Gradient Descent with Early Stopping is Provably Robust to Label Noise for Overparameterized Neural Networks

May 5, 2019

Abstract

This paper is focused on investigating and demystifying an intriguing robustness phenomena in over-parameterized neural network training. In particular we provide empirical and theoretical evidence that first order methods such as gradient descent are provably robust to noise/corruption on a constant fraction of the labels despite over-parameterization under a rich dataset model. In particular: i) First, we show that in the first few iterations where the updates are still in the vicinity of the initialization these algorithms only fit to the correct labels essentially ignoring the noisy labels. ii) Secondly, we prove that to start to overfit to the noisy labels these algorithms must stray rather far from from the initial model which can only occur after many more iterations. Together, these show that gradient descent with early stopping is provably robust to label noise and shed light on empirical robustness of deep networks as well as commonly adopted early-stopping heuristics.

1. Introduction

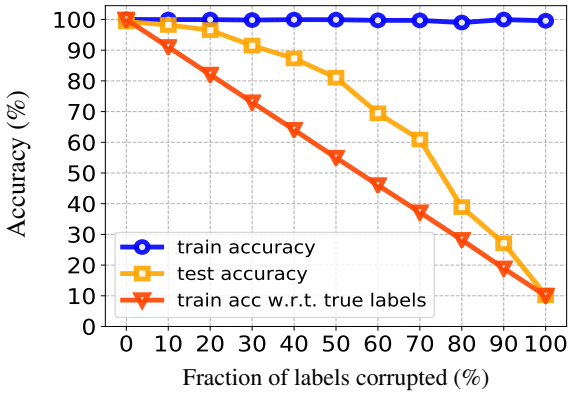
1.1. Motivation

This paper focuses on an intriguing phenomena: overparameterized neural networks are surprisingly robust to label noise when first order methods with early stopping is used to train them. To observe this phenomena consider Figure 1 where we perform experiments on the MNIST data set. Here, we corrupt a fraction of the labels of the training data by assigning their label uniformly at random. We then fit a four layer model via stochastic gradient descent and plot various performance metrics in Figures 1a and 1b. Figure 1a (blue curve) shows that indeed with a sufficiently large number of iterations the neural network does in fact perfectly fit the corrupted training data. However, Figure 1a also shows that such a model does not generalize to the test data (yellow curve) and the accuracy with respect to the ground truth labels degrades (orange curve). These plots clearly demonstrate that the model overfits with many iterations. In Figure 1b we repeat the same experiment but this time stop the updates after a few iterations (i.e. use early stopping). In this case the train accuracy degrades linearly (blue curve). However, perhaps unexpected, the test accuracy (yellow curve) remains high even with a significant amount of corruption. This suggests that with early stopping the model does not overfit and generalizes to new test data. Even more surprising, the train accuracy (orange curve) with respect to the ground truth labels continues to stay around $\%100$ even when $\%50$ of the labels are corrupted. That is, with early stopping overparameterized neural networks even correct the corrupted labels! These plots collectively demonstrate that overparameterized neural networks when combined with early stopping have unique generalization and robustness capabilities. As we detail further in Section D this phenomena holds (albeit less pronounced) for richer data models and architectures.

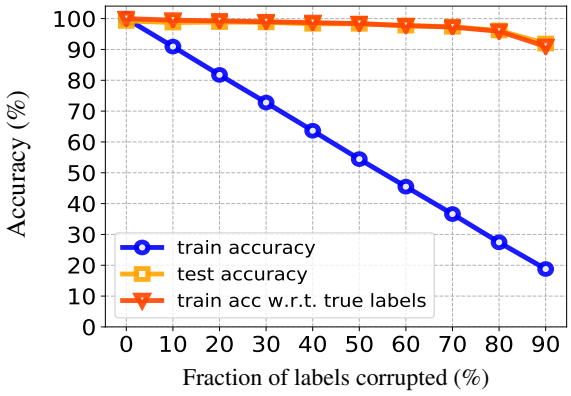
This paper aims to demonstrate and begin to demystify the surprising robustness of overparameterized neural networks when early stopping is used. We show that gradient descent is indeed provably robust to noise/corruption on a *constant fraction of the labels* in such overparametrized learning scenarios. In particular, under a fairly expressive dataset model and focusing on one-hidden layer networks, we show that after a few iterations (a.k.a. *early stopping*), gradient descent finds a model (i) that is within a small neighborhood of the point of initialization and (ii) only fits to the correct labels essentially ignoring the noisy labels. We complement these findings by proving that if the network is trained to overfit to the noisy labels, then the solution found by gradient descent must stray rather far from the initial model. Together, these results highlight the key features of a solution that *generalizes well* vs a solution that *fits well*.

1.2. Models

We now describe the dataset model used in our theoretical results. In this model we assume that the input samples $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ come from K clusters which are located on the unit Euclidian ball in $\mathbb{R}^d$. We also assume our data set consists of $\bar{K} \leq K$ classes where each class can be composed of multiple clusters. We consider a deterministic data set with n samples with roughly balanced clusters each



(a) Trained model after many iterations



(b) Trained model with early stopping

Figure 1. In these experiments we use a 4 layer neural network consisting of two convolution layers followed by two fully-connected layers to train a data set of 50,000 samples from MNIST with various amounts of random corruption on the labels. In this architecture the convolutional layers have width 64 and 128 kernels, and the fully-connected layers have 256 and 10 outputs, respectively. Overall, there are 4.8 million trainable parameters. We depict the training accuracy both w.r.t. the corrupted and uncorrupted training labels as well as the (uncorrupted) test accuracy. (a) Shows the performance after 200 epochs of Adadelta where near perfect fitting to the corrupted data is achieved. (b) Shows the performance with early stopping. We observe that with early stopping the trained neural network is robust to label corruption.

consisting on the order of n/K samples.[1] Finally, while we allow for multiple classes, in our theoretical model we assume the labels are scalars and take values in $[-1, 1]$ interval. We formally define our dataset model below and provide an illustration in Figure 2.

Definition 1.1 (Clusterable dataset) *Consider a data set of size n consisting of input/label pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$. We assume the input data have unit Euclidean norm and originate from K clusters with the ℓth cluster containing n_ℓ data points. We assume the number of points originating from each cluster is well-balanced in the sense that $c_{low}\frac{n}{K} \le n_\ell \le c_{up}\frac{n}{K}$ with c_{low} and c_{up} two numerical constants obeying $0 < c_{low} < c_{up} < 1$. We use $\{\boldsymbol{c}_\ell\}_{\ell=1}^K \subset \mathbb{R}^d$ to denote the cluster centers which are distinct unit Euclidian norm vectors. We assume the input data points $\boldsymbol{x}$ that belong to the ℓ-th cluster obey*

$$\|\boldsymbol{x} - \boldsymbol{c}_\ell\|_{\ell_2} \le \varepsilon_0,$$

with $\varepsilon_0 > 0$ denoting the input noise level.

We assume the labels y_i belong to one of $\bar{K} \le K$ classes. Specifically, we assume $y_i \in \{\alpha_1, \alpha_2, \ldots, \alpha_{\bar{K}}\}$ with $\{\alpha_\ell\}_{\ell=1}^{\bar{K}} \in [-1, 1]$ denoting the labels associated with each class. We assume all the elements of the same cluster belong to the same class and hence have the same label. However, a class can contain multiple clusters. Finally, we assume the labels are separated in the sense that

$$|\alpha_r - \alpha_s| \ge \delta \quad \text{for} \quad r \ne s, \tag{1.1}$$

with $\delta > 0$ denoting the class separation.

In the data model above $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$ are the K cluster centers that govern the input distribution. We note that in this model different clusters can be assigned to the same label. Hence, this setup is rich enough to model data which is not linearly separable: e.g. over $\mathbb{R}^2$, we can assign cluster centers $(0, 1)$ and $(0, -1)$ to label 1 and cluster centers $(1, 0)$ and $(-1, 0)$ to label -1. Note that the maximum number of classes are dictated by the separation δ. In particular, we can have at most $\bar{K} \le \frac{2}{\delta} + 1$ classes. We remark that this model is related to the setup of (4) which focuses on providing polynomial guarantees for learning shallow networks. Finally, note that, we need some sort of separation between the cluster centers to distinguish them. While Definition 1.1 doesn't specifies such separation explicitly, Definition 2.1 establishes a notion of separation in terms of how well a neural net can distinguish the cluster centers. Next, we introduce our noisy/corrupted dataset model.

Definition 1.2 ($(\rho, \varepsilon_0, \delta)$ corrupted dataset) *Let $\{(\boldsymbol{x}_i, \widetilde{y}_i)\}_{i=1}^n$ be an (ε_0, δ) clusterable dataset with $\alpha_1, \alpha_2, \ldots, \alpha_{\bar{K}}$ denoting the $\bar{K}$ possible class labels. A $(\rho, \varepsilon_0, \delta)$ noisy/corrupted dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is generated from $\{(\boldsymbol{x}_i, \widetilde{y}_i)\}_{i=1}^n$ as follows. For each cluster $1 \le \ell \le K$, at most ρn_ℓ of the labels associated with that cluster (which contains n_ℓ points) is assigned to another label value chosen from $\{\alpha_\ell\}_{\ell=1}^{\bar{K}}$. We shall refer to the initial labels $\{\widetilde{y}_i\}_{i=1}^n$ as the ground truth labels.*

We note that this definition allows for a fraction ρ of corruptions in each cluster.

[1] This is for ease of exposition rather than a particular challenge arising in the analysis.

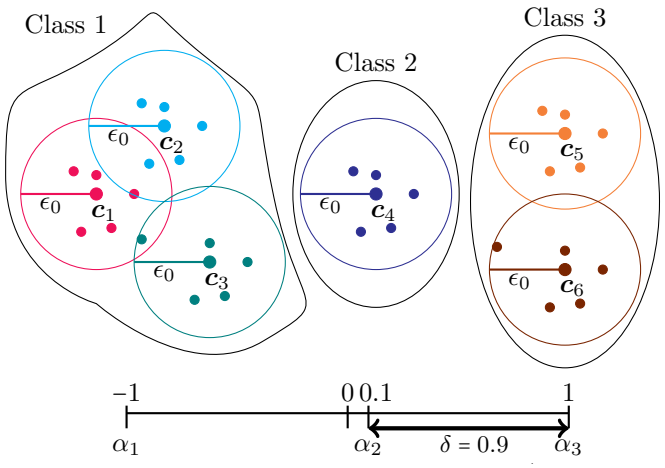


Figure 2. Visualization of the input/label samples and classes according to the clusterable dataset model in Definition 1.1. In the depicted example there are $K = 6$ clusters, $\bar{K} = 3$ classes. In this example the number of data points is $n = 30$ with each cluster containing 5 data points. The labels associated to classes 1, 2, and 3 are $\alpha_1 = -1$, $\alpha_2 = 0.1$, and $\alpha_3 = 1$, respectively so that $\delta = 0.9$. We note that the placement of points are exaggerated for clarity. In particular, per definition the cluster center and data points all have unit Euclidean norm. Also, there is no explicit requirements that the cluster centers be separated. The depicted separation is for exposition purposes only.

Network model: We will study the ability of neural networks to learn this corrupted dataset model. To proceed, let us introduce our neural network model. We consider a network with one hidden layer that maps $\mathbb{R}^d$ to $\mathbb{R}$. Denoting the number of hidden nodes by k, this network is characterized by an activation function ϕ, input weight matrix $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ and output weight vector $\boldsymbol{v} \in \mathbb{R}^k$. In this work, we will fix output $\boldsymbol{v}$ to be a unit vector where half the entries are $1/\sqrt{k}$ and other half are $-1/\sqrt{k}$ to simplify exposition.[2] We will only optimize over the weight matrix $\boldsymbol{W}$ which contains most of the network parameters and will be shown to be sufficient for robust learning. We will also assume ϕ has bounded first and second order derivatives, i.e. $|\phi'(z)|, |\phi''(z)| \le \Gamma$ for all z. The network's prediction at an input sample $\boldsymbol{x}$ is given by

$$\boldsymbol{x} \mapsto f(\boldsymbol{W}, \boldsymbol{x}) = \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}), \tag{1.2}$$

where the activation function ϕ applies entrywise. Given a dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$, we shall train the network via minimizing the empirical risk over the training data via a quadratic loss

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^{n}(y_i - f(\boldsymbol{x}_i, \boldsymbol{W}))^2. \tag{1.3}$$

In particular, we will run gradient descent with a constant learning rate η, starting from a random initialization $\boldsymbol{W}_0$ via the following updates

$$\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau). \tag{1.4}$$

[2] If the number of hidden units is odd we set one entry of $\boldsymbol{v}$ to zero.

2. Main results

Throughout, $\|\cdot\|$ denotes the largest singular value of a given matrix. The notation $\mathcal{O}(\cdot)$ denotes that a certain identity holds up to a fixed numerical constant. Also, c, c_0, C, C_0 etc. represent numerical constants.

Our main result shows that overparameterized neural networks, when trained via gradient descent using early stopping are fairly robust to label noise. The ability of neural networks to learn from the training data, even without label corruption, naturally depends on the diversity of the input training data. Indeed, if two input data are nearly the same but have different uncorrupted labels reliable learning is difficult. We will quantify this notion of diversity via a notion of condition number related to a covariance matrix involving the activation ϕ and the cluster centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$.

Definition 2.1 *Define the matrix of cluster centers*

$$\boldsymbol{C} = [\boldsymbol{c}_1 \ \dots \ \boldsymbol{c}_K]^T \in \mathbb{R}^{K\times d}.$$

Let $\boldsymbol{g} \sim \mathcal{N}(0, \boldsymbol{I}_d)$. Define the neural net covariance matrix $\boldsymbol{\Sigma}(\boldsymbol{C})$ as

$$\boldsymbol{\Sigma}(\boldsymbol{C}) = (\boldsymbol{C}\boldsymbol{C}^T) \odot \mathbb{E}_{\boldsymbol{g}}[\phi'(\boldsymbol{C}\boldsymbol{g})\phi'(\boldsymbol{C}\boldsymbol{g})^T].$$

Here $\odot$ denotes the elementwise product. Also denote the minimum eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{C})$ by $\lambda(\boldsymbol{C})$ and define the condition number associated with the cluster centers $\boldsymbol{C}$ as

$$\kappa(\boldsymbol{C}) = \sqrt{\frac{d}{K}}\frac{\|\boldsymbol{C}\|}{\lambda(\boldsymbol{C})}.$$

One can view $\boldsymbol{\Sigma}(\boldsymbol{C})$ as an empirical kernel matrix associated with the network where the kernel is given by

$\mathcal{K}(\boldsymbol{c}_i, \boldsymbol{c}_j) = \boldsymbol{\Sigma}_{ij}(\boldsymbol{C})$. Note that $\boldsymbol{\Sigma}(\boldsymbol{C})$ is trivially rank deficient if there are two cluster centers that are identical. In this sense, the minimum eigenvalue of $\boldsymbol{\Sigma}(\boldsymbol{C})$ will quantify the ability of the neural network to distinguish between distinct cluster centers. Therefore, one can think of $\kappa(\boldsymbol{C})$ as a condition number associated with the neural network which characterizes the distinctness/diversity of the cluster centers. The more distinct the cluster centers, the larger $\lambda(\boldsymbol{C})$ and smaller the condition number $\kappa(\boldsymbol{C})$ is. Indeed, based on results in (5) when the cluster centers are maximally diverse e.g. uniformly at random from the unit sphere $\kappa(\boldsymbol{C})$ scales like a constant. Throughout we shall assume that $\lambda(\boldsymbol{C})$ is strictly positive (and hence $\kappa(\boldsymbol{C}) < \infty$). This property is empirically verified to hold in earlier works (6) when ϕ is a standard activation (e.g. ReLU, softplus). As a concrete example, for ReLU activation, using results from (5) one can show if the cluster centers are separated by a distance $\nu > 0$, then $\lambda(\boldsymbol{C}) \geq \frac{\nu}{100K^2}$. We note that variations of the $\lambda(\boldsymbol{C}) > 0$ assumption based on the data points (i.e. $\lambda(\boldsymbol{X}) > 0$ not cluster centers) (5; 7; 8) are utilized to provide convergence guarantees for DNNs. Also see (9; 10) for other publications using related definitions. With a quantitative characterization of distinctiveness/diversity in place we are now ready to state our main result. Throughout we use c_Γ, C_Γ, etc. to denote constants only depending on Γ. We note that this Theorem is slightly simplified by ignoring logarithmic terms and precise dependencies on Γ. See Theorem E.13 for precise statements.

Theorem 2.2 (Robust learning with early stopping) *Consider an $(s, \varepsilon_0, \delta)$ clusterable corrupted data set of input/label pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ per Definition 1.2 with cluster centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$ aggregated as rows of a matrix $\boldsymbol{C} \in \mathbb{R}^{K\times d}$. Furthermore, let $\{\widetilde{y}_i\}_{i=1}^n$ be the corresponding uncorrupted ground truth labels. Also consider a one-hidden layer neural network of the form (1.2) where the activation ϕ obeys $|\phi(0)| \leq \Gamma$ and $|\phi'(z)|, |\phi''(z)| \leq \Gamma$ for all z and some $\Gamma \geq 1$. Furthermore, we set half of the entries of $\boldsymbol{v}$ to $1/\sqrt{k}$ and the other half to $-1/\sqrt{k}$[3] and train only over $\boldsymbol{W}$. Starting from an initial weight matrix $\boldsymbol{W}_0$ selected at random with i.i.d. $\mathcal{N}(0,1)$ entries we run Gradient Descent (GD) updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ on the least-squares loss (1.3) with step size $\eta = \bar{c}_\Gamma \frac{K}{n} \frac{1}{\|\boldsymbol{C}\|^2}$ with $\bar{c}_\Gamma$. Furthermore, assume the number of parameters obey*

$$kd \geq C_\Gamma \kappa^4(\boldsymbol{C}) \frac{K^4}{d},$$

with $\kappa(\boldsymbol{C})$ the neural net cluster condition number pre Definition 2.1. Then as long as $\epsilon_0 \leq \widetilde{c}_\Gamma/K^2$ and $\rho \leq \frac{\delta}{8}$ with probability at least $1 - 3/K^{100}$, after $\tau_0 = c_\Gamma \frac{K}{d}\lambda(\boldsymbol{C})\kappa^2(\boldsymbol{C})\log(\frac{1}{\rho})$ iterations, the neural network $f(\cdot, \boldsymbol{W}_{\tau_0})$ found by gradient descent assigns all the input samples $\boldsymbol{x}_i$ to the correct ground truth labels $\widetilde{y}_i$. That is,

$$\arg\min_{\alpha_\ell : 1\leq\ell\leq K} |f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) - \alpha_\ell| = \widetilde{y}_i, \tag{2.1}$$

holds for all $1 \leq i \leq n$. Furthermore, for all $0 \leq \tau \leq \tau_0$, the distance to the initial point obeys

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \leq \bar{C}_\Gamma \left(\sqrt{K} + \frac{K^2}{\|\boldsymbol{C}\|^2}\tau\varepsilon_0\right).$$

Theorem 2.2 shows that gradient descent with early stopping has a few intriguing properties:

Robustness. The solution found by gradient descent with early stopping degrades gracefully as the label corruption level ρ grows. In particular, as long as $\rho \leq \delta/8$, the final model is able to correctly classify all samples including the corrupted ones. In our setup, intuitively label gap obeys $\delta \sim \frac{1}{K}$, hence, we prove robustness to

$$\text{Total Number of corrupted labels} \lesssim \frac{n}{K}.$$

This result is independent of number of clusters and only depends on number of classes. An interesting future direction is to improve this result to allow on the order of n corrupted labels. Such a result maybe possible by using a multi-output classification neural network.

