal is to find conditions under which $h^+(\mathbf{W}) > -\eta$ for all $\mathbf{W}$ in a large set. To derive such conditions, we will need a Lipschitz constant for $h^+(\mathbf{W})$, which is no larger than the maximal Lipschitz constant of $h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)$ for all i, j. We have that $\|\nabla_{\mathbf{W}} f_i\| = \|\zeta_{\mathbf{x}_i}(\mathbf{W})\mathbf{x}_i\| \le \zeta_0$. Now we need to get a $\mathbf{W}$-Lipschitz constants for $\nabla_{\mathbf{x}_i} f_i = \zeta_{\mathbf{x}_i}(\mathbf{W})\mathbf{x}_i$. By lemma D.1, we have $\|\nabla_{\mathbf{W}}(\zeta_{\mathbf{x}_i}(\mathbf{W})\mathbf{x}_i)\| = \|(\nabla_{\mathbf{W}}\zeta_{\mathbf{x}_i}(\mathbf{W}))\mathbf{x}_i\| \le \zeta_0$. Using lemma D.2, we see that $2\zeta_0^2$ is a Lipschitz constant for $h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)$, and thus also $h^+(\mathbf{W})$.

Now, consider a minimizer $\mathbf{W}$ of the objective, and a ball $\mathcal{B}_r$ around this point of radius r. Define the constant $\epsilon = \frac{\eta}{4\zeta_0^2}$, and create an ϵ-net of points $\mathcal{N}_\epsilon = \{\mathbf{W}_i\}$ inside the ball. This net is sufficiently dense that any $\mathbf{W}' \in \mathcal{B}_r$ is at most ϵ units away from some $\mathbf{W}_i \in \mathcal{N}_\epsilon$. Furthermore, because $h^+(\mathbf{W})$ is Lipschitz in $\mathbf{W}$, $|h^+(\mathbf{W}') - h^+(\mathbf{W}_i)| \le 2\zeta_0^2\epsilon = \eta/2$.

We now know the following: if we can guarantee that

$$h^+(\mathbf{W}_i) \ge -\eta/2, \text{ for all } \mathbf{W}_i \in \mathcal{N}_\epsilon, \tag{11}$$

then we also know that $h^+(\mathbf{W}') \ge -\eta$ for all $\mathbf{W}' \in \mathcal{B}_r$. For this reason, we prove the result by bounding the probability that (11) holds. It is known that $\mathcal{N}_\epsilon$ can be constructed so that $|\mathcal{N}_\epsilon| \le (2r/\epsilon + 1)^d = (8\zeta_0^2 r/\eta + 1)^d$ (see Vershynin (2018), corollary 4.1.13). Theorem 5.1 provides a bound on the probability that each individual point in the net satisfies condition (11). Using a union bound, we see that all points in the net satisfy this condition with probability at least

$$1 - N^2\left(\frac{8\zeta_0^2 r}{\eta} + 1\right)^d \exp\left(-\frac{cd(\eta/2)^2}{16\zeta_0^4}\right) \tag{12}$$

$$= 1 - N^2 \exp(d\log(8\zeta_0^2 r/\eta + 1)) \exp\left(-\frac{cd\eta^2}{64\zeta_0^4}\right) \tag{13}$$

$$\ge 1 - N^2 \exp(8d\zeta_0^2 r/\eta) \exp\left(-\frac{cd\eta^2}{64\zeta_0^4}\right) \tag{14}$$

$$= 1 - N^2 \exp\left(-\frac{cd\eta^2}{64\zeta_0^4} + \frac{8d\zeta_0^2 r}{\eta}\right). \tag{15}$$

Finally, note that, if $r < \epsilon$, then we can form a net with $|\mathcal{N}_\epsilon| = 1$. In this case, the probability of satisfying (11) is at least

$$1 - N^2 \exp\left(-\frac{cd(\eta/2)^2}{64\zeta_0^4}\right). \qquad \square$$

D.4. Proof of theorem 4.1

Theorem 4.1. *Let $\mathbf{W}_0, \mathbf{W}_1, \ldots, \mathbf{W}_\beta$ be weight matrices chosen according to strategy 4.1. There exists fixed constants $c_1, c_2 > 0$ such that we have the following.*

1. *Consider a fixed but arbitrary dataset $\mathbf{x}_1, \mathbf{x}_2, \ldots, \mathbf{x}_N$ with $\|\mathbf{x}_i\| \le 1$ for every $i \in [N]$. For $\eta > 4$, the gradient confusion bound in equation 3 holds with probability at least*

$$1 - \beta \exp\left(-c_1\kappa^2\ell^2\right) - N^2 \exp\left(\tfrac{-c\ell^2\beta(\eta-4)^2}{64\zeta_0^4(\beta+2)^4}\right).$$

2. *If the dataset $\{\mathbf{x}_i\}_{i\in[N]}$ is such that each $\mathbf{x}_i$ is an i.i.d. sample from the surface of d-dimensional unit sphere, then for every $\eta > 0$ the gradient confusion bound in equation 3 holds with probability at least*

$$1 - \beta \exp\left(-c_1\kappa^2\ell^2\right) - N^2 \exp\left(\tfrac{-c_2(\ell d + \ell^2\beta)\eta^2}{16\zeta_0^4(\beta+2)^4}\right).$$

Both parts in theorem 4.1 depend on the following argument. From theorem 2.3.8 and Proposition 2.3.10 in Tao (2012) with appropriate scaling[7], we have for every $p = 1, \ldots, \beta$ we have that the matrix norm $\|\mathbf{W}_p\| \leq 1$ with probability at least $1 - \beta \exp\left(-c_1 \kappa^2 \ell^2\right)$ and $\|\mathbf{W}_0\| \leq 1$ with probability at least $1 - \exp\left(-c_1 \kappa^2 d^2\right)$ when the weight matrices are initialized according to strategy 4.1. Thus, conditioning on this event it implies that these matrices satisfy assumption 1. The proof strategy is similar to that of theorem 5.1. We will first show that the gradient of the function $h(.,.)$ as defined in equation (7) *with respect to the weights* is bounded. Note that in part (1) the random variable is the set of weight matrices $\{W_p\}_{p\in[\beta]}$. Thus, the dimension used to invoke theorem E.1 is at most $\ell^2\beta$. In part (2) along with the weights, the data $\mathbf{x} \in \mathbb{R}^d$ is also random. Thus, the dimension used to invoke theorem E.1 is at most $\ell d + \ell^2\beta$. Combining this with theorem E.1, the bound on the gradient of $h(.,.)$ and taking a union bound, we get the respective parts of the theorem. Thus, all it remains to prove is the bound on the gradient of the function $h(.,.)$ as defined in equation (7) *with respect to the weights* conditioning on the event that $\|\mathbf{W}_p\| \leq 1$ for every $p \in \{0, 1, \ldots, \beta\}$.

We obtain the following analogue of equation (8).

$$
\begin{aligned}
(\nabla_{\mathbf{W}} \zeta_{\mathbf{x}_i}(\mathbf{W}))\zeta_{\mathbf{x}_j}(\mathbf{W}) \left(\sum_{p\in[\beta]_0} \mathrm{Tr}[\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)^\top] \right) + \\
(\nabla_{\mathbf{W}} \zeta_{\mathbf{x}_j}(\mathbf{W}))\zeta_{\mathbf{x}_i}(\mathbf{W}) \left(\sum_{p\in[\beta]_0} \mathrm{Tr}[\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)^\top] \right) + \\
\zeta_{\mathbf{x}_i}(\mathbf{W})\zeta_{\mathbf{x}_j}(\mathbf{W}) \sum_{p\in[\beta]_0} \left[\nabla_{\mathbf{W}} \left(\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)\right)\right]^\top . \quad (16)
\end{aligned}
$$