Early stopping time. We show that gradient descent finds a model that is robust to outliers after a few iterations. In particular using the maximum allowed step size, the number of iterations is of the order of $\frac{K}{d}\lambda(\boldsymbol{C})\kappa^2(\boldsymbol{C})\log(\frac{1}{\rho})$ which scales with K/d up to condition numbers.

Modest overparameterization. Our result requires modest overparemetrization and apply as soon as the number of parameters exceed the number of classes to the power four ($kd \gtrsim K^4$). Interestingly, the amount of overparameterization is essentially independent of the size of the training data n (ignoring logarithmic terms) and conditioning of the data points, only depending on the number of clusters and conditioning of the cluster centers. This can be interpreted as ensuring that the network has enough capacity to fit the cluster centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$ and the associated true labels.

Distance from initialization. Another feature of Theorem 2.2 is that the network weights do not stray far from the initialization as the distance between the initial model and the final model (at most) grows with the square root of the number of clusters ($\sqrt{K}$). This $\sqrt{K}$ dependence implies that the more clusters there are, the updates travel further away but continue to stay within a certain radius. This dependence is intuitive as the *Rademacher complexity* of the function space is dictated by the distance to initialization and should grow with the square-root of the number of input clusters to ensure the model is expressive enough to learn the dataset.

[3] If k is odd we set one entry to zero $\lfloor \frac{k-1}{2} \rfloor$ to $1/\sqrt{k}$ and $\lfloor \frac{k-1}{2} \rfloor$ entries to $-1/\sqrt{k}$.

References

[1] Chiyuan Zhang, Samy Bengio, Moritz Hardt, Benjamin Recht, and Oriol Vinyals. Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*, 2016.

[2] Daniel Soudry, Elad Hoffer, Mor Shpigel Nacson, Suriya Gunasekar, and Nathan Srebro. The implicit bias of gradient descent on separable data. *arXiv preprint arXiv:1710.10345*, 2017.

[3] Pratik Chaudhari, Anna Choromanska, Stefano Soatto, Yann LeCun, Carlo Baldassi, Christian Borgs, Jennifer Chayes, Levent Sagun, and Riccardo Zecchina. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*, 2016.

[4] Yuanzhi Li and Yingyu Liang. Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8168–8177, 2018.

[5] Samet Oymak and Mahdi Soltanolkotabi. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*, 2019.

[6] Bo Xie, Yingyu Liang, and Le Song. Diverse neural network learns true target functions. *arXiv preprint arXiv:1611.03131*, 2016.

[7] Simon S Du, Xiyu Zhai, Barnabas Poczos, and Aarti Singh. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*, 2018.

[8] Simon S Du, Jason D Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*, 2018.

[9] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*, 2018.

[10] Difan Zou, Yuan Cao, Dongruo Zhou, and Quanquan Gu. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*, 2018.

[11] Levent Sagun, Utku Evci, V Ugur Guney, Yann Dauphin, and Leon Bottou. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*, 2017.

[12] Sepp Hochreiter and Jürgen Schmidhuber. Flat minima. *Neural Computation*, 9(1):1–42, 1997.

[13] Samet Oymak and Mahdi Soltanolkotabi. Overparameterized nonlinear learning: Gradient descent takes the shortest path? *arXiv preprint arXiv:1812.10004*, 2018.

[14] Guy Gur-Ari, Daniel A Roberts, and Ethan Dyer. Gradient descent happens in a tiny subspace. *arXiv preprint arXiv:1812.04754*, 2018.

[15] J. Schur. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik*, 140:1–28, 1911.

A. Improvements for perfectly cluster-able data

We would like to note that in the limit of $\epsilon_0 \to 0$ where the input data set is perfectly clustered one can improve the amount of overparamterization. Indeed, the result above is obtained via a perturbation argument from this more refined result stated below.

Theorem A.1 (Training with perfectly clustered data) *Consier the setting and assumptions of Theorem E.14 with $\epsilon_0 = 0$. Starting from an initial weight matrix $\boldsymbol{W}_0$ selected at random with i.i.d. $\mathcal{N}(0,1)$ entries we run gradient descent updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ on the least-squares loss (1.3) with step size $\eta \le \frac{K}{2c_{up}n\Gamma^2\|\boldsymbol{C}\|^2}$. Furthermore, assume the number of parameters obey*

$$kd \ge C\Gamma^4\kappa^2(\boldsymbol{C})K^2,$$

with $\kappa(\boldsymbol{C})$ the neural net cluster condition number per Definition 2.1. Then, with probability at least $1 - 2/K^{100}$ over randomly initialized $\boldsymbol{W}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, the iterates $\boldsymbol{W}_\tau$ obey the following properties.

- *The distance to initial point $\boldsymbol{W}_0$ is upper bounded by*

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \le c\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}}.$$

- *After $\tau \ge \tau_0 := c\frac{K}{\eta n\lambda(\boldsymbol{C})}\log\left(\frac{\Gamma\sqrt{n\log K}}{\rho}\right)$ iterations, the entrywise predictions of the learned network with respect to the* ground truth labels *$\{\widetilde{y}_i\}_{i=1}^n$ satisfy*

$$|f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) - \widetilde{y}_i| \le 4\rho,$$

for all $1 \le i \le n$. Furthermore, if the noise level ρ obeys $\rho \le \delta/8$ the network predicts the correct label for all samples i.e.

$$\arg\min_{\alpha_\ell : 1\le\ell\le K} |f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) - \alpha_\ell| = \widetilde{y}_i \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{A.1}$$

This result shows that in the limit $\epsilon_0 \to 0$ where the data points are perfectly clustered, the required amount of overparameterization can be reduced from $kd \gtrsim K^4$ to $kd \gtrsim K^2$. In this sense this can be thought of a nontrivial analogue of (5) where the number of data points are replaced with the number of clusters and the condition number of the data points is replaced with a cluster condition number. This can be interpreted as ensuring that the network has enough capacity to fit the cluster centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$ and the associated true labels. Interestingly, the robustness benefits continue to hold in this case. However, in this perfectly clustered scenario there is no need for early stopping and a robust network is trained as soon as the number of iterations are sufficiently large. Infact, in this case given the clustered nature of the input data the network never overfits to the corrupted data even after many iterations.

B. To (over)fit to corrupted labels requires straying far from initialization

In this section we wish to provide further insight into why early stopping enables robustness and generalizable solutions. Our main insight is that while a neural network maybe expressive enough to fit a corrupted dataset, the model has to travel a longer distance from the point of initialization as a function of the distance from the cluster centers ε_0 and the amount of corruption. We formalize this idea as follows. Suppose

1. two input points are close to each other (e.g. they are from the same cluster),
2. but their labels are different, hence the network has to map them to distant outputs.

Then, the network has to be large enough so that it can amplify the small input difference to create a large output difference. Our first result formalizes this for a randomly initialized network. Our random initialization picks $\boldsymbol{W}$ with i.i.d. standard normal entries which ensures that the network is isometric i.e. given input $\boldsymbol{x}$, $\mathbb{E}[f(\boldsymbol{W}, \boldsymbol{x})^2] = \mathcal{O}(\|\boldsymbol{x}\|_{\ell_2}^2)$.

Theorem B.1 *Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$ be two vectors with unit Euclidean norm obeying $\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2} \le \epsilon_0$. Let $f(\boldsymbol{W}, \boldsymbol{x}) = \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$ where $\boldsymbol{v}$ is fixed, $\boldsymbol{W} \in \mathbb{R}^{k\times d}$, and $k \ge cd$ with $c > 0$ a fixed constant. Assume $|\phi'|, |\phi''| \le \Gamma$. Let y_1 and y_2 be two scalars satisfying $|y_2 - y_1| \ge \delta$. Suppose $\boldsymbol{W}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Then, with probability at least $1 - 2e^{-(k+d)} - 2e^{-\frac{t^2}{2}}$, for any $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ such that $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \le c\sqrt{k}$ and*

$$f(\boldsymbol{W}, \boldsymbol{x}_1) = y_1 \quad \text{and} \quad f(\boldsymbol{W}, \boldsymbol{x}_2) = y_2,$$

holds, we have

$$\|\boldsymbol{W} - \boldsymbol{W}_0\| \ge \frac{\delta}{C\Gamma\varepsilon_0} - \frac{t}{1000}.$$

In words, this result shows that in order to fit to a data set with a *single corrupted label*, a randomly initialized network has to traverse a distance of at least δ/ε_0. The next lemma clarifies the role of the corruption amount s and shows that more label corruption within a fixed class requires a model with a larger norm in order to fit the labels. For this result we consider a randomized model with ε_0^2 input noise variance.

Lemma B.2 *Let $\boldsymbol{c} \in \mathbb{R}^d$ be a cluster center. Consider $2s$ data points $\{\boldsymbol{x}_i\}_{i=1}^s$ and $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^s$ in $\mathbb{R}^d$ generated i.i.d. around $\boldsymbol{c}$ according to the following distribution*

$$\boldsymbol{c} + \boldsymbol{g} \quad \text{with} \quad \boldsymbol{g} \sim \mathcal{N}(0, \frac{\varepsilon_0^2}{d}\boldsymbol{I}_d).$$

Assign $\{\boldsymbol{x}_i\}_{i=1}^s$ with labels $y_i = y$ and $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^s$ with labels $\widetilde{y}_i = \widetilde{y}$ and assume these two labels are δ separated i.e. $|y - \widetilde{y}| \geq \delta$. Also suppose $s \leq d$ and $|\phi'| \leq \Gamma$. Then, any $\boldsymbol{W} \in \mathbb{R}^{k \times d}$ satisfying

$$f(\boldsymbol{W}, \boldsymbol{x}_i) = y_i \quad \text{and} \quad f(\boldsymbol{W}, \widetilde{\boldsymbol{x}}_i) = \widetilde{y}_i \quad \text{for} \quad i = 1, \ldots, s,$$

obeys $\|\boldsymbol{W}\|_F \geq \frac{\sqrt{s}\delta}{5\Gamma\varepsilon_0}$ with probability at least $1 - e^{-d/2}$.

Unlike Theorem E.15 this result lower bounds the network norm in lieu of the distance to the initialization $\boldsymbol{W}_0$. However, using the triangular inequality we can in turn get a guarantee on the distance from initialization $\boldsymbol{W}_0$ via triangle inequality as long as $\|\boldsymbol{W}_0\|_F \lesssim \mathcal{O}(\sqrt{s}\delta/\varepsilon_0)$ (e.g. by choosing a small ε_0).

The above Theorem implies that the model has to traverse a distance of at least

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \gtrsim \sqrt{\frac{\rho n}{K}}\frac{\delta}{\varepsilon_0},$$

to perfectly fit corrupted labels. In contrast, we note that the conclusions of the upper bound in Theorem 2.2 show that to be able to fit to the uncorrupted true labels the distance to initialization grows at most by $\tau\varepsilon_0$ after τ iterates. This demonstrates that there is a gap in the required distance to initialization for *fitting enough to generalize* and *overfitting*. To sum up, our results highlight that, one can find a network with good generalization capabilities and robustness to label corruption within a small neighborhood of the initialization and that the size of this neighborhood is independent of the corruption. However, to fit to the corrupted labels, one has to travel much more, increasing the search space and likely decreasing generalization ability. Thus, early stopping can enable robustness without overfitting by restricting the distance to the initialization.

C. Technical Approach and General Theory

In this section, we outline our approach to proving robustness of overparameterized neural networks. Towards this goal, we consider a general formulation where we aim to fit a general nonlinear model of the form $\boldsymbol{x} \mapsto f(\boldsymbol{\theta}, \boldsymbol{x})$ with $\boldsymbol{\theta} \in \mathbb{R}^p$ denoting the parameters of the model. For instance in the case of neural networks $\boldsymbol{\theta}$ represents its weights. Given a data set of n input/label pairs $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, we fit to this data by minimizing a nonlinear least-squares loss of the form

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\sum_{i=1}^n (y_i - f(\boldsymbol{\theta}, \boldsymbol{x}_i))^2.$$

which can also be written in the more compact form

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|f(\boldsymbol{\theta}) - \boldsymbol{y}\|_{\ell_2}^2 \quad \text{with} \quad f(\boldsymbol{\theta}) := \begin{bmatrix} f(\boldsymbol{\theta}, \boldsymbol{x}_1) \\ f(\boldsymbol{\theta}, \boldsymbol{x}_2) \\ \vdots \\ f(\boldsymbol{\theta}, \boldsymbol{x}_n) \end{bmatrix}.$$

To solve this problem we run gradient descent iterations with a constant learning rate η starting from an initial point $\boldsymbol{\theta}_0$. These iterations take the form

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau) \quad \text{with} \quad \nabla\mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}^T(\boldsymbol{\theta})\left(f(\boldsymbol{\theta}) - \boldsymbol{y}\right). \tag{C.1}$$

Here, $\mathcal{J}(\boldsymbol{\theta})$ is the $n \times p$ Jacobian matrix associated with the nonlinear mapping f defined via

$$\mathcal{J}(\boldsymbol{\theta}) = \left[\frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_1)}{\partial \boldsymbol{\theta}} \ \ldots \ \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_n)}{\partial \boldsymbol{\theta}}\right]^T. \tag{C.2}$$

C.1. Bimodal jacobian structure

Our approach is based on the hypothesis that the nonlinear model has a Jacobian matrix with *bimodal spectrum where few singular values are large and remaining singular values are small*. This assumption is inspired by the fact that realistic datasets are clusterable in a proper, possibly nonlinear, representation space. Indeed, one may argue that one reason for using neural networks is to automate the learning of such a representation (essentially the input to the softmax layer). We formalize the notion of bimodal spectrum below.

Assumption 1 (Bimodal Jacobian) *Let $\beta \geq \alpha \geq \epsilon > 0$ be scalars. Let $f : \mathbb{R}^p \to \mathbb{R}^n$ be a nonlinear mapping and consider a set $\mathcal{D} \subset \mathbb{R}^p$ containing the initial point $\boldsymbol{\theta}_0$ (i.e. $\boldsymbol{\theta}_0 \in \mathcal{D}$). Let $\mathcal{S}_+ \subset \mathbb{R}^n$ be a subspace and $\mathcal{S}_-$ be its complement. We say the mapping f has a Bimodal Jacobian with respect to the complementary subpspaces $\mathcal{S}_+$ and $\mathcal{S}_-$ as long as the following two assumptions hold for all $\boldsymbol{\theta} \in \mathcal{D}$.*

- **Spectrum over $\mathcal{S}_+$:** *For all $\boldsymbol{v} \in \mathcal{S}_+$ with unit Euclidian norm we have*

$$\alpha \leq \left\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{v}\right\|_{\ell_2} \leq \beta.$$

- **Spectrum over $\mathcal{S}_-$:** *For all $\boldsymbol{v} \in \mathcal{S}_-$ with unit Euclidian norm we have*

$$\left\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{v}\right\|_{\ell_2} \leq \epsilon.$$

We will refer to $\mathcal{S}_+$ as the signal subspace *and $\mathcal{S}_-$ as the* noise subspace*.*

When $\epsilon << \alpha$ the Jacobian is approximately low-rank. An extreme special case of this assumption is where $\epsilon = 0$ so that the Jacobian matrix is exactly low-rank. We formalize this assumption below for later reference.

Assumption 2 (Low-rank Jacobian) *Let $\beta \geq \alpha > 0$ be scalars. Consider a set $\mathcal{D} \subset \mathbb{R}^p$ containing the initial point $\boldsymbol{\theta}_0$ (i.e. $\boldsymbol{\theta}_0 \in \mathcal{D}$). Let $\mathcal{S}_+ \subset \mathbb{R}^n$ be a subspace and $\mathcal{S}_-$ be its complement. For all $\boldsymbol{\theta} \in \mathcal{D}$, $\boldsymbol{v} \in \mathcal{S}_+$ and $\boldsymbol{w} \in \mathcal{S}_-$ with unit Euclidian norm, we have that*

$$\alpha \leq \left\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{v}\right\|_{\ell_2} \leq \beta \quad \text{and} \quad \left\|\mathcal{J}^T(\boldsymbol{\theta})\boldsymbol{w}\right\|_{\ell_2} = 0.$$

Our dataset model in Definition 1.2 naturally has a low-rank Jacobian when $\epsilon_0 = 0$ and each input example is equal to one of the K cluster centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$. In this case, the Jacobian will be at most rank K since each row will be in the span of $\left\{\frac{\partial f(\boldsymbol{c}_\ell,\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}\right\}_{\ell=1}^K$. The subspace $\mathcal{S}_+$ is dictated by the *membership* of each cluster as follows: Let $\Lambda_\ell \subset \{1,\ldots,n\}$ be the set of coordinates i such that $\boldsymbol{x}_i = \boldsymbol{c}_\ell$. Then, subspace is characterized by

$$\mathcal{S}_+ = \{\boldsymbol{v} \in \mathbb{R}^n \;\big|\; \boldsymbol{v}_{i_1} = \boldsymbol{v}_{i_2} \text{ for all } i_1, i_2 \in \Lambda_\ell \text{ and } 1 \leq \ell \leq K\}.$$

When $\epsilon_0 > 0$ and the data points of each cluster are not the same as the cluster center we have the bimodal Jacobian structure of Assumption 1 where over $\mathcal{S}_-$ the spectral norm is small but nonzero.

In Section D, we verify that the Jacobian matrix of real datasets indeed have a bimodal structure i.e. there are few large singular values and the remaining singular values are small which further motivate Assumption 2. This is inline with earlier papers which observed that Hessian matrices of deep networks have bimodal spectrum (approximately low-rank) (11) and is related to various results demonstrating that there are flat directions in the loss landscape (12).

C.2. Meta result on learning with label corruption

Define the n-dimensional residual vector $\boldsymbol{r}$ where $\boldsymbol{r}(\boldsymbol{\theta}) = \begin{bmatrix} f(\boldsymbol{x}_1,\boldsymbol{\theta}) - \boldsymbol{y}_1 & \ldots & f(\boldsymbol{x}_n,\boldsymbol{\theta}) - \boldsymbol{y}_n \end{bmatrix}^T$. A key idea in our approach is that we argue that (1) in the absence of any corruption $\boldsymbol{r}(\boldsymbol{\theta})$ approximately lies on the subspace $\mathcal{S}_+$ and (2) if the labels are corrupted by a vector $\boldsymbol{e}$, then $\boldsymbol{e}$ approximately lies on the complement space. Before we state our general result we need to discuss another assumption and definition.