As in the case of the proof for theorem 5.1, we will upper-bound the ℓ_2-norm of the above expression. In particular, we show the following.

$$
\left\| (\nabla_{\mathbf{W}} \zeta_{\mathbf{x}_i}(\mathbf{W}))\zeta_{\mathbf{x}_j}(\mathbf{W}) \left(\sum_{p\in[\beta]_0} \mathrm{Tr}[\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)^\top] \right) \right\|_2 \leq 2\zeta_0^2(\beta+2)^2. \quad (17)
$$

$$
\left\| (\nabla_{\mathbf{W}} \zeta_{\mathbf{x}_j}(\mathbf{W}))\zeta_{\mathbf{x}_i}(\mathbf{W}) \left(\sum_{p\in[\beta]_0} \mathrm{Tr}[\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)^\top] \right) \right\|_2 \leq 2\zeta_0^2(\beta+2)^2. \quad (18)
$$

$$
\left\| \zeta_{\mathbf{x}_i}(\mathbf{W})\zeta_{\mathbf{x}_j}(\mathbf{W}) \sum_{p\in[\beta]_0} \left[\nabla_{\mathbf{W}} \left(\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)\right)\right]^\top \right\|_2 \leq 4\zeta_0^2(\beta+2)^2. \quad (19)
$$

Equations (17) and 18 follow from the the fact that $\|(\nabla_{\mathbf{W}} \zeta_{\mathbf{x}_i}(\mathbf{W}))\|_2 \leq \zeta_0$ and the arguments in the proof for theorem 5.1. We will now show the proof sketch for equation (19). For every $p \in [\beta]_0$, consider $\|\nabla_{\mathbf{W}} \left(\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)\right) \|$. Using the symmetry between $\mathbf{x}_i$ and $\mathbf{x}_j$, the expression can be upper-bounded by,

$$
2\|\nabla_{\mathbf{W}} \mathbf{H}_{p-1}(\mathbf{x}_i)\|\|\mathbf{g}_p(\mathbf{x}_i)^\top\|\|\mathbf{g}_p(\mathbf{x}_j)\|\|\mathbf{H}_{p-1}(\mathbf{x}_i)\| + 2\|\mathbf{H}_{p-1}(\mathbf{x}_i)\|\|\nabla_{\mathbf{W}} \mathbf{g}_p(\mathbf{x}_i)^\top\|\|\mathbf{g}_p(\mathbf{x}_j)\|\|\mathbf{H}_{p-1}(\mathbf{x}_i)\|.
$$

As before we can use an inductive argument to find the upper-bound and thus, we obtain the following which implies equation (19).

$$
\|\nabla_{\mathbf{W}} \left(\mathbf{H}_{p-1}(\mathbf{x}_i) \cdot \mathbf{g}_p(\mathbf{x}_i)^\top \cdot \mathbf{g}_p(\mathbf{x}_j) \cdot \mathbf{H}_{p-1}(\mathbf{x}_i)\right) \| \leq 4(\beta+2)^2.
$$

Next, we show that the expected value can be lower-bounded by -4 as in the case of theorem 4.1 above. Combining these two gives us the desired result. Consider $\mathbb{E}_{\mathbf{W}}[h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)]$. We compute this expectation iteratively as follows.

$$
\begin{aligned}
&\mathbb{E}_{\mathbf{W}}[h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)] \\
&= \mathbb{E}_{\mathbf{W}_0}[\mathbb{E}_{\mathbf{W}_1}[\ldots \mathbb{E}_{\mathbf{W}_\beta}[h_{\mathbf{W}}(\mathbf{x}_i, \mathbf{x}_j)]
\end{aligned}
$$

[7] In particular, each entry has to be scaled by $\frac{1}{\ell}$ for matrices $\{\mathbf{W}_p\}_{p\in[\beta]}$ and $\frac{1}{d}$ for the matrix $\mathbf{W}_0$.

$$\geq -4\mathbb{E}_{\mathbf{W}_0}\left[\mathbb{E}_{\mathbf{W}_1}\left[\dots\mathbb{E}_{\mathbf{W}_\beta}\left[\sum_{p\in[\beta]_0}\mathrm{Tr}(\mathbf{H}_{p-1}(\mathbf{x}_i)\cdot\mathbf{g}_p(\mathbf{x}_i)^\top\cdot\mathbf{g}_p(\mathbf{x}_j)\cdot\mathbf{H}_{p-1}(\mathbf{x}_i)^\top)\right]\right]\right].$$

The inequality combines equation 7 with Lemma D.1. We now prove the following inequality.

$$\mathbb{E}_{\mathbf{W}_0}\left[\mathbb{E}_{\mathbf{W}_1}\left[\dots\mathbb{E}_{\mathbf{W}_\beta}\left[\sum_{p\in[\beta]_0}\mathrm{Tr}(\mathbf{H}_{p-1}(\mathbf{x}_i)\cdot\mathbf{g}_p(\mathbf{x}_i)^\top\cdot\mathbf{g}_p(\mathbf{x}_j)\cdot\mathbf{H}_{p-1}(\mathbf{x}_i)^\top)\right]\right]\right]\leq 1. \tag{20}$$

Consider the inner-most expectation. Note that the only random variable is $\mathbf{W}_\beta$. Moreover, the term inside the trace is *scalar*. Note that the activation function σ satisfies $|\sigma'(x)|\leq 1$. Using the linearity of expectation, the LHS in equation (20) can be upper-bounded by the following.

$$\mathbb{E}_{\mathbf{W}_0}\left[\mathbb{E}_{\mathbf{W}_1}\left[\dots\mathbb{E}_{\mathbf{W}_{\beta-1}}\left[\mathrm{Tr}(\mathbf{H}_{\beta-1}(\mathbf{x}_i)\cdot\mathbf{H}_{\beta-1}(\mathbf{x}_i)^\top)\right]\right]\right] \tag{21}$$

$$+\,\mathbb{E}_{\mathbf{W}_0}\left[\mathbb{E}_{\mathbf{W}_1}\left[\dots\mathbb{E}_{\mathbf{W}_\beta}\left[\sum_{p\in[\beta]_0\setminus\{\beta\}}\mathrm{Tr}(\mathbf{H}_{p-1}(\mathbf{x}_i)\cdot\mathbf{g}_p(\mathbf{x}_i)^\top\cdot\mathbf{g}_p(\mathbf{x}_j)\cdot\mathbf{H}_{p-1}(\mathbf{x}_i)^\top)\right]\right]\right]. \tag{22}$$

The first sum in the above expression can be upper-bounded by 1, since $|\sigma(x)|\leq 1$. We will now show that the second sum is 0. Consider the inner-most expectation. The weights $\mathbf{W}_\beta$ appears only in the expression $\mathbf{g}_p(\mathbf{x}_i)^\top\cdot\mathbf{g}_p(\mathbf{x}_j)$. Moreover, note that every entry in $\mathbf{W}_\beta$ is an i.i.d. normal random variable with mean 0. Thus, the second summand simplifies to,

$$\mathbb{E}_{\mathbf{W}_0}\left[\mathbb{E}_{\mathbf{W}_1}\left[\dots\mathbb{E}_{\mathbf{W}_{\beta-1}}\left[\sum_{p\in[\beta]_0\setminus\{\beta,\beta-1\}}\mathrm{Tr}(\mathbf{H}_{p-1}(\mathbf{x}_i)\cdot\mathbf{g}_p(\mathbf{x}_i)^\top\cdot\mathbf{g}_p(\mathbf{x}_j)\cdot\mathbf{H}_{p-1}(\mathbf{x}_i)^\top)\right]\right]\right].$$

Applying the above argument repeatedly we obtain that the second summand (equation (22)) is 0.