Assumption 3 (Smoothness) *The Jacobian mapping $\mathcal{J}(\boldsymbol{\theta})$ associated to a nonlinear mapping $f : \mathbb{R}^p \to \mathbb{R}^n$ is L-smooth if for all $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathbb{R}^p$ we have $\|\mathcal{J}(\boldsymbol{\theta}_2) - \mathcal{J}(\boldsymbol{\theta}_1)\| \leq L\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_{\ell_2}$.*[4]

Additionally, to connect our results to the number of corrupted labels, we introduce the notion of subspace diffusedness defined below.

[4]Note that, if $\frac{\partial \mathcal{J}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$ is continuous, the smoothness condition holds over any compact domain (albeit for a possibly large L).

Definition C.1 (Diffusedness) *$\mathcal{S}_+$ is γ diffused if for any vector $\boldsymbol{v} \in \mathcal{S}_+$*

$$\|\boldsymbol{v}\|_{\ell_\infty} \leq \sqrt{\gamma/n}\|\boldsymbol{v}\|_{\ell_2},$$

holds for some $\gamma > 0$.

The following theorem is our meta result on the robustness of gradient descent to sparse corruptions on the labels when the Jacobian mapping is exactly low-rank. Theorem E.14 for the perfectly clustered data ($\epsilon_0 = 0$) is obtained by combining this result with specific estimates developed for neural networks.

Theorem C.2 (Gradient descent with label corruption) *Consider a nonlinear least squares problem of the form $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|f(\boldsymbol{\theta}) - \boldsymbol{y})\|_{\ell_2}^2$ with the nonlinear mapping $f : \mathbb{R}^p \to \mathbb{R}^n$ obeying assumptions 2 and 3 over a unit Euclidian ball of radius $\frac{4\|\boldsymbol{r}_0\|_{\ell_2}}{\alpha}$ around an initial point $\boldsymbol{\theta}_0$ and $\boldsymbol{y} = \begin{bmatrix} y_1 & \ldots & y_n \end{bmatrix} \in \mathbb{R}^n$ denoting the corrupted labels. Also let $\widetilde{\boldsymbol{y}} = \begin{bmatrix} \widetilde{y}_1 & \ldots & \widetilde{y}_n \end{bmatrix} \in \mathbb{R}^n$ denote the uncorrupted labels and $\boldsymbol{e} = \boldsymbol{y} - \widetilde{\boldsymbol{y}}$ the corruption. Furthermore, suppose the initial residual $f(\boldsymbol{\theta}_0) - \widetilde{\boldsymbol{y}}$ with respect to the uncorrupted labels obey $f(\boldsymbol{\theta}_0) - \widetilde{\boldsymbol{y}} \in \mathcal{S}_+$. Then, running gradient descent updates of the from (C.1) with a learning rate $\eta \leq \frac{1}{2\beta^2}\min\left(1, \frac{\alpha\beta}{L\|\boldsymbol{r}_0\|_{\ell_2}}\right)$, all iterates obey*

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} \leq \frac{4\|\boldsymbol{r}_0\|_{\ell_2}}{\alpha}.$$

Furthermore, assume $\nu > 0$ is a precision level obeying $\nu \geq \|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}$. Then, after $\tau \geq \frac{5}{\eta\alpha^2}\log\left(\frac{\|\boldsymbol{r}_0\|_{\ell_2}}{\nu}\right)$ iterations, $\boldsymbol{\theta}_\tau$ achieves the following error bound with respect to the true labels

$$\|f(\boldsymbol{\theta}_\tau) - \widetilde{\boldsymbol{y}}\|_{\ell_\infty} \leq 2\nu.$$

Furthermore, if $\boldsymbol{e}$ has at most s nonzeros and $\mathcal{S}_+$ is γ diffused per Definition C.1, then using $\nu = \|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}$

$$\|f(\boldsymbol{\theta}_\tau) - \widetilde{\boldsymbol{y}}\|_{\ell_\infty} \leq 2\|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty} \leq \frac{\gamma\sqrt{s}}{n}\|\boldsymbol{e}\|_{\ell_2}.$$

This result shows that when the Jacobian of the nonlinear mapping is low-rank, gradient descent enjoys two intriguing properties. First, gradient descent iterations remain rather close to the initial point. Second, the estimated labels of the algorithm enjoy *sample-wise* robustness guarantees in the sense that the noise in the estimated labels are gracefully distributed over the dataset and the effects on individual label estimates are negligible. This theorem is the key result that allows us to prove Theorem E.14 when the data points are perfectly clustered ($\epsilon_0 = 0$). Furthermore, this theorem when combined with a perturbation analysis allows us to deal with data that is not perfectly clustered ($\epsilon_0 > 0$) and to conclude that with early stopping neural networks are rather robust to label corruption (Theorem 2.2).

Finally, we note that a few recent publication (7; 9; 13) require the Jacobian to be well-conditioned to fit labels perfectly. In contrast, our low-rank model cannot perfectly fit the corrupted labels. Furthermore, when the Jacobian is bimodal (as seems to be the case for many practical data sets and neural network models) it would take a very long time to perfectly fit the labels and as demonstrated earlier such a model does not generalize and is not robust to corruptions. Instead we focus on proving robustness with early stopping.

C.3. To (over)fit to corrupted labels requires straying far from initialization

In this section we state a result that provides further justification as to why early stopping of gradient descent leads to more robust models without overfitting to corrupted labels. This is based on the observation that while finding an estimate that fits the uncorrupted labels one does not have to move far from the initial estimate in the presence of corruption one has to stray rather far from the initialization with the distance from initialization increasing further in the presence of more corruption. We make this observation rigorous below by showing that it is more difficult to fit to the portion of the residual that lies on the noise space compared to the portion on the signal space (assuming $\alpha \gg \epsilon$).

Theorem C.3 *Denote the residual at initialization $\boldsymbol{\theta}_0$ by $\boldsymbol{r}_0 = f(\boldsymbol{\theta}_0) - \boldsymbol{y}$. Define the residual projection over the signal and noise space as*

$$E_+ = \|\Pi_{\mathcal{S}_+}(\boldsymbol{r}_0)\|_{\ell_2} \quad \text{and} \quad E_- = \|\Pi_{\mathcal{S}_-}(\boldsymbol{r}_0)\|_{\ell_2}.$$

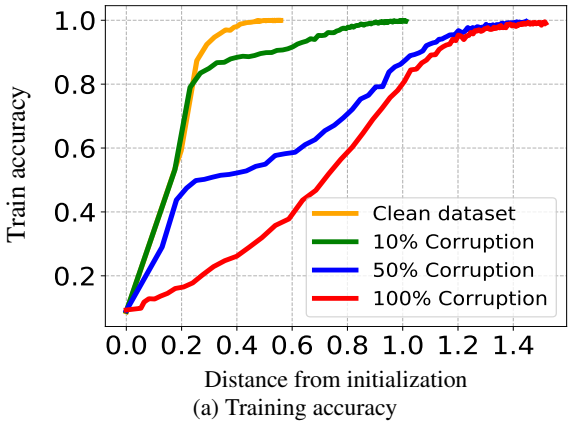
Suppose Assumption 1 holds over an Euclidian ball $\mathcal{D}$ of radius $R < \max\left(\frac{E_+}{\beta}, \frac{E_-}{\varepsilon}\right)$ around the initial point $\boldsymbol{\theta}_0$ with $\alpha \geq \epsilon$. Then, over $\mathcal{D}$ there exists no $\boldsymbol{\theta}$ that achieves zero training loss. In particular, if $\mathcal{D} = \mathbb{R}^p$, any parameter $\boldsymbol{\theta}$ achieving zero training loss ($f(\boldsymbol{\theta}) = \boldsymbol{y}$) satisfies the distance bound

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \geq \max\left(\frac{E_+}{\beta}, \frac{E_-}{\varepsilon}\right).$$

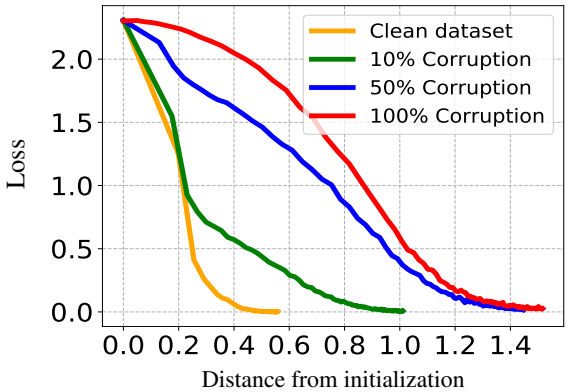
This theorem shows that the higher the corruption (and hence E_-) the further the iterates need to stray from the initial model to fit the corrupted data.

D. Numerical experiments

We conduct several experiments to investigate the robustness capabilities of deep networks to label corruption. In our first set of experiments, we explore the relationship between loss, accuracy, and amount of label corruption on the MNIST dataset to corroborate our theory. Our next experiments study the distribution of the loss and the Jacobian on the CIFAR-10 dataset. Finally, we simulate our theoretical model by generating data according to the corrupted data model of Definition 1.2 and verify the robustness capability of gradient descent with early stopping in this model.



(a) Training accuracy



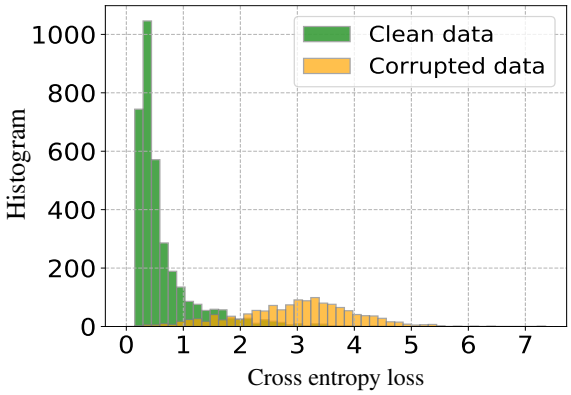
(b) Training loss

Figure 3. We depict the training accuracy of a LENET model trainined on 3000 samples from MNIST as a function of relative distance from initialization. Here, the x-axis keeps track of the distance between the current and initial weights of all layers combined.

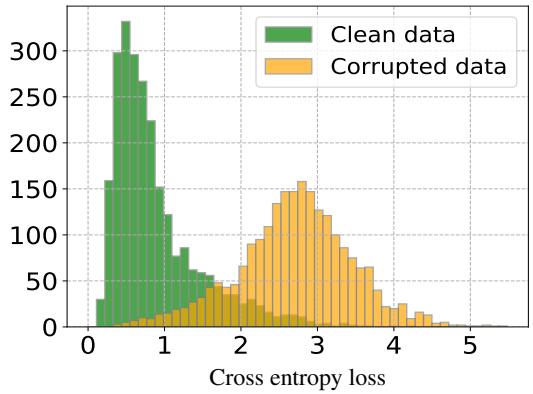
In Figure 3, we train the same model used in Figure 1 with $n = 3,000$ MNIST samples for different amounts of corruption. Our theory predicts that more label corruption leads to a larger distance to initialization. To probe this hypothesis, Figure 3a and 3b visualizes training accuracy and training loss as a function of the distance from the initialization. These results demonstrate that the distance from initialization gracefully increase with more corruption.

Next, we study the distribution of the individual sample losses on the CIFAR-10 dataset. We conducted two experiments using Resnet-20 with cross entropy loss[5]. In Figure 4 we assess the noise robustness of gradient descent where we used all 50,000 samples with either 30% random corruption

[5]We opted for cross entropy as it is the standard classification loss however least-squares loss achieves similar accuracy.



(a) 30% corruption



(b) 50% corruption

Figure 4. Histogram of the cross entropy loss of individual data points based on a model trained on 50,000 samples from CIFAR-10 with early stopping. Plot depicts 5000 random samples from these 50,000 samples. The loss distribution of clean and corrupted data are separated but gracefully overlap as the corruption level increases.

or 50% random corruption. Theorem E.14 predicts that when the corruption level is small, the loss distribution of corrupted vs clean samples should be separable. Figure 4 shows that when 30% of the data is corrupted the distributions are approximately separable. When we increase the shuffling amount to 50% the training loss on the clean data increases as predicted by our theory and the distributions start to gracefully overlap.

As described in Section C, our technical framework utilizes a bimodal prior on the Jacobian matrix (C.2) of the model. We now further investigate this hypothesis. For a multiclass task, the Jacobian matrix is essentially a 3-way tensor where dimensions are sample size (n), total number of parameters in the model (p), and the number of classes ($\bar{K}$). The neural network model we used for CIFAR 10 has around 270,000 parameters in total. In Figure 5 we illustrate the singular value spectrum of the two multiclass Jacobian models where

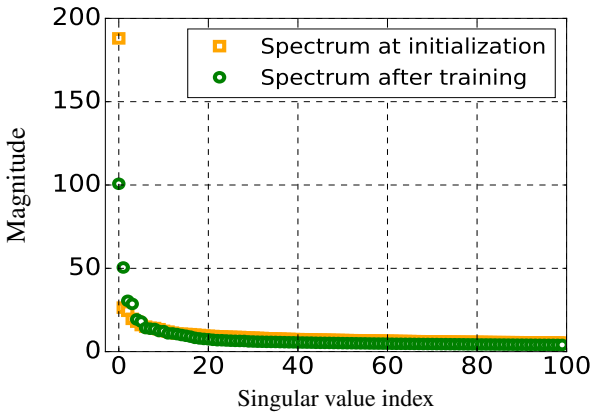
# $>0.1\times$ top singular	At initialization	After training
All classes	4	14
Correct class	15	16

Table 1. Jacobian of the network has few singular values that are significantly large i.e. larger than $0.1\times$ the spectral norm. This is true whether we consider the initial network or final network.

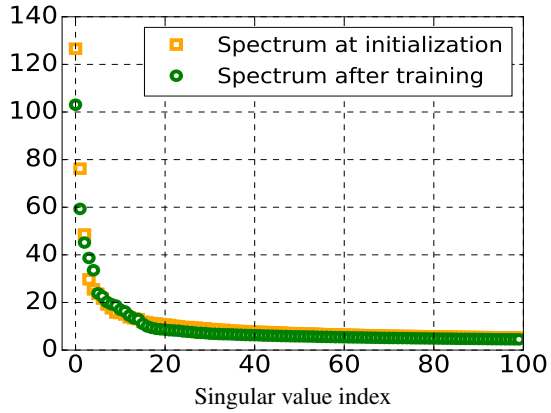
we form the Jacobian from all layers except the five largest (in total we use $\bar{p} \approx 90,000$ parameters).[6] We train the model with all samples and focus on the spectrum before and after the training. In Figure 5a, we picked $n = 1000$ samples and unfolded this tensor along parameters to obtain a $10,000 \times 90,000$ matrix which verifies our intuition on bimodality. In particular, only 10 to 20 singular values are larger than $0.1\times$ the top one. This is consistent with earlier works that studied the Hessian spectrum. However, focusing on the Jacobian has the added advantage of requiring only first order information (11; 14). A disadvantage is that the size of Jacobian grows with number of classes. Intuitively, cross entropy loss focuses on the class associated with the label hence in Figure 5b, we only picked the partial derivative associated with the correct class so that each sample is responsible for a single (size $\bar{p}$) vector. This allowed us to scale to $n = 10000$ samples and the corresponding spectrum is strikingly similar. Another intriguing finding is that the spectrums of before and after training are fairly close to each other highlighting that even at random initialization, spectrum is bimodal.

In Figure 6, we turn our attention to verifying our findings for the corrupted dataset model of Definition 1.2. We generated $K = 2$ classes where the associated clusters centers are generated uniformly at random on the unit sphere of $\mathbb{R}^{d=20}$. We also generate the input samples at random around these two clusters uniformly at random on a sphere of radius $\varepsilon_0 = 0.5$ around the corresponding cluster center. Hence, the clusters are guaranteed to be at least 1 distance from each other to prevent overlap. Overall we generate $n = 400$ samples (200 per class/cluster). Here, $\bar{K} = K = 2$ and the class labels are 0 and 1. We picked a network with $k = 1000$ hidden units and trained on a data set with 400 samples where 30% of the labels were corrupted. Figure 6a plots the trajectory of training error and highlights the model achieves good classification in the first few iterations and ends up overfitting later on. In Figures 6b and 6c, we focus on the loss distribution of 6a at iterations 80 and 4500. In this figure, we visualize the loss distribution of clean and corrupted data. Figure 6b highlights the loss distribution with early stopping and implies that the gap between corrupted and clean loss distributions is surprisingly resilient despite a large amount of corruption and the high-

[6] We depict the smaller Jacobian due to the computational cost of calculating the full Jacobian.



(a) All classes, 1k samples



(b) Correct class, 10k samples

Figure 5. Spectrum of the Jacobian obtained by plotting the singular values. (a) is obtained by forming the Jacobian by taking partial derivatives of all classes associated with a sample for 1000 samples. (b) is obtained by taking the class corresponding to the label for 10000 samples.

capacity of the model. In Figure 6c, we repeat plot after many more iterations at which point the model overfits. This plot shows that the distribution of the two classes overlap demonstrating that the model has overfit the corruption and lacks generalization/robustness.

E. Proofs

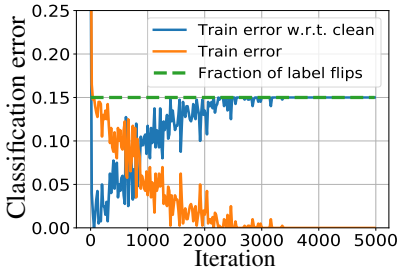
E.1. Proofs for General Theory

We begin by defining the average Jacobian which will be used throughout our analysis.