Thus, we obtain the inequality in equation (20) which implies that $\mathbb{E}_{\mathbf{W}}[h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)]\geq -4$.

D.5. Proof of Theorem 6.1

In this section, we prove Theorem 6.1. The proof follows similar to those in previous sub-sections; we prove a bound on the gradient of the gradient inner-product and show that the expectation is non-negative. Combining these two with an argument similar to equation 10 we get the theorem.

Note that the dataset is obtained by considering i.i.d. samples from a d-dimensional unit sphere. Thus, the lower-bound on the expectation (*i.e.,* non-negative expectation of the gradient inner-product) follows from equation 9. Thus, it remains to prove an upper-bound on the norm of the gradient of the gradient inner-product term.

Throughout this proof, we will use $g(\mathbf{x})$ as a short-hand to denote $g_{\mathbf{W}}(\mathbf{x})$. Consider the gradient $\nabla_{\mathbf{W}}g(\mathbf{x})$. The the i^{th} component of this can be written as follows.

$$[\nabla_{\mathbf{W}}g(\mathbf{x})]_i=\gamma^2\zeta_{\mathbf{x}}(\mathbf{W})\left(\mathbf{W}_\beta^T\cdot\dots\mathbf{W}_{i+1}^T\cdot\mathbf{x}^T\cdot\mathbf{W}_1^T\cdot\dots\mathbf{W}_{i-1}^T\right). \tag{23}$$

Now consider, the gradient inner-product $h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)$. We want to upper-bound the quantity $\|\nabla_{(\mathbf{x}_i,\mathbf{x}_j)}h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\|$. From symmetry, this can be upper-bounded by $2\|\nabla_{\mathbf{x}_i}h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\|$. Consider the k^{th} coordinate of $\nabla_{\mathbf{x}_i}h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)$. Using equation 23, the assumption that $\{\mathbf{W}_i\}_{i\in[\beta]}$ are orthogonal matrices and taking the gradient, this can be written as,

$$[\nabla_{\mathbf{x}_i}h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)]_k=\gamma^2\zeta_{\mathbf{x}_i}(\mathbf{W})\mathbf{x}_j+\alpha^2\left(\mathbf{W}_\beta^T\cdot\dots\mathbf{W}_{i+1}^T\cdot\mathbf{x}^T\cdot\mathbf{W}_1^T\cdot\dots\mathbf{W}_{i-1}^T\right)(\nabla_{\mathbf{x}_i}\zeta_{\mathbf{x}_i}(\mathbf{W})). \tag{24}$$

Combining assumption 1 with equation 24 we have that $\|\nabla_{\mathbf{x}_i}h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\|$ is at most $2\gamma^2\beta\|\mathbf{x}_j\|\leq 2\gamma^2\beta$. For the definition of the scaling factor $\gamma=\frac{1}{\sqrt{2\beta}}$, we have that $2\gamma^2\beta=1$. Thus, $\|\nabla_{(\mathbf{x}_i,\mathbf{x}_j)}h_{\mathbf{W}}(\mathbf{x}_i,\mathbf{x}_j)\|\leq 2$.

E. Technical lemmas

We will briefly describe some technical lemmas we require in our analysis. The following Chernoff-style concentration bound is proved in Chapter 5 of Vershynin (2018).

Lemma E.1 (Concentration of Lipshitz function over a sphere)**.** *Let $\mathbf{x} \in \mathbb{R}^d$ be sampled uniformly from the surface of a d-dimensional sphere. Consider a Lipshitz function $\ell : \mathbb{R}^d \to \mathbb{R}$ which is differentiable everywhere. Let $\|\nabla \ell\|_2$ denote $\sup_{\mathbf{x} \in \mathbb{R}^d} \|\nabla \ell(\mathbf{x})\|_2$. Then for any $t \geq 0$ and some fixed constant $c \geq 0$, we have the following.*

$$\Pr\left[\left|\ell(\mathbf{x}) - \mathbb{E}[\ell(\mathbf{x})]\right| \geq t\right] \leq 2\exp\left(-\frac{cdt^2}{\rho^2}\right), \tag{25}$$

where $\rho \geq \|\nabla \ell\|_2$.