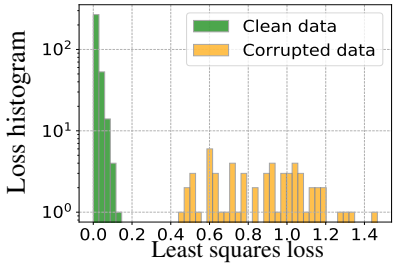
Definition E.1 (Average Jacobian) *We define the average Jacobian along the path connecting two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ as*

$$\mathcal{J}(\boldsymbol{y}, \boldsymbol{x}) := \int_0^1 \mathcal{J}(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x}))d\alpha. \tag{E.1}$$

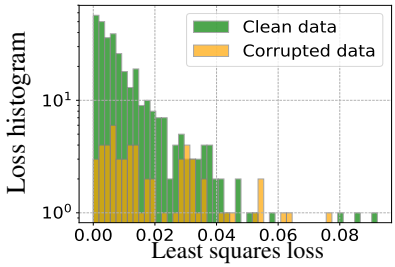
Lemma E.2 (Linearization of the residual) *Given gradient descent iterate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} - \eta\nabla\mathcal{L}(\boldsymbol{\theta})$, define*



(a) Fraction of incorrect predictions



(b) Loss histogram at iteration 80



(c) Loss histogram at iteration 4500

Figure 6. We experiment with the corrupted dataset model of Definition 1.2. We picked $K = 2$ classes and set $n = 400$ and $\varepsilon_0 = 0.5$. Trained 30% corrupted data with $k = 1000$ hidden units. Each corruption has 50% chance to remain in the correct class hence around 15% of the labels are actually flipped which corresponds to the dashed green line.

$$\boldsymbol{C}(\boldsymbol{\theta}) = \mathcal{J}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T.$$

The residuals $\hat{\boldsymbol{r}} = f(\hat{\boldsymbol{\theta}}) - \boldsymbol{y}$, $\boldsymbol{r} = f(\boldsymbol{\theta}) - \boldsymbol{y}$ obey the following equation

$$\hat{\boldsymbol{r}} = (\boldsymbol{I} - \eta\boldsymbol{C}(\boldsymbol{\theta}))\boldsymbol{r}.$$

Proof Following Definition E.1, denoting $f(\hat{\boldsymbol{\theta}}) - \boldsymbol{y} = \hat{\boldsymbol{r}}$ and $f(\boldsymbol{\theta}) - \boldsymbol{y} = \boldsymbol{r}$, we find that

$$\begin{aligned}
\hat{\boldsymbol{r}} &= \boldsymbol{r} - f(\boldsymbol{\theta}) + f(\hat{\boldsymbol{\theta}}) \\
&\overset{(a)}{=} \boldsymbol{r} + \mathcal{J}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&\overset{(b)}{=} \boldsymbol{r} - \eta\mathcal{J}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T\boldsymbol{r} \\
&= (\boldsymbol{I} - \eta\boldsymbol{C}(\boldsymbol{\theta}))\boldsymbol{r}.
\end{aligned} \tag{E.2}$$

Here (a) uses the fact that Jacobian is the derivative of f and (b) uses the fact that $\nabla \mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}(\boldsymbol{\theta})^T \boldsymbol{r}$. ■

Using Assumption C.1, one can show that sparse vectors have small projection on $\mathcal{S}_+$.

Lemma E.3 *Suppose Assumption C.1 holds. If $\boldsymbol{r} \in \mathbb{R}^n$ is a vector with s nonzero entries, we have that*

$$\|\Pi_{\mathcal{S}_+}(\boldsymbol{r})\|_{\ell_\infty} \leq \frac{\gamma\sqrt{s}}{n}\|\boldsymbol{r}\|_{\ell_2}. \tag{E.3}$$

Proof First, we bound the ℓ_2 projection of $\boldsymbol{r}$ on $\mathcal{S}_+$ as follows

$$\|\Pi_{\mathcal{S}_+}(\boldsymbol{r})\|_{\ell_2} = \sup_{\boldsymbol{v}\in\mathcal{S}_+} \frac{\boldsymbol{v}^T\boldsymbol{r}}{\|\boldsymbol{v}\|_{\ell_2}} \leq \sqrt{\frac{\gamma}{n}}\|\boldsymbol{r}\|_{\ell_1} \leq \sqrt{\frac{\gamma s}{n}}\|\boldsymbol{r}\|_{\ell_2}.$$

where we used the fact that $|\boldsymbol{v}_i| \leq \sqrt{\gamma}\|\boldsymbol{v}\|_{\ell_2}/\sqrt{n}$. Next, we conclude with

$$\|\Pi_{\mathcal{S}_+}(\boldsymbol{r})\|_{\ell_\infty} \leq \sqrt{\frac{\gamma}{n}}\|\Pi_{\mathcal{S}_+}(\boldsymbol{r})\|_{\ell_2} \leq \frac{\gamma\sqrt{s}}{n}\|\boldsymbol{r}\|_{\ell_2}.$$

■

E.1.1. Proof of Theorem C.2

Proof The proof will be done inductively over the properties of gradient descent iterates and is inspired from the recent work (13). In particular, (13) requires a well-conditioned Jacobian to fit labels perfectly. In contrast, we have a low-rank Jacobian model which cannot fit the noisy labels (or it would have trouble fitting if the Jacobian was approximately low-rank). Despite this, we wish to prove that gradient descent satisfies desirable properties such as robustness and closeness to initialization. Let us introduce the notation related to the residual. Set $\boldsymbol{r}_\tau = f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}$ and let $\boldsymbol{r}_0 = f(\boldsymbol{\theta}_0) - \boldsymbol{y}$ be the initial residual. We keep track of the growth of the residual by partitioning the residual as $\boldsymbol{r}_\tau = \bar{\boldsymbol{r}}_\tau + \bar{\boldsymbol{e}}_\tau$ where

$$\bar{\boldsymbol{e}}_\tau = \Pi_{\mathcal{S}_-}(\boldsymbol{r}_\tau) \quad , \quad \bar{\boldsymbol{r}}_\tau = \Pi_{\mathcal{S}_+}(\boldsymbol{r}_\tau).$$

We claim that for all iterations $\tau \geq 0$, the following conditions hold.

$$\bar{\boldsymbol{e}}_\tau = \bar{\boldsymbol{e}}_0 \tag{E.4}$$

$$\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 \leq \left(1 - \frac{\eta\alpha^2}{2}\right)^\tau \|\bar{\boldsymbol{r}}_0\|_{\ell_2}^2, \tag{E.5}$$

$$\frac{1}{4}\alpha\,\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} \leq \|\bar{\boldsymbol{r}}_0\|_{\ell_2} \leq \|\boldsymbol{r}_0\|_{\ell_2}. \tag{E.6}$$

Assuming these conditions hold till some $\tau > 0$, inductively, we focus on iteration $\tau + 1$. First, note that these conditions imply that for all $\tau \geq i \geq 0$, $\boldsymbol{\theta}_i \in \mathcal{D}$ where $\mathcal{D}$ is the Euclidian ball around $\boldsymbol{\theta}_0$ of radius $\frac{4\|\boldsymbol{r}_0\|_{\ell_2}}{\alpha}$. This directly follows from (E.6) induction hypothesis. Next, we claim that $\boldsymbol{\theta}_{\tau+1}$ is still within the set $\mathcal{D}$. This can be seen as follows:

Claim 1 *Under the induction hypothesis (E.4), $\boldsymbol{\theta}_{\tau+1} \in \mathcal{D}$.*

Proof Since range space of Jacobian is in $\mathcal{S}_+$ and $\eta \leq 1/\beta^2$, we begin by noting that

$$\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} = \eta\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\,(f(\boldsymbol{\theta}_\tau) - \boldsymbol{y})\|_{\ell_2} \tag{E.7}$$

$$\overset{(a)}{=} \eta\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\,(\Pi_{\mathcal{S}_+}(f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}))\|_{\ell_2} \tag{E.8}$$

$$\overset{(b)}{=} \eta\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2} \tag{E.9}$$

$$\overset{(c)}{\leq} \eta\beta\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} \tag{E.10}$$

$$\overset{(d)}{\leq} \frac{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}{\beta} \tag{E.11}$$

$$\overset{(e)}{\leq} \frac{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}{\alpha} \tag{E.12}$$

In the above, (a) follows from the fact that row range space of Jacobian is subset of $\mathcal{S}_+$ via Assumption 2. (b) follows from the definition of $\bar{\boldsymbol{r}}_\tau$. (c) follows from the upper bound on the spectral norm of the Jacobian over $\mathcal{D}$ per Assumption 2, (d) from the fact that $\eta \leq \frac{1}{\beta^2}$, (e) from $\alpha \leq \beta$. The latter combined with the triangular inequality and induction hypothesis (E.6) yields (after scaling (E.6) by $4/\alpha$)

$$\begin{aligned}\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\|_{\ell_2} \leq& \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} + \|\boldsymbol{\theta}_0 - \boldsymbol{\theta}_\tau\|_{\ell_2} \\ \leq& \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \frac{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}{\alpha} \leq \frac{4\|\boldsymbol{r}_0\|_{\ell_2}}{\alpha},\end{aligned}$$

concluding the proof of $\boldsymbol{\theta}_{\tau+1} \in \mathcal{D}$. ■

To proceed, we shall verify that (E.6) holds for $\tau + 1$ as well. Note that, following Lemma E.2, gradient descent iterate can be written as

$$\boldsymbol{r}_{\tau+1} = (\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{\theta}_\tau))\boldsymbol{r}_\tau.$$

Since both column and row space of $\boldsymbol{C}(\boldsymbol{\theta}_\tau)$ is subset of $\mathcal{S}_+$, we have that

$$\bar{\boldsymbol{e}}_{\tau+1} = \Pi_{\mathcal{S}_-}((\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{\theta}_\tau))\boldsymbol{r}_\tau) \tag{E.13}$$

$$= \Pi_{\mathcal{S}_-}(\boldsymbol{r}_\tau) \tag{E.14}$$

$$= \bar{\boldsymbol{e}}_\tau, \tag{E.15}$$

This shows the first statement of the induction. Next, over $\mathcal{S}_+$, we have

$$\bar{\boldsymbol{r}}_{\tau+1} = \Pi_{\mathcal{S}_+}((\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{\theta}_\tau))\boldsymbol{r}_\tau) \tag{E.16}$$

$$= \Pi_{\mathcal{S}_+}((\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{r}}_\tau) + \Pi_{\mathcal{S}_+}((\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{e}}_\tau) \tag{E.17}$$

$$= \Pi_{\mathcal{S}_+}((\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{r}}_\tau) \tag{E.18}$$

$$= (\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{r}}_\tau \tag{E.19}$$

where the second line uses the fact that $\bar{\boldsymbol{e}}_\tau \in \mathcal{S}_-$ and last line uses the fact that $\bar{\boldsymbol{r}}_\tau \in \mathcal{S}_+$. To proceed, we need to prove that $\boldsymbol{C}(\boldsymbol{\theta}_\tau)$ has desirable properties over $\mathcal{S}_+$, in particular, it contracts this space.

Claim 2 *let $\boldsymbol{P}_{\mathcal{S}_+} \in \mathbb{R}^{n\times n}$ be the projection matrix to $\mathcal{S}_+$ i.e. it is a positive semi-definite matrix whose eigenvectors over $\mathcal{S}_+$ is 1 and its complement is 0. Under the induction hypothesis and setup of the theorem, we have that*[7]

$$\beta^2 \boldsymbol{P}_{\mathcal{S}_+} \succeq \boldsymbol{C}(\boldsymbol{\theta}_\tau) \succeq \frac{1}{2}\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T \succeq \frac{\alpha^2}{2}\boldsymbol{P}_{\mathcal{S}_+}. \quad \text{(E.20)}$$

Proof The proof utilizes the upper bound on the learning rate. The argument is similar to the proof of Lemma 9.7 of (13). Suppose Assumption 3 holds. Then, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{D}$ we have

$$\begin{aligned}
&\|\mathcal{J}(\boldsymbol{\theta}_2, \boldsymbol{\theta}_1) - \mathcal{J}(\boldsymbol{\theta}_1)\| \\
&\quad = \left\| \int_0^1 \left(\mathcal{J}(\boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)) - \mathcal{J}(\boldsymbol{\theta}_1)\right) dt \right\|, \\
&\quad \le \int_0^1 \|\mathcal{J}(\boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1)) - \mathcal{J}(\boldsymbol{\theta}_1)\| \, dt, \\
&\quad \le \int_0^1 tL \|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_{\ell_2} \, dt \le \frac{L}{2}\|\boldsymbol{\theta}_2 - \boldsymbol{\theta}_1\|_{\ell_2}.
\end{aligned} \quad \text{(E.21)}$$

Thus, for $\eta \le \frac{\alpha}{L\beta\|\boldsymbol{r}_0\|_{\ell_2}}$,

$$\begin{aligned}
\|\mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau) - \mathcal{J}(\boldsymbol{\theta}_\tau)\| &\le \frac{L}{2}\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} \\
&= \frac{\eta L}{2}\left\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\left(f(\boldsymbol{\theta}_{\boldsymbol{\tau}}) - \boldsymbol{y}\right)\right\|_{\ell_2} \\
&\le \frac{\eta\beta L}{2}\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} \\
&\overset{(a)}{\le} \frac{\eta\beta L}{2}\|\bar{\boldsymbol{r}}_0\|_{\ell_2} \overset{(b)}{\le} \frac{\alpha}{2}.
\end{aligned} \quad \text{(E.22)}$$

where for (a) we utilized the induction hypothesis (E.6) and (b) follows from the upper bound on η. Now that (E.22) is established, using following lemma, we find

$$\mathcal{C}(\boldsymbol{\theta}_\tau) = \mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T \succeq (1/2)\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T.$$

The β^2 upper bound directly follows from Assumption 2 by again noticing range space of Jacobian is subset of $\mathcal{S}_+$.

Lemma E.4 (Asymmetric PSD perturbation) *Consider the matrices $\boldsymbol{A}, \boldsymbol{C} \in \mathbb{R}^{n\times p}$ obeying $\|\boldsymbol{A} - \boldsymbol{C}\| \le \alpha/2$. Also suppose $\boldsymbol{C}\boldsymbol{C}^T \succeq \alpha^2\boldsymbol{P}_{\mathcal{S}_+}$. Furthermore, assume range spaces of $\boldsymbol{A}, \boldsymbol{C}$ lies in $\mathcal{S}_+$. Then,*

$$\boldsymbol{A}\boldsymbol{C}^T \succeq \frac{\boldsymbol{C}\boldsymbol{C}^T}{2} \succeq \frac{\alpha^2}{2}\boldsymbol{P}_{\mathcal{S}_+}.$$

[7] We say $\boldsymbol{A} \succeq \boldsymbol{B}$ if $\boldsymbol{A} - \boldsymbol{B}$ is a positive semi-definite matrix in the sense that for any real vector $\boldsymbol{v}$, $\boldsymbol{v}^T(\boldsymbol{A} - \boldsymbol{B})\boldsymbol{v} \ge 0$.

Proof For $\boldsymbol{r} \in \mathcal{S}_+$ with unit Euclidian norm, we have

$$\begin{aligned}
\boldsymbol{r}^T\boldsymbol{A}\boldsymbol{C}^T\boldsymbol{r} &= \|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2}^2 + \boldsymbol{r}^T(\boldsymbol{A} - \boldsymbol{C})\boldsymbol{C}^T\boldsymbol{r} \\
&\ge \|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2}^2 - \|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2}\|\boldsymbol{r}^T(\boldsymbol{A} - \boldsymbol{C})\|_{\ell_2} \\
&= (\|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2} - \|\boldsymbol{r}^T(\boldsymbol{A} - \boldsymbol{C})\|_{\ell_2})\|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2} \\
&\ge (\|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2} - \alpha/2)\|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2} \\
&\ge \|\boldsymbol{C}^T\boldsymbol{r}\|_{\ell_2}^2/2.
\end{aligned}$$

Also, for any $\boldsymbol{r}$, by range space assumption $\boldsymbol{r}^T\boldsymbol{A}\boldsymbol{C}^T\boldsymbol{r} = \Pi_{\mathcal{S}_+}(\boldsymbol{r})^T\boldsymbol{A}\boldsymbol{C}^T\Pi_{\mathcal{S}_+}(\boldsymbol{r})$ (same for $\boldsymbol{C}\boldsymbol{C}^T$). Combined with above, this concludes the claim. ■

■

What remains is proving the final two statements of the induction (E.6). Note that, using the claim above and recalling (E.19) and using the fact that $\|\mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau)\| \le \beta$, the residual satisfies

$$\begin{aligned}
\|\bar{\boldsymbol{r}}_{\tau+1}\|_{\ell_2}^2 &= \|(\boldsymbol{I} - \eta\boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 \\
&= \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - 2\eta\bar{\boldsymbol{r}}_\tau^T\boldsymbol{C}_\tau\bar{\boldsymbol{r}}_\tau + \eta^2\bar{\boldsymbol{r}}_\tau^T\boldsymbol{C}_\tau^T\boldsymbol{C}_\tau\bar{\boldsymbol{r}}_\tau \\
&\le \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - \eta\bar{\boldsymbol{r}}_\tau^T\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau \\
&\quad + \eta^2\beta^2\bar{\boldsymbol{r}}_\tau^T\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau \\
&\le \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - (\eta - \eta^2\beta^2)\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 \\
&\le \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - \frac{\eta}{2}\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2.
\end{aligned} \quad \text{(E.23)}$$

where we used the fact that $\eta \le \frac{1}{2\beta^2}$. Now, using the fact that $\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T \succeq \alpha^2\boldsymbol{P}_{\mathcal{S}_+}$, we have

$$\begin{aligned}
\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - \frac{\eta}{2}\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 &\le (1 - \frac{\eta\alpha^2}{2})\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 \\
&\le (1 - \frac{\eta\alpha^2}{2})^{\tau+1}\|\bar{\boldsymbol{r}}_0\|_{\ell_2}^2,
\end{aligned}$$

which establishes the second statement of the induction (E.6). What remains is obtaining the last statement of (E.6). To address this, completing squares, observe that

$$\begin{aligned}
\|\bar{\boldsymbol{r}}_{\tau+1}\|_{\ell_2} &\le \sqrt{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - \frac{\eta}{2}\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2} \\
&\le \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} - \frac{\eta}{4}\frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2}{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}.
\end{aligned}$$

On the other hand, the distance to initial point satisfies

$$\begin{aligned}
\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\|_{\ell_2} &\le \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} + \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} \\
&\le \|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \eta\|\mathcal{J}(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2}.
\end{aligned}$$

Combining the last two lines (by scaling the second line by $\frac{1}{4}\alpha$) and using induction hypothesis (E.6), we find that

$$\begin{aligned}
&\frac{1}{4}\alpha\left\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_0\right\|_{\ell_2}+\left\|\bar{\boldsymbol{r}}_{\tau+1}\right\|_{\ell_2} \\
&\quad\leq \frac{1}{4}\alpha(\left\|\boldsymbol{\theta}_{\tau}-\boldsymbol{\theta}_0\right\|_{\ell_2}+\eta\|\mathcal{J}(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2})+\left\|\bar{\boldsymbol{r}}_{\tau}\right\|_{\ell_2} \\
&\qquad-\frac{\eta}{4}\frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2}{\left\|\bar{\boldsymbol{r}}_{\tau}\right\|_{\ell_2}} \\
&\quad\leq \left[\frac{1}{4}\alpha\|\boldsymbol{\theta}_{\tau}-\boldsymbol{\theta}_0\|_{\ell_2}+\left\|\bar{\boldsymbol{r}}_{\tau}\right\|_{\ell_2}\right] \\
&\qquad+\frac{\eta}{4}\left[\alpha\|\mathcal{J}(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2}-\frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2}{\left\|\bar{\boldsymbol{r}}_{\tau}\right\|_{\ell_2}}\right] \\
&\quad\leq \left[\frac{1}{4}\alpha\|\boldsymbol{\theta}_{\tau}-\boldsymbol{\theta}_0\|_{\ell_2}+\left\|\bar{\boldsymbol{r}}_{\tau}\right\|_{\ell_2}\right] \\
&\qquad+\frac{\eta}{4}\|\mathcal{J}(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2}\left[\alpha-\frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}{\left\|\bar{\boldsymbol{r}}_{\tau}\right\|_{\ell_2}}\right] \\
&\quad\leq \frac{1}{4}\alpha\|\boldsymbol{\theta}_{\tau}-\boldsymbol{\theta}_0\|_{\ell_2}+\left\|\bar{\boldsymbol{r}}_{\tau}\right\|_{\ell_2} \\
&\quad\leq \left\|\bar{\boldsymbol{r}}_0\right\|_{\ell_2}\leq\left\|\boldsymbol{r}_0\right\|_{\ell_2}.
\end{aligned}$$

This establishes the final line of the induction and concludes the proof of the upper bound on $\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2}$. To proceed, we shall bound the infinity norm of the residual. Using $\Pi_{\mathcal{S}_-}(\boldsymbol{e})=\Pi_{\mathcal{S}_-}(\boldsymbol{r}_0)=\bar{\boldsymbol{e}}_\tau$, note that

$$\begin{aligned}
\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}-\boldsymbol{e}\|_{\ell_\infty} &= \|\boldsymbol{r}_\tau-\boldsymbol{e}\|_{\ell_\infty} && \text{(E.24)}\\
&\leq \|\bar{\boldsymbol{r}}_\tau\|_{\ell_\infty}+\|\boldsymbol{e}-\bar{\boldsymbol{e}}_\tau\|_{\ell_\infty} && \text{(E.25)}\\
&= \|\bar{\boldsymbol{r}}_\tau\|_{\ell_\infty}+\|\boldsymbol{e}-\Pi_{\mathcal{S}_-}(\boldsymbol{e})\|_{\ell_\infty} && \text{(E.26)}\\
&= \|\bar{\boldsymbol{r}}_\tau\|_{\ell_\infty}+\|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}. && \text{(E.27)}
\end{aligned}$$

What remains is controlling $\|\bar{\boldsymbol{r}}_\tau\|_{\ell_\infty}$. For this term, we shall use the naive upper bound $\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}$. Using the rate of convergence of the algorithm (E.6), we have that

$$\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}\leq(1-\frac{\eta\alpha^2}{4})^\tau\|\boldsymbol{r}_0\|_{\ell_2}.$$

We wish the right hand side to be at most $\nu>0$ where $\nu\geq\|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}$. This implies that we need

$$(1-\frac{\eta\alpha^2}{4})^\tau\|\boldsymbol{r}_0\|_{\ell_2}\leq\nu \iff \tau\log(1-\frac{\eta\alpha^2}{4})\leq\log(\frac{\nu}{\|\boldsymbol{r}_0\|_{\ell_2}}) \quad \text{(E.28)}$$

$$\iff \tau\log(\frac{1}{1-\frac{\eta\alpha^2}{4}})\geq\log(\frac{\|\boldsymbol{r}_0\|_{\ell_2}}{\nu}) \quad \text{(E.29)}$$

To conclude, note that since $\frac{\eta\alpha^2}{4}\leq 1/8$ (as $\eta\leq 1/2\beta^2$), we have

$$\log(\frac{1}{1-\frac{\eta\alpha^2}{4}})\geq\log(1+\frac{\eta\alpha^2}{4})\geq\frac{\eta\alpha^2}{5}.$$

Consequently, if $\tau\geq\frac{5}{\eta\alpha^2}\log(\frac{\|\boldsymbol{r}_0\|_{\ell_2}}{\nu})$, we find that $\|\bar{\boldsymbol{r}}_\tau\|_{\ell_\infty}\leq\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}\leq\nu$, which guarantees

$$\|\boldsymbol{r}_\tau-\boldsymbol{e}\|_{\ell_\infty}\leq 2\nu.$$

which is the advertised result. If $\boldsymbol{e}$ is s sparse and $\mathcal{S}_+$ is diffused, applying Lemma C.1 we have

$$\|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}\leq\frac{\gamma\sqrt{s}}{n}\|\boldsymbol{e}\|_{\ell_2}.$$

■

E.1.2. Proof of Generic Lower Bound – Theorem C.3

Proof Suppose $\boldsymbol{\theta}\in\mathcal{D}$ satisfies $\boldsymbol{y}=f(\boldsymbol{\theta})$. Define $\boldsymbol{J}_\tau=\mathcal{J}((1-\tau)\boldsymbol{\theta}+\tau\boldsymbol{\theta}_0)$ and $\boldsymbol{J}=\mathcal{J}(\boldsymbol{\theta},\boldsymbol{\theta}_0)=\int_0^1\boldsymbol{J}_\tau d\tau$. Since Jacobian is derivative of f, we have that

$$f(\boldsymbol{\theta})-f(\boldsymbol{\theta}_0)=\int_0^1\boldsymbol{J}_\tau(\boldsymbol{\theta}-\boldsymbol{\theta}_0)d\tau=\boldsymbol{J}(\boldsymbol{\theta}-\boldsymbol{\theta}_0).$$

Now, define the matrices $\boldsymbol{J}_+=\Pi_{\mathcal{S}_+}(\boldsymbol{J})$ and $\boldsymbol{J}_-=\Pi_{\mathcal{S}_-}(\boldsymbol{J})$. Using Assumption 1, we bound the spectral norms via

$$\begin{aligned}
\|\boldsymbol{J}_+\| &= \sup_{\boldsymbol{v}\in\mathcal{S}_+,\|\boldsymbol{v}\|_{\ell_2}\leq 1}\|\boldsymbol{J}^T\boldsymbol{v}\|_{\ell_2}\leq\beta\\
\|\boldsymbol{J}_-\| &= \sup_{\boldsymbol{v}\in\mathcal{S}_-,\|\boldsymbol{v}\|_{\ell_2}\leq 1}\|\boldsymbol{J}^T\boldsymbol{v}\|_{\ell_2}\leq\epsilon.
\end{aligned}$$

To proceed, projecting the residual on $\mathcal{S}_+$, we find for any $\boldsymbol{\theta}$ with $f(\boldsymbol{\theta})=\boldsymbol{y}$

$$\begin{aligned}
&\Pi_{\mathcal{S}_+}(f(\boldsymbol{\theta})-f(\boldsymbol{\theta}_0))=\Pi_{\mathcal{S}_+}(\boldsymbol{J})(\boldsymbol{\theta}-\boldsymbol{\theta}_0)\implies\\
&\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\geq\frac{\|\Pi_{\mathcal{S}_+}(f(\boldsymbol{\theta})-f(\boldsymbol{\theta}_0))\|_{\ell_2}}{\beta}\geq\frac{E_+}{\beta}.
\end{aligned}$$

The identical argument for $\mathcal{S}_-$ yields $\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\geq\frac{E_-}{\epsilon}$. Together this implies

$$\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\geq\max(\frac{E_-}{\epsilon},\frac{E_+}{\beta}). \quad \text{(E.30)}$$

If R is strictly smaller than right hand side, we reach a contradiction as $\boldsymbol{\theta}\notin\mathcal{D}$. If $\mathcal{D}=\mathbb{R}^p$, we still find (E.30). ■

This shows that if ϵ is small and E_- is nonzero, gradient descent has to traverse a long distance to find a good model. Intuitively, if the projection over the noise space indeed contains the label noise, we actually don't want to fit that. Algorithmically, our idea fits the residual over the signal space and not worries about fitting over the noise space. Approximately speaking, this intuition corresponds to the ℓ_2 regularized problem

$$\min_{\boldsymbol{\theta}}\mathcal{L}(\boldsymbol{\theta})\qquad\|\boldsymbol{\theta}-\boldsymbol{\theta}_0\|_{\ell_2}\leq R.$$

If we set $R=\frac{E_+}{\beta}$, we can hope that solution will learn only the signal and does not overfit to the noise. The next section builds on this intuition and formalizes our algorithmic guarantees.

E.2. Proofs for Neural Networks

Throughout, $\sigma_{\min}(\cdot)$ denotes the smallest singular value of a given matrix. We first introduce helpful definitions that will be used in our proofs.

Definition E.5 (Support subspace) *Let $\{\boldsymbol{x}_i\}_{i=1}^n$ be an input dataset generated according to Definition 1.1. Also let $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^n$ be the associated cluster centers, that is, $\widetilde{\boldsymbol{x}}_i = \boldsymbol{c}_\ell$ iff $\boldsymbol{x}_i$ is from the ℓth cluster. We define the support subspace $\mathcal{S}_+$ as a subspace of dimension K, dictated by the cluster membership as follows. Let $\Lambda_\ell \subset \{1,\dots,n\}$ be the set of coordinates i such that $\tilde{\boldsymbol{x}}_i = \boldsymbol{c}_\ell$. Then, $\mathcal{S}_+$ is characterized by*

$$\mathcal{S}_+ = \{\boldsymbol{v} \in \mathbb{R}^n \,\big|\, \boldsymbol{v}_{i_1} = \boldsymbol{v}_{i_2} \quad \text{for all} \quad i_1, i_2 \in \Lambda_\ell,\ 1 \le \ell \le K\}.$$

Definition E.6 (Neural Net Jacobian) *Given input samples $(\boldsymbol{x}_i)_{i=1}^n$, form the input matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$. The Jacobian of the learning problem (1.3), at a matrix $\boldsymbol{W}$ is denoted by $\mathcal{J}(\boldsymbol{W},\boldsymbol{X}) \in \mathbb{R}^{n\times kd}$ and is given by*

$$\mathcal{J}(\boldsymbol{W},\boldsymbol{X})^T = (diag(\boldsymbol{v})\phi'(\boldsymbol{W}\boldsymbol{X}^T)) * \boldsymbol{X}^T.$$

Here $*$ denotes the Khatri-Rao product.

The following theorem is borrowed from (5) and characterizes three key properties of the neural network Jacobian. These are smoothness, spectral norm, and minimum singular value at initialization which correspond to Lemmas 6.6, 6.7, and 6.8 in that paper.

Theorem E.7 (Jacobian Properties at Cluster Center) *Suppose $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$ be an input dataset satisfying $\lambda(\boldsymbol{X}) > 0$. Suppose $|\phi'|, |\phi''| \le \Gamma$. The Jacobian mapping with respect to the input-to-hidden weights obey the following properties.*

- *Smoothness is bounded by*

$$\left\|\mathcal{J}(\widetilde{\boldsymbol{W}},\boldsymbol{X}) - \mathcal{J}(\boldsymbol{W},\boldsymbol{X})\right\| \le \frac{\Gamma}{\sqrt{k}}\|\boldsymbol{X}\|\left\|\widetilde{\boldsymbol{W}} - \boldsymbol{W}\right\|_F \quad \text{for all} \quad \widetilde{\boldsymbol{W}}, \boldsymbol{W} \in \mathbb{R}^{k\times d}$$

- *Top singular value is bounded by*

$$\|\mathcal{J}(\boldsymbol{W},\boldsymbol{X})\| \le \Gamma\|\boldsymbol{X}\|.$$

- *Let $C > 0$ be an absolute constant. As long as*

$$k \ge \frac{C\Gamma^2 \log n \|\boldsymbol{X}\|^2}{\lambda(\boldsymbol{X})}$$

At random Gaussian initialization $\boldsymbol{W}_0 \sim \mathcal{N}(0,1)^{k\times d}$, with probability at least $1 - 1/K^{100}$, we have

$$\sigma_{\min}\left(\mathcal{J}(\boldsymbol{W}_0,\boldsymbol{X})\right) \ge \sqrt{\lambda(\boldsymbol{X})/2}.$$

In our case, the Jacobian is **not** well-conditioned. However, it is pretty well-structured as described previously. To proceed, given a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and a subspace $\mathcal{S} \subset \mathbb{R}^n$, we define the minimum singular value of the matrix over this subspace by $\sigma_{\min}(\boldsymbol{X},\mathcal{S})$ which is defined as

$$\sigma_{\min}(\boldsymbol{X},\mathcal{S}) = \sup_{\|\boldsymbol{v}\|_{\ell_2}=1, \boldsymbol{U}\boldsymbol{U}^T = \boldsymbol{P}_{\mathcal{S}}} \|\boldsymbol{v}^T\boldsymbol{U}^T\boldsymbol{X}\|_{\ell_2}.$$

Here, $\boldsymbol{P}_{\mathcal{S}} \in \mathbb{R}^{n\times n}$ is the projection operator to the subspace. Hence, this definition essentially projects the matrix on $\mathcal{S}$ and then takes the minimum singular value over that projected subspace. The following theorem states the properties of the Jacobian at a clusterable dataset.

Theorem E.8 (Jacobian Properties at Clusterable Dataset) *Let input samples $(\boldsymbol{x}_i)_{i=1}^n$ be generated according to (ε_0,δ) clusterable dataset model of Definition 1.1 and define $\boldsymbol{X} = [\boldsymbol{x}_1 \ \dots \ \boldsymbol{x}_n]^T$. Let $\mathcal{S}_+$ be the support space and $(\tilde{\boldsymbol{x}}_i)_{i=1}^n$ be the associated clean dataset as described by Definition E.5. Set $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1 \ \dots \ \tilde{\boldsymbol{x}}_n]^T$. Assume $|\phi'|, |\phi''| \le \Gamma$ and $\lambda(\boldsymbol{C}) > 0$. The Jacobian mapping at $\tilde{\boldsymbol{X}}$ with respect to the input-to-hidden weights obey the following properties.*

- *Smoothness is bounded by*

$$\left\|\mathcal{J}(\widetilde{\boldsymbol{W}},\tilde{\boldsymbol{X}}) - \mathcal{J}(\boldsymbol{W},\tilde{\boldsymbol{X}})\right\| \le \Gamma\sqrt{\frac{c_{up}n}{kK}}\|\boldsymbol{C}\|\left\|\widetilde{\boldsymbol{W}} - \boldsymbol{W}\right\|_F \quad \text{for all} \quad \widetilde{\boldsymbol{W}}$$

- *Top singular value is bounded by*

$$\left\|\mathcal{J}(\boldsymbol{W},\tilde{\boldsymbol{X}})\right\| \le \sqrt{\frac{c_{up}n}{K}}\Gamma\|\boldsymbol{C}\|.$$

- *As long as*

$$k \ge \frac{C\Gamma^2 \log K \|\boldsymbol{C}\|^2}{\lambda(\boldsymbol{C})}$$

At random Gaussian initialization $\boldsymbol{W}_0 \sim \mathcal{N}(0,1)^{k\times d}$, with probability at least $1 - 1/K^{100}$, we have

$$\sigma_{\min}\left(\mathcal{J}(\boldsymbol{W}_0,\tilde{\boldsymbol{X}}),\mathcal{S}_+\right) \ge \sqrt{\frac{c_{low}n\lambda(\boldsymbol{C})}{2K}}$$

- *The range space obeys range$(\mathcal{J}(\boldsymbol{W}_0,\tilde{\boldsymbol{X}})) \subset \mathcal{S}_+$ where $\mathcal{S}_+$ is given by Definition E.5.*

Proof Let $\mathcal{J}(\boldsymbol{W},\boldsymbol{C})$ be the Jacobian at the cluster center matrix. Applying Theorem E.7, this matrix already obeys the properties described in the conclusions of this theorem with desired probability (for the last conclusion). We prove our theorem by relating the cluster center Jacobian to the clean dataset Jacobian matrix $\mathcal{J}(\boldsymbol{W},\tilde{\boldsymbol{X}})$.

Note that $\tilde{\boldsymbol{X}}$ is obtained by duplicating the rows of the cluster center matrix $\boldsymbol{C}$. This implies that $\mathcal{J}(\boldsymbol{W}, \tilde{\boldsymbol{X}})$ is obtained by duplicating the rows of the cluster center Jacobian. The critical observation is that, by construction in Definition 1.1, each row is duplicated somewhere between $c_{low}n/K$ and $c_{up}n/K$.

To proceed, fix a vector $\boldsymbol{v}$ and let $\tilde{\boldsymbol{p}} = \mathcal{J}(\boldsymbol{W}, \tilde{\boldsymbol{X}})\boldsymbol{v} \in \mathbb{R}^n$ and $\boldsymbol{p} = \mathcal{J}(\boldsymbol{W}, \boldsymbol{C})\boldsymbol{v} \in \mathbb{R}^K$. Recall the definition of the support sets Λ_ℓ from Definition E.5. We have the identity

$$\tilde{\boldsymbol{p}}_i = \boldsymbol{p}_\ell \quad \text{for all} \quad i \in \Lambda_\ell.$$

This implies $\tilde{\boldsymbol{p}} \in \mathcal{S}_+$ hence $\text{range}(\mathcal{J}(\boldsymbol{W}, \tilde{\boldsymbol{X}})) \subset \mathcal{S}_+$. Furthermore, the entries of $\tilde{\boldsymbol{p}}$ repeats the entries of $\boldsymbol{p}$ somewhere between $c_{low}n/K$ and $c_{up}n/K$. This implies that,

$$\sqrt{\frac{c_{up}n}{K}}\|\boldsymbol{p}\|_{\ell_2} \geq \|\tilde{\boldsymbol{p}}\|_{\ell_2} \geq \sqrt{\frac{c_{low}n}{K}}\|\boldsymbol{p}\|_{\ell_2},$$

and establishes the upper and lower bounds on the singular values of $\mathcal{J}(\boldsymbol{W}, \tilde{\boldsymbol{X}})$ over $\mathcal{S}_+$ in terms of the singular values of $\mathcal{J}(\boldsymbol{W}, \boldsymbol{C})$. Finally, the smoothness can be established similarly. Given matrices $\boldsymbol{W}, \tilde{\boldsymbol{W}}$, the rows of the difference

$$\left\|\mathcal{J}(\widetilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}}) - \mathcal{J}(\boldsymbol{W}, \tilde{\boldsymbol{X}})\right\|$$

is obtained by duplicating the rows of $\left\|\mathcal{J}(\widetilde{\boldsymbol{W}}, \boldsymbol{C}) - \mathcal{J}(\boldsymbol{W}, \boldsymbol{C})\right\|$ by at most $c_{up}n/K$ times. Hence the spectral norm is scaled by at most $\sqrt{c_{up}n/K}$. ■

Lemma E.9 (Upper bound on initial misfit) *Consider a one-hidden layer neural network model of the form $\boldsymbol{x} \mapsto \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$ where the activation ϕ has bounded derivatives obeying $|\phi(0)|, |\phi'(z)| \leq \Gamma$. Suppose entries of $\boldsymbol{v} \in \mathbb{R}^k$ are half $1/\sqrt{k}$ and half $-1/\sqrt{k}$ so that $\|\boldsymbol{v}\|_{\ell_2} = 1$. Also assume we have n data points $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_n \in \mathbb{R}^d$ with unit euclidean norm ($\|\boldsymbol{x}_i\|_{\ell_2} = 1$) aggregated as rows of a matrix $\boldsymbol{X} \in \mathbb{R}^{n\times d}$ and the corresponding labels given by $\boldsymbol{y} \in \mathbb{R}^n$ generated accoring to $(\rho, \varepsilon_0 = 0, \delta)$ noisy dataset (Definition 1.2). Then for $\boldsymbol{W}_0 \in \mathbb{R}^{k\times d}$ with i.i.d. $\mathcal{N}(0,1)$ entries*

$$\left\|\boldsymbol{v}^T\phi\left(\boldsymbol{W}_0\boldsymbol{X}^T\right) - \boldsymbol{y}\right\|_{\ell_2} \leq \mathcal{O}(\Gamma\sqrt{n\log K}),$$

holds with probability at least $1 - K^{-100}$.