We will rely on the following generalization of lemma E.1. We would like to point out that the underlying metric is the Euclidean metric and thus we use the $\|.\|_2$-norm.

Corollary E.1. *Let $\mathbf{x}, \mathbf{y} \in \mathbb{R}^d$ be two mutually independent vectors sampled uniformly from the surface of a d-dimensional sphere. Consider a Lipshitz function $\ell : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$ which is differentiable everywhere. Let $\|\nabla \ell\|_2$ denote $\sup_{(\mathbf{x},\mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^d} \|\nabla \ell(\mathbf{x}, \mathbf{y})\|_2$. Then for any $t \geq 0$ and some fixed constant $c \geq 0$, we have the following.*

$$\Pr\left[\left|\ell(\mathbf{x}, \mathbf{y}) - \mathbb{E}[\ell(\mathbf{x}, \mathbf{y})]\right| \geq t\right] \leq 2\exp\left(-\frac{cdt^2}{\rho^2}\right), \tag{26}$$

where $\rho \geq \|\nabla \ell\|_2$.

Proof. This corollary can be derived from lemma E.1 as follows. Note that for every fixed $\tilde{\mathbf{y}} \in \mathbb{R}^d$, equation 25 holds. Additionally, we have that the vectors $\mathbf{x}$ and $\mathbf{y}$ are mutually independent. Hence we can write the LHS of equation 26 as the following.

$$\int_{(\tilde{\mathbf{y}})_1=-\infty}^{(\tilde{\mathbf{y}})_1=\infty} \dots \int_{(\tilde{\mathbf{y}})_d=-\infty}^{(\tilde{\mathbf{y}})_d=\infty} \Pr\left[\left|\ell(\mathbf{x}, \mathbf{y}) - \mathbb{E}[\ell(\mathbf{x}, \mathbf{y})]\right| \geq t \;\middle|\; \mathbf{y} = \tilde{\mathbf{y}}\right] \phi(\tilde{\mathbf{y}}) d(\tilde{\mathbf{y}})_1 \dots d(\tilde{\mathbf{y}})_d.$$

Here $\phi(\tilde{\mathbf{y}})$ refers to the pdf of the distribution of $\mathbf{y}$. From independence, the inner term in the integral evaluates to $\Pr\left[\left|\ell(\mathbf{x}, \tilde{\mathbf{y}}) - \mathbb{E}[\ell(\mathbf{x}, \tilde{\mathbf{y}})]\right| \geq t\right]$. We know this is less than or equal to $2\exp\left(-\frac{cdt^2}{\|\nabla \ell\|_2^2}\right)$. Therefore, the integral can be upper bounded by the following.

$$\int_{(\tilde{\mathbf{y}})_1=-\infty}^{(\tilde{\mathbf{y}})_1=\infty} \dots \int_{(\tilde{\mathbf{y}})_d=-\infty}^{(\tilde{\mathbf{y}})_d=\infty} 2\exp\left(-\frac{cdt^2}{\|\nabla \ell\|_2^2}\right) \phi(\tilde{\mathbf{y}}) d(\tilde{\mathbf{y}})_1 \dots d(\tilde{\mathbf{y}})_d.$$

Since $\phi(\tilde{\mathbf{y}})$ is a valid pdf, we get the required equation 26. □

Additionally, we will use the following facts about a normalized Gaussian random variable.