Proof This lemma is based on a fairly straightforward union bound. First, by construction $\|\boldsymbol{y}\|_{\ell_2} \leq \sqrt{n}$. What remains is bounding $\|\boldsymbol{v}^T\phi\left(\boldsymbol{W}_0\boldsymbol{X}^T\right)\|_{\ell_2}$. Since $\varepsilon_0 = 0$ there are K unique rows. We will show that each of the unique rows is bounded with probability $1 - K^{-101}$ and union bounding will give the final result. Let $\boldsymbol{w}$ be a row of $\boldsymbol{W}_0$ and $\boldsymbol{x}$ be a row of $\boldsymbol{X}$. Since ϕ is Γ Lipschitz and $|\phi(0)| \leq \Gamma$, each entry of $\phi(\boldsymbol{X}\boldsymbol{w})$ is $\mathcal{O}(\Gamma)$-subgaussian. Hence $\boldsymbol{v}^T\phi(\boldsymbol{W}_0\boldsymbol{x})$ is weighted average of k i.i.d. subgaussians which are entries of $\phi(\boldsymbol{W}_0\boldsymbol{x})$. Additionally it is zero mean since $\sum_{i=1}^n \boldsymbol{v}_i = 0$. This means $\boldsymbol{v}^T\phi(\boldsymbol{W}_0\boldsymbol{x})$ is also $\mathcal{O}(\Gamma)$ subgaussian and obeys

$$\mathbb{P}(|\boldsymbol{v}^T\phi(\boldsymbol{W}_0\boldsymbol{x})| \geq c\Gamma\sqrt{\log K}) \leq K^{-101},$$

for some constant $c > 0$, concluding the proof. ■

E.2.1. Proof of Theorem E.14

We first prove a lemma regarding the projection of label noise on the cluster induced subspace.

Lemma E.10 *Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be an $(\rho, \varepsilon_0 = 0, \delta)$ clusterable noisy dataset as described in Definition 1.2. Let $\{\tilde{y}_i\}_{i=1}^n$ be the corresponding noiseless labels. Let $\mathcal{J}(\boldsymbol{W}, \boldsymbol{C})$ be the Jacobian at the cluster center matrix which is rank K and $\mathcal{S}_+$ be its column space. Then, the difference between noiseless and noisy labels satisfy the bound*

$$\|\Pi_{\mathcal{S}_+}(\boldsymbol{y} - \tilde{\boldsymbol{y}})\|_{\ell_\infty} \leq 2\rho.$$

Proof Let $\boldsymbol{e} = \boldsymbol{y} - \tilde{\boldsymbol{y}}$. Observe that by assumption, ℓth cluster has at most $s_\ell = \rho n_\ell$ errors. Let $\mathcal{I}_\ell$ denote the membership associated with cluster ℓ i.e. $\mathcal{I}_\ell \subset \{1, \ldots, n\}$ and $i \in \mathcal{I}_\ell$ if and only if $\boldsymbol{x}_i$ belongs to ℓth cluster. Let $\mathbf{1}(\ell) \in \mathbb{R}^n$ be the indicator function of the ℓth class where ith entry is 1 if $i \in \mathcal{I}_\ell$ and 0 else for $1 \leq i \leq n$. Then, denoting the size of the ℓth cluster by n_ℓ, the projection to subspace $\mathcal{S}_+$ can be written as the $\boldsymbol{P}$ matrix where

$$\boldsymbol{P} = \sum_{\ell=1}^K \frac{1}{n_\ell}\mathbf{1}(\ell)\mathbf{1}(\ell)^T.$$

Let $\boldsymbol{e}_\ell$ be the error pattern associated with ℓth cluster i.e. $\boldsymbol{e}_\ell$ is equal to $\boldsymbol{e}$ over $\mathcal{I}_\ell$ and zero outside. Since cluster membership is non-overlapping, we have that

$$\boldsymbol{P}\boldsymbol{e} = \sum_{\ell=1}^K \frac{1}{n_\ell}\mathbf{1}(\ell)\mathbf{1}(\ell)^T\boldsymbol{e}_\ell.$$

Similarly since supports of $\mathbf{1}(\ell)$ are non-overlapping, we have that

$$\|\boldsymbol{P}\boldsymbol{e}\|_{\ell_\infty} = \max_{1\leq\ell\leq K} \frac{1}{n_\ell}\mathbf{1}(\ell)\mathbf{1}(\ell)^T\boldsymbol{e}_\ell.$$

Now, using $\|\boldsymbol{e}\|_{\ell_\infty} \leq 2$ (max distance between two labels), observe that

$$\|\mathbf{1}(\ell)\mathbf{1}(\ell)^T\boldsymbol{e}_\ell\|_{\ell_\infty} \leq 2\|\mathbf{1}(\ell)\|_{\ell_\infty}\|\boldsymbol{e}_\ell\|_{\ell_1} = 2\|\boldsymbol{e}_\ell\|_{\ell_1}.$$

Since number of errors within cluster ℓ is at most $n_\ell\rho$, we find that

$$\|\boldsymbol{P}\boldsymbol{e}\|_{\ell_\infty} = \sum_{\ell=1}^K \|\frac{1}{n_\ell}\mathbf{1}(\ell)\mathbf{1}(\ell)^T\boldsymbol{e}_\ell\|_{\ell_\infty} \leq \frac{\|\boldsymbol{e}_\ell\|_{\ell_1}}{n_\ell} \leq 2\rho.$$

The final line yields the bound

$$\|\mathcal{P}_{\mathcal{S}_+}(\boldsymbol{y}-\tilde{\boldsymbol{y}})\|_{\ell_\infty} = \|\mathcal{P}_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty} = \|\boldsymbol{P}\boldsymbol{e}\|_{\ell_\infty} \le 2\rho.$$

■

With this, we are ready to state the proof of Theorem E.14.

Proof The proof is based on the meta Theorem C.2, hence we need to verify its Assumptions 2 and 3 with proper values and apply Lemma E.10 to get $\|\mathcal{P}_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}$. We will also make significant use of Corollary E.8.

Using Corollary E.8, Assumption 3 holds with $L = \Gamma\sqrt{\frac{c_{up}n}{kK}}\|\boldsymbol{C}\|$ where L is the Lipschitz constant of Jacobian spectrum. Denote $\boldsymbol{r}_\tau = f(\boldsymbol{W}_\tau) - \boldsymbol{y}$. Using Lemma E.9 with probability $1 - K^{-100}$, we have that $\|\boldsymbol{r}_0\|_{\ell_2} = \|\boldsymbol{y} - f(\boldsymbol{W}_0)\|_{\ell_2} \le \Gamma\sqrt{c_0 n \log K/128}$ for some $c_0 > 0$. Corollary E.8 guarantees a uniform bound for β, hence in Assumption 2, we pick

$$\beta \le \sqrt{\frac{c_{up}n}{K}}\Gamma\|\boldsymbol{C}\|.$$

We shall also pick the minimum singular value over $\mathcal{S}_+$ to be

$$\alpha = \frac{\alpha_0}{2} \quad \text{where} \quad \alpha_0 = \sqrt{\frac{c_{low}n\lambda(\boldsymbol{C})}{2K}},$$

We wish to verify Assumption 2 over the radius of

$$\begin{aligned} R =& \frac{4\|f(\boldsymbol{W}_0)-\boldsymbol{y}\|_{\ell_2}}{\alpha} \\ \le& \frac{\Gamma\sqrt{c_0 n\log K/8}}{\alpha} \\ =& \Gamma\sqrt{\frac{c_0 n\log K/2}{\frac{c_{low}n\lambda(\boldsymbol{C})}{2K}}} \\ =& \Gamma\sqrt{\frac{c_0 K\log K}{c_{low}\lambda(\boldsymbol{C})}}, \end{aligned}$$

neighborhood of $\boldsymbol{W}_0$. What remains is ensuring that Jacobian over $\mathcal{S}_+$ is lower bounded by α. Our choice of k guarantees that at the initialization, with probability $1 - K^{-100}$, we have

$$\sigma_{\min}(\mathcal{J}(\boldsymbol{W}_0, \boldsymbol{X}), \mathcal{S}_+) \ge \alpha_0.$$

Suppose $LR \le \alpha = \alpha_0/2$. Using triangle inequality on Jacobian spectrum, for any $\boldsymbol{W} \in \mathcal{D}$, using $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \le R$, we would have

$$\begin{aligned} \sigma_{\min}(\mathcal{J}(\boldsymbol{W}, \boldsymbol{X}), \mathcal{S}_+) \ge& \sigma_{\min}(\mathcal{J}(\boldsymbol{W}_0, \boldsymbol{X}), \mathcal{S}_+) - LR \\ \ge& \alpha_0 - \alpha = \alpha. \end{aligned}$$

Now, observe that

$$\begin{aligned} LR =& \Gamma\sqrt{\frac{c_{up}n}{kK}}\|\boldsymbol{C}\|\Gamma\sqrt{\frac{c_0 K\log(K)}{c_{low}\lambda(\boldsymbol{C})}} \\ =& \Gamma^2\|\boldsymbol{C}\|\sqrt{\frac{c_{up}c_0 n\log K}{c_{low}k\lambda(\boldsymbol{C})}} \\ \le& \frac{\alpha_0}{2} \\ =& \sqrt{\frac{c_{low}n\lambda(\boldsymbol{C})}{8K}}, \end{aligned}$$

as k satisfies

$$k \ge \mathcal{O}(\Gamma^4\|\boldsymbol{C}\|^2\frac{c_{up}K\log(K)}{c_{low}^2\lambda(\boldsymbol{C})^2}) \ge \mathcal{O}(\frac{\Gamma^4 K\log(K)\|\boldsymbol{C}\|^2}{\lambda(\boldsymbol{C})^2}).$$

Finally, since $LR = 4L\|\boldsymbol{r}_0\|_{\ell_2}/\alpha \le \alpha$, the learning rate is

$$\eta \le \frac{1}{2\beta^2}\min(1, \frac{\alpha\beta}{L\|\boldsymbol{r}_0\|_{\ell_2}}) = \frac{1}{2\beta^2} = \frac{K}{2c_{up}n\Gamma^2\|\boldsymbol{C}\|^2}.$$

Overall, the assumptions of Theorem C.2 holds with stated α, β, L with probability $1 - 2K^{-100}$ (union bounding initial residual and minimum singular value events). This implies for all $\tau > 0$ the distance of current iterate to initial obeys

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \le R.$$

The final step is the properties of the label corruption. Using Lemma E.10, we find that

$$\|\Pi_{\mathcal{S}_+}(\tilde{\boldsymbol{y}} - \boldsymbol{y})\|_{\ell_\infty} \le 2\rho.$$

Substituting the values corresponding to α, β, L yields that, for all gradient iterations with

$$\begin{aligned} \frac{5}{\eta\alpha^2}\log\left(\frac{\|\boldsymbol{r}_0\|_{\ell_2}}{2\rho}\right) \le& \frac{5}{\eta\alpha^2}\log\left(\frac{\Gamma\sqrt{c_0 n\log K/32}}{2\rho}\right) \\ =& \mathcal{O}(\frac{K}{\eta n\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})) \le \tau, \end{aligned}$$

denoting the clean labels by $\tilde{\boldsymbol{y}}$ and applying Theorem C.2, we have that, the infinity norm of the residual obeys (using $\|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty} \le 2\rho$)

$$\|f(\boldsymbol{W}) - \tilde{\boldsymbol{y}}\|_{\ell_\infty} \le 4\rho.$$

This implies that if $\rho \le \delta/8$, the network will miss the correct label by at most $\delta/2$, hence all labels (including noisy ones) will be correctly classified. ■

E.2.2. Proof of Theorem E.15

Consider

$$f(\boldsymbol{W}, \boldsymbol{x}) = \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$$

and note that

$$\nabla_{\boldsymbol{x}} f(\boldsymbol{W},\boldsymbol{x}) = \boldsymbol{W}^T \operatorname{diag}\left(\phi'\left(\boldsymbol{W}\boldsymbol{x}\right)\right)\boldsymbol{v}$$

Thus

$$\begin{aligned}\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{W},\boldsymbol{x})\boldsymbol{u} =& \boldsymbol{v}^T \operatorname{diag}\left(\phi'\left(\boldsymbol{W}\boldsymbol{x}\right)\right)\boldsymbol{W}\boldsymbol{u}\\ =& \sum_{\ell=1}^{k} \boldsymbol{v}_\ell \phi'\left(\langle \boldsymbol{w}_\ell, \boldsymbol{x}\rangle\right)\boldsymbol{w}_\ell^T\boldsymbol{u}\end{aligned}$$

Thus

$$\nabla_{\boldsymbol{w}_\ell}\left(\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{W},\boldsymbol{x})\boldsymbol{u}\right) = \boldsymbol{v}_\ell\left(\phi''(\boldsymbol{w}_\ell^T\boldsymbol{x})(\boldsymbol{w}_\ell^T\boldsymbol{u})\boldsymbol{x} + \phi'(\boldsymbol{w}_\ell^T\boldsymbol{x})\boldsymbol{u}\right)$$

Thus, denoting vectorization of a matrix by $\operatorname{vect}(\cdot)$

$$\begin{aligned}&\operatorname{vect}(\boldsymbol{U})^T\left(\frac{\partial}{\partial \operatorname{vect}(\boldsymbol{W})}\frac{\partial}{\partial \boldsymbol{x}} f(\boldsymbol{W},\boldsymbol{x})\right)\boldsymbol{u}\\ =&\sum_{\ell=1}^{k}\boldsymbol{v}_\ell\left(\phi''(\boldsymbol{w}_\ell^T\boldsymbol{x})(\boldsymbol{w}_\ell^T\boldsymbol{u})(\boldsymbol{u}_\ell^T\boldsymbol{x}) + \phi'(\boldsymbol{w}_\ell^T\boldsymbol{x})(\boldsymbol{u}_\ell^T\boldsymbol{u})\right)\\ =&\boldsymbol{u}^T\boldsymbol{W}^T\operatorname{diag}\left(\boldsymbol{v}\right)\operatorname{diag}\left(\phi''(\boldsymbol{W}\boldsymbol{x})\right)\boldsymbol{U}\boldsymbol{x} + \boldsymbol{v}^T\operatorname{diag}\left(\phi'\left(\boldsymbol{W}\boldsymbol{x}\right)\right)\boldsymbol{U}\boldsymbol{u}\end{aligned}$$

Thus by the general mean value theorem there exists a point $(\widetilde{\boldsymbol{W}},\widetilde{\boldsymbol{x}})$ in the square $(\boldsymbol{W}_0,\boldsymbol{x}_1),(\boldsymbol{W}_0,\boldsymbol{x}_2),(\boldsymbol{W},\boldsymbol{x}_1)$ and $(\boldsymbol{W},\boldsymbol{x}_2)$ such that

$$\begin{aligned}&\left(f(\boldsymbol{W},\boldsymbol{x}_2) - f(\boldsymbol{W}_0,\boldsymbol{x}_2)\right) - \left(f(\boldsymbol{W},\boldsymbol{x}_1) - f(\boldsymbol{W}_0,\boldsymbol{x}_1)\right)\\ &\quad= (\boldsymbol{x}_2-\boldsymbol{x}_1)^T\widetilde{\boldsymbol{W}}^T\operatorname{diag}\left(\boldsymbol{v}\right)\operatorname{diag}\left(\phi''(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{x}})\right)(\boldsymbol{W}-\boldsymbol{W}_0)\widetilde{\boldsymbol{x}}\\ &\qquad+\boldsymbol{v}^T\operatorname{diag}\left(\phi'\left(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{x}}\right)\right)(\boldsymbol{W}-\boldsymbol{W}_0)(\boldsymbol{x}_2-\boldsymbol{x}_1)\end{aligned}$$

Using the above we have that

$$\begin{aligned}&\left|\left(f(\boldsymbol{W},\boldsymbol{x}_2) - f(\boldsymbol{W}_0,\boldsymbol{x}_2)\right) - \left(f(\boldsymbol{W},\boldsymbol{x}_1) - f(\boldsymbol{W}_0,\boldsymbol{x}_1)\right)\right|\\ &\overset{(a)}{\le}\left|(\boldsymbol{x}_2-\boldsymbol{x}_1)^T\widetilde{\boldsymbol{W}}^T\operatorname{diag}\left(\boldsymbol{v}\right)\operatorname{diag}\left(\phi''(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{x}})\right)(\boldsymbol{W}-\boldsymbol{W}_0)\widetilde{\boldsymbol{x}}\right|\\ &\quad+\left|\boldsymbol{v}^T\operatorname{diag}\left(\phi'\left(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{x}}\right)\right)(\boldsymbol{W}-\boldsymbol{W}_0)(\boldsymbol{x}_2-\boldsymbol{x}_1)\right|\\ &\overset{(b)}{\le}\left(\|\boldsymbol{v}\|_{\ell_\infty}\|\widetilde{\boldsymbol{x}}\|_{\ell_2}\left\|\widetilde{\boldsymbol{W}}\right\| + \|\boldsymbol{v}\|_{\ell_2}\right)\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\overset{(c)}{\le}\left(\frac{1}{\sqrt{k}}\|\widetilde{\boldsymbol{x}}\|_{\ell_2}\left\|\widetilde{\boldsymbol{W}}\right\| + 1\right)\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\overset{(d)}{\le}\left(\frac{1}{\sqrt{k}}\left\|\widetilde{\boldsymbol{W}}\right\| + 1\right)\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\overset{(e)}{\le}\left(\frac{1}{\sqrt{k}}\|\boldsymbol{W}_0\| + \frac{1}{\sqrt{k}}\left\|\widetilde{\boldsymbol{W}}-\boldsymbol{W}_0\right\| + 1\right)\\ &\quad\cdot\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\overset{(f)}{\le}\left(\frac{1}{\sqrt{k}}\|\boldsymbol{W}_0\| + \frac{1}{\sqrt{k}}\left\|\widetilde{\boldsymbol{W}}-\boldsymbol{W}_0\right\|_F + 1\right)\\ &\quad\cdot\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\overset{(g)}{\le}\left(\frac{1}{\sqrt{k}}\left\|\widetilde{\boldsymbol{W}}-\boldsymbol{W}_0\right\|_F + 3 + 2\sqrt{\frac{d}{k}}\right)\\ &\quad\cdot\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\overset{(h)}{\le} C\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\| \qquad \text{(E.31)}\end{aligned}$$

Here, (a) follows from the triangle inequality, (b) from simple algebraic manipulations along with the fact that $|\phi'(z)| \le \Gamma$ and $|\phi''(z)| \le \Gamma$, (c) from the fact that $\boldsymbol{v}_\ell = \pm\frac{1}{\sqrt{k}}$, (d) from $\|\boldsymbol{x}_2\|_{\ell_2} = \|\boldsymbol{x}_1\|_{\ell_2} = 1$ which implies $\|\widetilde{\boldsymbol{x}}\|_{\ell_2} \le 1$, (e) from triangular inequality, (f) from the fact that Frobenius norm dominates the spectral norm, (g) from the fact that with probability at least $1-2e^{-(d+k)}$, $\|\boldsymbol{W}_0\| \le 2(\sqrt{k}+\sqrt{d})$, and (h) from the fact that $\left\|\widetilde{\boldsymbol{W}}-\boldsymbol{W}_0\right\| \le \|\boldsymbol{W}-\boldsymbol{W}_0\|_F \le \widetilde{c}\sqrt{k}$ and $k \ge cd$.