Lemma E.2. *For a normalized Gaussian $\mathbf{x}$ (*i.e.*, an $\mathbf{x}$ sampled uniformly from the surface of a unit d-dimensional sphere) the following statements are true.*

1. *$\forall p \in [d]$ we have that $\mathbb{E}[(\mathbf{x})_p] = 0$.*
2. *$\forall p \in [d]$ we have that $\mathbb{E}[(\mathbf{x})_p^2] = 1/d$.*

Proof. Part (1) can be proved by observing that the normalized Gaussian random variable is spherically symmetric about the origin. In other words, for every $p \in [d]$ the vectors $(x_1, x_2, \dots, x_p, \dots, x_d)$ and $(x_1, x_2, \dots, -x_p, \dots, x_d)$ are identically distributed. Hence $\mathbb{E}[x_p] = \mathbb{E}[-x_p]$ which implies that $\mathbb{E}[x_p] = 0$.

Part (2) can be proved by observing that for any $p, p' \in [d]$, x_p and $x_{p'}$ are identically distributed. Fix any $p \in [d]$. We have that $\sum_{p' \in [d]} \mathbb{E}[x_{p'}^2] = d \times \mathbb{E}[x_p^2]$. Note that we have

$$\sum_{p' \in [d]} \mathbb{E}[x_{p'}^2] = \int_{(\mathbf{x})_1=-\infty}^{(\mathbf{x})_1=\infty} \dots \int_{(\mathbf{x})_d=-\infty}^{(\mathbf{x})_d=\infty} \frac{\sum_{p' \in [d]} x_{p'}^2}{\sum_{p'' \in [d]} x_{p''}^2} \phi(\mathbf{x}) d(\mathbf{x})_1 \dots d(\mathbf{x})_d = 1.$$

Therefore $\mathbb{E}[x_p^2] = 1/d$. □

We use the following well-known Gaussian concentration inequality in our proofs (*e.g.,* Chapter 5 in Boucheron et al. (2013)).

Lemma E.3 (Gaussian Concentration)**.** *Let* $\mathbf{x} = (x_1, x_2, \ldots, x_d)$ *be* i.i.d. $\mathcal{N}(0, \nu^2)$ *random variables. Consider a Lipshitz function* $\ell : \mathbb{R}^d \to \mathbb{R}$ *which is differentiable everywhere. Let* $\|\nabla \ell\|_2$ *denote* $\sup_{\mathbf{x} \in \mathbb{R}^d} \|\nabla \ell(\mathbf{x})\|_2$*. Then for any* $t \geq 0$*, we have the following.*

$$\Pr\left[\left|\ell(\mathbf{x}) - \mathbb{E}[\ell(\mathbf{x})]\right| \geq t\right] \leq 2\exp\left(-\frac{t^2}{2\nu^2\rho^2}\right), \tag{27}$$

where $\rho \geq \|\nabla \ell\|_2$.

F. Additional discussion of the small weights assumption (assumption 1)

Without the small-weights assumption, the signal propagated forward or the gradients $\nabla_{\mathbf{W}} f_i$ could potentially blow up in magnitude, making the network untrainable. Proving non-vacuous bounds in case of such blow-ups in magnitude of the signal or the gradient is not possible in general, and thus, we assume this restricted class of weights.

Note that the small-weights assumption is not just a theoretical concern, but also usually holds in practice. Neural networks are often trained with *weight decay* regularizers of the form $\sum_i \|W_i\|_F^2$, which keep the weights small during optimization. The operator norm of convolutional layers have also recently been used as an effective regularizer for image classification tasks by Sedghi et al. (2018).

In the proof of theorem 4.1 we showed that assumption 1 holds with high probability at standard Gaussian initializations used in practice. While, in general, there is no reason to believe that such a small-weights assumption would continue to hold during optimization without explicit regularizers like weight decay, some recent work has shown evidence that the weights do not move too far away during training from the random initialization point for overparameterized neural networks (Neyshabur et al., 2018; Dziugaite & Roy, 2017; Nagarajan & Kolter, 2019; Zou et al., 2018; Allen-Zhu et al., 2018; Du et al., 2018; Oymak & Soltanolkotabi, 2018). It is worth noting though that all these results have been shown under some restrictive assumptions, such as the width requiring to be much larger than generally used by practitioners.