Next we note that for a Gaussian random vector $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0},\boldsymbol{I}_d)$ we have

$$\begin{aligned}\|\phi(\boldsymbol{g}^T\boldsymbol{x}_2) - \phi(\boldsymbol{g}^T\boldsymbol{x}_1)\|_{\psi_2} \qquad &\text{(E.32)}\\ =&\|\phi(\boldsymbol{g}^T\boldsymbol{x}_2) - \phi(\boldsymbol{g}^T\boldsymbol{x}_1)\|_{\psi_2}\\ =&\|\phi'\left(t\boldsymbol{g}^T\boldsymbol{x}_2 + (1-t)\boldsymbol{g}^T\boldsymbol{x}_1\right)\boldsymbol{g}^T(\boldsymbol{x}_2-\boldsymbol{x}_1)\|_{\psi_2}\\ \le&\Gamma\|\boldsymbol{g}^T(\boldsymbol{x}_2-\boldsymbol{x}_1)\|_{\psi_2}\\ \le&c\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}. \qquad \text{(E.33)}\end{aligned}$$

Also note that

$$\begin{aligned}f(\boldsymbol{W}_0,\boldsymbol{x}_2) - f(\boldsymbol{W}_0,\boldsymbol{x}_1) =& \boldsymbol{v}^T\left(\phi\left(\boldsymbol{W}_0\boldsymbol{x}_2\right) - \phi\left(\boldsymbol{W}_0\boldsymbol{x}_1\right)\right)\\ \sim& \sum_{\ell=1}^{k}\boldsymbol{v}_\ell\left(\phi(\boldsymbol{g}_\ell^T\boldsymbol{x}_2) - \phi(\boldsymbol{g}_\ell^T\boldsymbol{x}_1)\right)\end{aligned}$$

where $\boldsymbol{g}_1,\boldsymbol{g}_2,\ldots,\boldsymbol{g}_k$ are i.i.d. vectors with $\mathcal{N}(0,\boldsymbol{I}_d)$ distribution. Also for $\boldsymbol{v}$ obeying $\boldsymbol{1}^T\boldsymbol{v} = 0$ this random variable

has mean zero. Hence, using the fact that weighted sum of subGaussian random variables are subgaussian combined with (G.2) we conclude that $f(\boldsymbol{W}_0, \boldsymbol{x}_2) - f(\boldsymbol{W}_0, \boldsymbol{x}_1)$ is also subGaussian obeying $\|f(\boldsymbol{W}_0, \boldsymbol{x}_2) - f(\boldsymbol{W}_0, \boldsymbol{x}_1)\|_{\psi_2} \leq c\Gamma \|\boldsymbol{v}\|_{\ell_2} \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2}$. Thus

$$\begin{aligned}|f(\boldsymbol{W}_0, \boldsymbol{x}_2) - f(\boldsymbol{W}_0, \boldsymbol{x}_1)| \leq & ct\Gamma \|\boldsymbol{v}\|_{\ell_2} \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2} \\ = & ct\Gamma \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2}, \end{aligned} \tag{E.34}$$

with probability at least $1 - e^{-\frac{t^2}{2}}$.

Now combining (G.1) and (G.3) we have

$$\begin{aligned}\delta \leq & |y_2 - y_2| \\ = & |f(\boldsymbol{W}, \boldsymbol{x}_1) - f(\boldsymbol{W}, \boldsymbol{x}_2)| \\ = & \left|\boldsymbol{v}^T \left(\phi(\boldsymbol{W}\boldsymbol{x}_2) - \phi(\boldsymbol{W}\boldsymbol{x}_1)\right)\right| \\ \leq & |(f(\boldsymbol{W}, \boldsymbol{x}_2) - f(\boldsymbol{W}_0, \boldsymbol{x}_2)) - (f(\boldsymbol{W}, \boldsymbol{x}_1) - f(\boldsymbol{W}_0, \boldsymbol{x}_1))| \\ & + \left|\boldsymbol{v}^T \left(\phi(\boldsymbol{W}_0\boldsymbol{x}_2) - \phi(\boldsymbol{W}_0\boldsymbol{x}_1)\right)\right| \\ \leq & C\Gamma \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2} \|\boldsymbol{W} - \boldsymbol{W}_0\| + ct\Gamma \|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2} \\ \leq & C\Gamma\varepsilon_0 \left(\|\boldsymbol{W} - \boldsymbol{W}_0\| + \frac{1}{1000}t\right)\end{aligned}$$

Thus

$$\|\boldsymbol{W} - \boldsymbol{W}_0\| \geq \frac{\delta}{C\Gamma\varepsilon_0} - \frac{t}{1000},$$

with high probability.

E.3. Perturbation analysis for perfectly clustered data (Proof of Theorem 2.2)

Denote average neural net Jacobian at data $\boldsymbol{X}$ via

$$\mathcal{J}(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{X}) = \int_0^1 \mathcal{J}(\alpha\boldsymbol{W}_1 + (1-\alpha)\boldsymbol{W}_2, \boldsymbol{X}) d\alpha.$$

Lemma E.11 (Perturbed Jacobian Distance) *Let $\boldsymbol{X} = [\boldsymbol{x}_1 \ \ldots \ \boldsymbol{x}_n]^T$ be the input matrix obtained from Definition 1.1. Let $\tilde{\boldsymbol{X}}$ be the noiseless inputs where $\tilde{\boldsymbol{x}}_i$ is the cluster center corresponding to $\boldsymbol{x}_i$. Given weight matrices $\boldsymbol{W}_1, \boldsymbol{W}_2, \tilde{\boldsymbol{W}}_1, \tilde{\boldsymbol{W}}_2$, we have that*

$$\begin{aligned}&\|\mathcal{J}(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}_1, \tilde{\boldsymbol{W}}_2, \tilde{\boldsymbol{X}})\| \\ &\qquad \leq \Gamma\sqrt{n}\left(\frac{\|\tilde{\boldsymbol{W}}_1 - \boldsymbol{W}_1\|_F + \|\tilde{\boldsymbol{W}}_2 - \boldsymbol{W}_2\|_F}{2\sqrt{k}} + \varepsilon_0\right).\end{aligned}$$

Proof Given $\boldsymbol{W}, \tilde{\boldsymbol{W}}$, we write

$$\begin{aligned}\|\mathcal{J}(\boldsymbol{W}, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}})\| \leq & \|\mathcal{J}(\boldsymbol{W}, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}, \boldsymbol{X})\| \\ & + \|\mathcal{J}(\tilde{\boldsymbol{W}}, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}})\|.\end{aligned}$$

We first bound

$$\begin{aligned}&\|\mathcal{J}(\boldsymbol{W}, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}, \boldsymbol{X})\| \\ &= \|\mathrm{diag}(\boldsymbol{v})\phi'(\boldsymbol{W}\boldsymbol{X}^T) * \boldsymbol{X}^T - \mathrm{diag}(\boldsymbol{v})\phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T) * \boldsymbol{X}^T\| \\ &= \frac{1}{\sqrt{k}}\|(\phi'(\boldsymbol{W}\boldsymbol{X}^T) - \phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T)) * \boldsymbol{X}^T\|\end{aligned}$$

To proceed, we use the results on the spectrum of Hadamard product of matrices due to Schur (15). Given $\boldsymbol{A} \in \mathbb{R}^{k\times d}, \boldsymbol{B} \in \mathbb{R}^{n\times d}$ matrices where $\boldsymbol{B}$ has unit length rows, we have

$$\begin{aligned}\|\boldsymbol{A} * \boldsymbol{B}\| = & \sqrt{\|(\boldsymbol{A} * \boldsymbol{B})^T(\boldsymbol{A} * \boldsymbol{B})\|} \\ = & \sqrt{\|(\boldsymbol{A}^T\boldsymbol{A}) \odot (\boldsymbol{B}^T\boldsymbol{B})\|} \\ \leq & \sqrt{\|\boldsymbol{A}^T\boldsymbol{A}\|} \\ = & \|\boldsymbol{A}\|.\end{aligned}$$

Substituting $\boldsymbol{A} = \phi'(\boldsymbol{W}\boldsymbol{X}^T) - \phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T)$ and $\boldsymbol{B} = \boldsymbol{X}^T$, we find

$$\begin{aligned}&\|(\phi'(\boldsymbol{W}\boldsymbol{X}^T) - \phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T)) * \boldsymbol{X}^T\| \\ &\qquad \leq \|\phi'(\boldsymbol{W}\boldsymbol{X}^T) - \phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T)\| \\ &\qquad \leq \Gamma\|(\tilde{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{X}^T\|_F \\ &\qquad \leq \Gamma\sqrt{n}\|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F.\end{aligned}$$

Secondly,

$$\|\mathcal{J}(\tilde{\boldsymbol{W}}, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}})\| = \frac{1}{\sqrt{k}}\|\phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T) * (\boldsymbol{X} - \tilde{\boldsymbol{X}})\|$$

where reusing Schur's result and boundedness of $|\phi'| \leq \Gamma$

$$\begin{aligned}\|\phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T) * (\boldsymbol{X} - \tilde{\boldsymbol{X}})\| \leq & \Gamma\sqrt{k}\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\| \\ \leq & \Gamma\sqrt{kn}\varepsilon_0.\end{aligned}$$

Combining both estimates yields

$$\|\mathcal{J}(\boldsymbol{W}, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{X}})\| \leq \Gamma\sqrt{n}\left(\frac{\|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F}{\sqrt{k}} + \varepsilon_0\right).$$

To get the result on $\|\mathcal{J}(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}_1, \tilde{\boldsymbol{W}}_2, \tilde{\boldsymbol{X}})\|$, we integrate

$$\begin{aligned}&\|\mathcal{J}(\boldsymbol{W}_1, \boldsymbol{W}_2, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}_1, \tilde{\boldsymbol{W}}_2, \tilde{\boldsymbol{X}})\| \\ &\leq \int_0^1 \Gamma\sqrt{n}\frac{\|\alpha(\tilde{\boldsymbol{W}}_1 - \boldsymbol{W}_1) + (1-\alpha)(\tilde{\boldsymbol{W}}_1 - \boldsymbol{W}_1)\|_F}{\sqrt{k}} d\alpha \\ &\quad + \int_0^1 \Gamma\sqrt{n}\varepsilon_0 d\alpha \\ &\leq \Gamma\sqrt{n}\left(\frac{\|\tilde{\boldsymbol{W}}_1 - \boldsymbol{W}_1\|_F + \|\tilde{\boldsymbol{W}}_2 - \boldsymbol{W}_2\|_F}{2\sqrt{k}} + \varepsilon_0\right).\end{aligned}$$

■

Theorem E.12 (Robustness of gradient path) *Generate samples $(\boldsymbol{x}_i, y_i)_{i=1}^n$ according to $(\rho, \varepsilon_0, \delta)$ noisy dataset model and form the concatenated input/labels $\boldsymbol{X} \in \mathbb{R}^{d\times n}, \boldsymbol{y} \in \mathbb{R}^n$. Let $\tilde{\boldsymbol{X}}$ be the clean input sample matrix obtained by mapping $\boldsymbol{x}_i$ to its associated cluster center. Set learning rate $\eta \le \frac{K}{2c_{up}n\Gamma^2\|\boldsymbol{C}\|^2}$ and maximum iterations τ_0 satisfying*

$$\eta\tau_0 = C_1\frac{K}{n\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho}).$$

where $C_1 \ge 1$ is a constant of our choice. Suppose input noise level ε_0 and number of hidden nodes obey

$$\varepsilon_0 \le \mathcal{O}(\frac{\lambda(\boldsymbol{C})}{\Gamma^2 K\log(\frac{\Gamma\sqrt{n\log K}}{\rho})})$$
$$k \ge \mathcal{O}(\Gamma^{10}\frac{K^2\|\boldsymbol{C}\|^4}{\lambda(\boldsymbol{C})^4}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^6).$$

Set $\boldsymbol{W}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Starting from $\boldsymbol{W}_0 = \tilde{\boldsymbol{W}}_0$ consider the gradient descent iterations over the losses

$$\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau) \quad \text{where}$$
$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^n (y_i - f(\boldsymbol{W}, \tilde{\boldsymbol{x}}_i))^2$$
$$\tilde{\boldsymbol{W}}_{\tau+1} = \tilde{\boldsymbol{W}}_\tau - \nabla\tilde{\mathcal{L}}(\tilde{\boldsymbol{W}}_\tau) \quad \text{where}$$
$$\tilde{\mathcal{L}}(\tilde{\boldsymbol{W}}) = \frac{1}{2}\sum_{i=1}^n (y_i - f(\tilde{\boldsymbol{W}}, \tilde{\boldsymbol{x}}_i))^2$$

Then, for all gradient descent iterations satisfying $\tau \le \tau_0$, we have that

$$\|f(\boldsymbol{W}_\tau, \boldsymbol{X}) - f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{X}})\|_{\ell_2} \le c_0\tau\eta\varepsilon_0\Gamma^3 n^{3/2}\sqrt{\log K},$$

and

$$\|\boldsymbol{W}_\tau - \tilde{\boldsymbol{W}}_\tau\|_F \le \mathcal{O}(\tau\eta\varepsilon_0\frac{\Gamma^4 Kn}{\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^2).$$

Proof Since $\tilde{\boldsymbol{W}}_\tau$ are the noiseless iterations, with probability $1 - 2K^{-100}$, the statements of Theorem E.14 hold on $\tilde{\boldsymbol{W}}_\tau$. To proceed with proof, we first introduce short hand notations. We use

$$\begin{aligned}
\boldsymbol{r}_i &= f(\boldsymbol{W}_i, \boldsymbol{X}) - \boldsymbol{y}, \quad \tilde{\boldsymbol{r}}_i = f(\tilde{\boldsymbol{W}}_i, \tilde{\boldsymbol{X}}_i) - \boldsymbol{y},\\
\mathcal{J}_i &= \mathcal{J}(\boldsymbol{W}_i, \boldsymbol{X}), \quad \mathcal{J}_{i+1,i} = \mathcal{J}(\boldsymbol{W}_{i+1}, \boldsymbol{W}_i, \boldsymbol{X}),\\
\tilde{\mathcal{J}}_i &= \mathcal{J}(\tilde{\boldsymbol{W}}_i, \tilde{\boldsymbol{X}}), \quad \tilde{\mathcal{J}}_{i+1,i} = \mathcal{J}(\tilde{\boldsymbol{W}}_{i+1}, \tilde{\boldsymbol{W}}_i, \tilde{\boldsymbol{X}}),\\
d_i &= \|\boldsymbol{W}_i - \tilde{\boldsymbol{W}}_i\|_F, \quad p_i = \|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i\|_F,\\
\beta &= \Gamma\|\boldsymbol{C}\|\sqrt{c_{up}n/K}, \quad \text{and} \quad L = \Gamma\|\boldsymbol{C}\|\sqrt{c_{up}n/Kk}.
\end{aligned}$$

Here β is the upper bound on the Jacobian spectrum and L is the spectral norm Lipschitz constant as in Theorem E.8.

Applying Lemma E.11, note that

$$\begin{aligned}
\|\mathcal{J}(\boldsymbol{W}_\tau, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{X}})\| &\le L\|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F + \Gamma\sqrt{n}\varepsilon_0\\
&\le Ld_\tau + \Gamma\sqrt{n}\varepsilon_0\\
\|\mathcal{J}(\boldsymbol{W}_{\tau+1}, \boldsymbol{W}_\tau, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}_{\tau+1}, \tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{X}})\|&\\
&\le L(d_\tau + d_{\tau+1})/2 + \Gamma\sqrt{n}\varepsilon_0.
\end{aligned}$$

Following this and using that noiseless residual is non-increasing and satisfies $\|\tilde{\boldsymbol{r}}_\tau\|_{\ell_2} \le \|\tilde{\boldsymbol{r}}_0\|_{\ell_2}$, note that parameter satisfies

$$\boldsymbol{W}_{i+1} = \boldsymbol{W}_i - \eta\mathcal{J}_i\boldsymbol{r}_i \quad , \quad \tilde{\boldsymbol{W}}_{i+1} = \tilde{\boldsymbol{W}}_i - \eta\tilde{\mathcal{J}}_i^T\tilde{\boldsymbol{r}}_i \quad \text{(E.35)}$$
$$\begin{aligned}
\|\boldsymbol{W}_{i+1} - \tilde{\boldsymbol{W}}_{i+1}\|_F &\le \|\boldsymbol{W}_i - \tilde{\boldsymbol{W}}_i\|_F + \eta\|\mathcal{J}_i - \tilde{\mathcal{J}}_i\|\|\tilde{\boldsymbol{r}}_i\|_{\ell_2}\\
&\quad + \eta\|\mathcal{J}_i\|\|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i\|_{\ell_2} \quad \text{(E.36)}
\end{aligned}$$
$$d_{i+1} \le d_i + \eta((Ld_i + \Gamma\sqrt{n}\varepsilon_0)\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} + \beta p_i), \quad \text{(E.37)}$$

and residual satisfies (using $\boldsymbol{I} \succeq \tilde{\mathcal{J}}_{i+1,i}\tilde{\mathcal{J}}_i^T/\beta^2 \succeq 0$)

$$\begin{aligned}
\boldsymbol{r}_{i+1} &= \boldsymbol{r}_i - \eta\mathcal{J}_{i+1,i}\mathcal{J}_i^T\boldsymbol{r}_i \implies\\
\boldsymbol{r}_{i+1} - \tilde{\boldsymbol{r}}_{i+1} &= (\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i) - \eta(\mathcal{J}_{i+1,i} - \tilde{\mathcal{J}}_{i+1,i})\mathcal{J}_i^T\boldsymbol{r}_i\\
&\quad - \eta\tilde{\mathcal{J}}_{i+1,i}(\mathcal{J}_i^T - \tilde{\mathcal{J}}_i^T)\boldsymbol{r}_i - \eta\tilde{\mathcal{J}}_{i+1,i}\tilde{\mathcal{J}}_i^T(\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i).\\
\boldsymbol{r}_{i+1} - \tilde{\boldsymbol{r}}_{i+1} &= (\boldsymbol{I} - \eta\tilde{\mathcal{J}}_{i+1,i}\tilde{\mathcal{J}}_i^T)(\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i)\\
&\quad - \eta(\mathcal{J}_{i+1,i} - \tilde{\mathcal{J}}_{i+1,i})\mathcal{J}_i^T\boldsymbol{r}_i\\
&\quad - \eta\tilde{\mathcal{J}}_{i+1,i}(\mathcal{J}_i^T - \tilde{\mathcal{J}}_i^T)\boldsymbol{r}_i.\\
\|\boldsymbol{r}_{i+1} - \tilde{\boldsymbol{r}}_{i+1}\|_{\ell_2} &\le \|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i\|_{\ell_2}\\
&\quad + \eta\beta\|\boldsymbol{r}_i\|_{\ell_2}(L(3d_\tau + d_{\tau+1})/2 + 2\Gamma\sqrt{n}\varepsilon_0).\\
\|\boldsymbol{r}_{i+1} - \tilde{\boldsymbol{r}}_{i+1}\|_{\ell_2} &\le \|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i\|_{\ell_2}\\
&\quad + \frac{\eta\beta L}{2}(\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} + p_i)(3d_\tau + d_{\tau+1})\\
&\quad + 2\Gamma\sqrt{n}\eta\beta(\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} + p_i)\varepsilon_0. \quad \text{(E.38)}
\end{aligned}$$

where we used $\|\boldsymbol{r}_i\|_{\ell_2} \le p_i + \|\tilde{\boldsymbol{r}}_0\|_{\ell_2}$ and $\|(\boldsymbol{I} - \eta\tilde{\mathcal{J}}_{i+1,i}\tilde{\mathcal{J}}_i^T)\boldsymbol{v}\|_{\ell_2} \le \|\boldsymbol{v}\|_{\ell_2}$ which follows from (E.23). This implies

$$p_{i+1} \le p_i + \eta\beta(\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} + p_i)(L(3d_\tau + d_{\tau+1})/2 + 2\Gamma\sqrt{n}\varepsilon_0). \quad \text{(E.39)}$$

Finalizing proof: Next, using Lemma E.9, we have $\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} \le \Theta := C_0\Gamma\sqrt{n\log K}$. We claim that if

$$\boxed{\varepsilon_0 \le \mathcal{O}(\frac{1}{\tau_0\eta\Gamma^2 n}) \le \frac{1}{8\tau_0\eta\beta\Gamma\sqrt{n}}}$$
$$\boxed{L \le \frac{2}{5\tau_0\eta\Theta(1 + 8\eta\tau_0\beta^2)} \le \frac{1}{30(\tau_0\eta\beta)^2\Theta}}$$

(where we used $\eta\tau_0\beta^2 \ge 1$), for all $t \le \tau_0$, we have that

$$\begin{aligned}
p_t &\le 8t\eta\Gamma\sqrt{n}\varepsilon_0\Theta\beta \le \Theta\\
d_t &\le 2t\eta\Gamma\sqrt{n}\varepsilon_0\Theta(1 + 8\eta\tau_0\beta^2). \quad \text{(E.40)}
\end{aligned}$$

The proof is by induction. Suppose it holds until $t \le \tau_0 - 1$. At $t+1$, via (E.37) we have that

$$\begin{aligned}\frac{d_{t+1}-d_t}{\eta} \le& L d_t \Theta + \Gamma\sqrt{n}\varepsilon_0\Theta + 8\tau_0\eta\beta^2\Gamma\sqrt{n}\varepsilon_0\Theta \\ \overset{?}{\le}& 2\Gamma\sqrt{n}\varepsilon_0\Theta(1+8\eta\tau_0\beta^2).\end{aligned}$$

Right hand side holds since $L \le \frac{1}{2\eta\tau_0\Theta}$. This establishes the induction for d_{t+1}.

Next, we show the induction on p_t. Observe that $3d_t + d_{t+1} \le 10\tau_0\eta\Gamma\sqrt{n}\varepsilon_0\Theta(1+8\eta\tau_0\beta^2)$. Following (E.39) and using $p_t \le \Theta$, we need

$$\begin{aligned}\frac{p_{t+1}-p_t}{\eta} \le& \beta\Theta(L(3d_\tau + d_{\tau+1}) + 4\Gamma\sqrt{n}\varepsilon_0) \\ \overset{?}{\le}& 8\Gamma\sqrt{n}\varepsilon_0\Theta\beta \\ \iff& \\ L(3d_\tau + d_{\tau+1}) + 4\Gamma\sqrt{n}\varepsilon_0 \overset{?}{\le}& 8\Gamma\sqrt{n}\varepsilon_0 \iff \\ L(3d_\tau + d_{\tau+1}) \overset{?}{\le}& 4\Gamma\sqrt{n}\varepsilon_0 \iff \\ 10L\tau_0\eta(1+8\eta\tau_0\beta^2)\Theta \overset{?}{\le}& 4 \iff \\ L \overset{?}{\le}& \frac{2}{5\tau_0\eta(1+8\eta\tau_0\beta^2)\Theta}.\end{aligned}$$

Concluding the induction since L satisfies the final line. Consequently, for all $0 \le t \le \tau_0$, we have that

$$p_t \le 8t\eta\Gamma\sqrt{n}\varepsilon_0\Theta\beta = c_0 t\eta\varepsilon_0\Gamma^3 n^{3/2}\sqrt{\log K}.$$

Next, note that, condition on L is implied by

$$\begin{aligned}k \ge& 1000\Gamma^2 n(\tau_0\eta\beta)^4\Theta^2 \\ =& \mathcal{O}(\Gamma^4 n\frac{K^4}{n^4\lambda(\boldsymbol{C})^4}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^4 \\ & \cdot\left(\|\boldsymbol{C}\|\Gamma\sqrt{n/K}\right)^4(\Gamma\sqrt{n\log K})^2) \\ =& \mathcal{O}(\Gamma^{10}\frac{K^2\|\boldsymbol{C}\|^4}{\lambda(\boldsymbol{C})^4}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^4\log^2(K))\end{aligned}$$

which is implied by $k \ge \mathcal{O}(\Gamma^{10}\frac{K^2\|\boldsymbol{C}\|^4}{\lambda(\boldsymbol{C})^4}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^6)$.

Finally, following (E.40), distance satisfies

$$\begin{aligned}d_t \le& 20t\eta^2\tau_0\Gamma\sqrt{n}\varepsilon_0\Theta\beta^2 \\ \le& \mathcal{O}(t\eta\varepsilon_0\frac{\Gamma^4 Kn}{\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^2).\end{aligned}$$

■

E.3.1. Completing the Proof of Theorem 2.2

Theorem 2.2 is obtained by the theorem below when we ignore the log terms, and treating Γ, $\lambda(\boldsymbol{C})$ as constants. We also plug in $\eta = \frac{K}{2c_{up}n\Gamma^2\|\boldsymbol{C}\|^2}$.

Theorem E.13 (Training neural nets with corrupted labels) *Let $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ be an $(s, \varepsilon_0, \delta)$ clusterable noisy dataset as described in Definition 1.2. Let $\{\tilde{y}_i\}_{i=1}^n$ be the corresponding noiseless labels. Suppose $|\phi(0)|, |\phi'|, |\phi''| \le \Gamma$ for some $\Gamma \ge 1$, input noise and the number of hidden nodes satisfy*

$$\begin{aligned}\varepsilon_0 \le& \mathcal{O}(\frac{\lambda(\boldsymbol{C})}{\Gamma^2 K\log(\frac{\Gamma\sqrt{n\log K}}{\rho})}) \\ k \ge& \mathcal{O}(\Gamma^{10}\frac{K^2\|\boldsymbol{C}\|^4}{\lambda(\boldsymbol{C})^4}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^6).\end{aligned}$$

where $\boldsymbol{C} \in \mathbb{R}^{K\times d}$ is the matrix of cluster centers. Set learning rate $\eta \le \frac{K}{2c_{up}n\Gamma^2\|\boldsymbol{C}\|^2}$ and randomly initialize $\boldsymbol{W}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. With probability $1 - 3/K^{100}$, after $\tau = \mathcal{O}(\frac{K}{\eta n\lambda(\boldsymbol{C})})\log(\frac{\Gamma\sqrt{n\log K}}{\rho})$ iterations, for all $1 \le i \le n$, we have that

- *The per sample normalized ℓ_2 norm bound satisfies*

$$\begin{aligned}\frac{\|f(\boldsymbol{W}_\tau, \boldsymbol{X}) - \tilde{\boldsymbol{y}}\|_{\ell_2}}{\sqrt{n}} \le& c\frac{\varepsilon_0\Gamma^3 K\sqrt{\log K}}{\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho}) \\ &+ 4\rho.\end{aligned}$$

- *Suppose $\rho \le \delta/8$. Denote the total number of prediction errors with respect to true labels (i.e. not satisfying (E.46)) by $err(\boldsymbol{W})$. With same probability, $err(\boldsymbol{W}_\tau)$ obeys*

$$\frac{err(\boldsymbol{W}_\tau)}{n} \le c\frac{\varepsilon_0 K}{\delta}\frac{\Gamma^3\sqrt{\log K}}{\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho}).$$

- *Suppose $\rho \le \delta/8$ and $\varepsilon_0 \le c'\frac{\delta\lambda(\boldsymbol{C})^2}{\Gamma^5 K^2\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3}$, then, $\boldsymbol{W}_\tau$ assigns all input samples $\boldsymbol{x}_i$ to correct ground truth labels $\tilde{y}_i$ i.e. (E.46) holds for all $1 \le i \le n$.*

- *Finally, for any iteration count $0 \le t \le \tau$ the total distance to initialization is bounded as*

$$\begin{aligned}\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \le& \mathcal{O}(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}} \\ &+ t\eta\varepsilon_0\frac{\Gamma^4 Kn}{\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^2).\end{aligned} \tag{E.41}$$

Proof Note that proposed number of iterations τ is set so that it is large enough for Theorem E.14 to achieve small error in the clean input model ($\varepsilon_0 = 0$) and it is small enough so that Theorem E.12 is applicable. In light of Theorems E.12 and E.14 consider two gradient descent iterations starting from $\boldsymbol{W}_0$ where one uses clean dataset (as if input vectors are perfectly cluster centers) $\tilde{\boldsymbol{X}}$ and other uses the

original dataset $\boldsymbol{X}$. Denote the prediction residual vectors of the noiseless and original problems at time τ with respect true ground truth labels $\tilde{\boldsymbol{y}}$ by $\tilde{\boldsymbol{r}}_\tau = f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{X}}) - \tilde{\boldsymbol{y}}$ and $\boldsymbol{r}_\tau = f(\boldsymbol{W}_\tau, \boldsymbol{X}) - \tilde{\boldsymbol{y}}$ respectively. Applying Theorems E.12 and E.14, under the stated conditions, we have that

$$\|\tilde{\boldsymbol{r}}_\tau\|_{\ell_\infty} \le 4\rho \quad \text{and} \tag{E.42}$$

$$\|\boldsymbol{r}_\tau - \tilde{\boldsymbol{r}}_\tau\|_{\ell_2} \le c\varepsilon_0 \frac{K}{n\lambda(\boldsymbol{C})} \log(\frac{\Gamma\sqrt{n\log K}}{\rho})\Gamma^3 n^{3/2}\sqrt{\log K} \tag{E.43}$$

$$= c\frac{\varepsilon_0\Gamma^3 K\sqrt{n\log K}}{\lambda(\boldsymbol{C})} \log(\frac{\Gamma\sqrt{n\log K}}{\rho}) \tag{E.44}$$

First statement: The latter two results imply the ℓ_2 error bounds on $\boldsymbol{r}_\tau = f(\boldsymbol{W}_\tau, \boldsymbol{X}) - \tilde{\boldsymbol{y}}$.

Second statement: To assess the classification rate we count the number of entries of $\boldsymbol{r}_\tau = f(\boldsymbol{W}_\tau, \boldsymbol{X}) - \tilde{\boldsymbol{y}}$ that is larger than the class margin $\delta/2$ in absolute value. Suppose $\rho \le \delta/8$. Let $\mathcal{I}$ be the set of entries obeying this. For $i \in \mathcal{I}$ using $\|\tilde{\boldsymbol{r}}_\tau\|_{\ell_\infty} \le 4\rho \le \delta/4$, we have

$$\begin{aligned} |r_{\tau,i}| \ge \delta/2 &\implies |r_{\tau,i}| + |r_{\tau,i} - \bar{r}_{\tau,i}| \ge \delta/2 \\ &\implies |r_{\tau,i} - \bar{r}_{\tau,i}| \ge \delta/4. \end{aligned}$$

Consequently, we find that

$$\|\boldsymbol{r}_\tau - \bar{\boldsymbol{r}}_\tau\|_{\ell_1} \ge |\mathcal{I}|\delta/4.$$

Converting ℓ_2 upper bound on the left hand side to ℓ_1, we obtain

$$c\sqrt{n}\frac{\varepsilon_0\Gamma^3 K\sqrt{n\log K}}{\lambda(\boldsymbol{C})} \log(\frac{\Gamma\sqrt{n\log K}}{\rho}) \ge |\mathcal{I}|\delta/4.$$

Hence, the total number of errors is at most

$$|\mathcal{I}| \le c'\frac{\varepsilon_0 nK}{\delta}\frac{\Gamma^3\sqrt{\log K}}{\lambda(\boldsymbol{C})} \log(\frac{\Gamma\sqrt{n\log K}}{\rho})$$

Third statement – Showing zero error: Pick an input sample $\boldsymbol{x}$ from dataset and its clean version $\tilde{\boldsymbol{x}}$. We will argue that $f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{x}})$ is smaller than $\delta/4$ when ε_0 is small enough. We again write

$$\begin{aligned} |f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{x}})| \le &|f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x})| \\ &+ |f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{x}})| \end{aligned}$$

The first term can be bounded via

$$\begin{aligned} |f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x})| =& |\boldsymbol{v}^T\phi(\boldsymbol{W}_\tau\boldsymbol{x}) - \boldsymbol{v}^T\phi(\tilde{\boldsymbol{W}}_\tau\boldsymbol{x})| \\ \le& \|\boldsymbol{v}\|_{\ell_2}\|\phi(\boldsymbol{W}_\tau\boldsymbol{x}) - \phi(\tilde{\boldsymbol{W}}_\tau\boldsymbol{x})\|_{\ell_2} \\ \le& \Gamma\|\boldsymbol{W}_\tau - \tilde{\boldsymbol{W}}_\tau\|_F \\ \le& \mathcal{O}(\varepsilon_0\frac{\Gamma^5 K^2}{\lambda(\boldsymbol{C})^2}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3) \end{aligned}$$

Next, we need to bound

$$|f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{x}})| \le |\boldsymbol{v}^T\phi(\tilde{\boldsymbol{W}}_\tau\boldsymbol{x}) - \boldsymbol{v}^T\phi(\tilde{\boldsymbol{W}}_\tau\tilde{\boldsymbol{x}})| \tag{E.45}$$

where $\|\tilde{\boldsymbol{W}}_\tau - \boldsymbol{W}_0\|_F \le \mathcal{O}(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}})$, $\|\boldsymbol{x} - \tilde{\boldsymbol{x}}\|_{\ell_2} \le \varepsilon_0$ and $\boldsymbol{W}_0 \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \boldsymbol{I})$. Consequently, using by assumption we have

$$k \ge \mathcal{O}(\|\tilde{\boldsymbol{W}} - \boldsymbol{W}_0\|_F^2) = \mathcal{O}(\Gamma^2\frac{K\log K}{\lambda(\boldsymbol{C})}),$$

and applying an argument similar to Theorem E.15 (detailed in Appendix G), with probability at $1 - 1/n^{100}$, we find that

$$\begin{aligned} |f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{x}})| &\le C'\Gamma\varepsilon_0(\|\tilde{\boldsymbol{W}}_\tau - \boldsymbol{W}_0\|_F + \sqrt{\log n}), \\ &\le C\Gamma\varepsilon_0(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}} + \sqrt{\log n}). \end{aligned}$$

Combining the two bounds above we get

$$\begin{aligned} |f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{x}})| \le& \varepsilon_0\mathcal{O}(\frac{\Gamma^5 K^2}{\lambda(\boldsymbol{C})^2}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3 \\ &+ \Gamma(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}} + \sqrt{\log n})) \\ \le& \varepsilon_0\mathcal{O}(\frac{\Gamma^5 K^2}{\lambda(\boldsymbol{C})^2}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3). \end{aligned}$$

Hence, if $\varepsilon_0 \le c'\frac{\delta\lambda(\boldsymbol{C})^2}{\Gamma^5 K^2\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3}$, we obtain that, for all $1 \le i \le n$,

$$\begin{aligned} |f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) - \tilde{y}_i| <& |f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{x}}_i) - f(\boldsymbol{W}_\tau, \boldsymbol{x}_i)| \\ &+ |f(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{x}}_i) - \tilde{y}_i| \tilde{y}_i| \le 4\rho + \frac{\delta}{4}. \end{aligned}$$

If $\rho \le \delta/8$, we obtain

$$|f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) - \tilde{y}_i| < \delta/2$$

hence, $\boldsymbol{W}_\tau$ outputs the correct decision for all samples.

Fourth statement – Distance: This follows from the triangle inequality

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \le \|\boldsymbol{W}_\tau - \tilde{\boldsymbol{W}}_\tau\|_F + \|\tilde{\boldsymbol{W}}_\tau - \boldsymbol{W}_0\|_F$$

We have that right hand side terms are at most $\mathcal{O}(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}})$ and $\mathcal{O}(t\eta\varepsilon_0\frac{\Gamma^4 Kn}{\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^2)$ from Theorems E.12 and E.14 respectively. This implies (E.41). ■

Before we end this section we would like to note that in the limit of $\epsilon_0 \to 0$ where the input data set is perfectly clustered one can improve the amount of overparamterization. Indeed, the result above is obtained via a perturbation argument from this more refined result stated below.

Theorem E.14 (Training with perfectly clustered data) *Consier the setting and assumptions of Theorem E.14 with $\epsilon_0 = 0$. Starting from an initial weight matrix $\boldsymbol{W}_0$ selected at random with i.i.d. $\mathcal{N}(0,1)$ entries we run Gradient Descent (GD) updates of the form $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ on the least-squares loss (1.3) with step size $\eta \le \frac{K}{2c_{up}n\Gamma^2\|\boldsymbol{C}\|^2}$. Furthermore, assume the number of parameters obey*

$$kd \ge C\Gamma^4\kappa^2(\boldsymbol{C})K^2,$$

with $\kappa(\boldsymbol{C})$ the neural net cluster condition number per Definition 2.1. Then, with probability at least $1 - 2/K^{100}$ over randomly initialized $\boldsymbol{W}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, the iterates $\boldsymbol{W}_\tau$ obey the following properties.

- *The distance to initial point $\boldsymbol{W}_0$ is upper bounded by*

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \le c\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}}.$$

- *After $\tau \ge \tau_0 := c\frac{K}{\eta n\lambda(\boldsymbol{C})}\log\left(\frac{\Gamma\sqrt{n\log K}}{\rho}\right)$ iterations, the entrywise predictions of the learned network with respect to the* ground truth labels *$\{\widetilde{y}_i\}_{i=1}^n$ satisfy*

$$|f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) - \widetilde{y}_i| \le 4\rho,$$

for all $1 \le i \le n$. Furthermore, if the noise level ρ obeys $\rho \le \delta/8$ the network predicts the correct label for all samples i.e.

$$\arg\min_{\alpha_\ell : 1\le\ell\le K} |f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) - \alpha_\ell| = \widetilde{y}_i \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{E.46}$$

This result shows that in the limit $\epsilon_0 \to 0$ where the data points are perfectly clustered, the required amount of overparameterization can be reduced from $kd \gtrsim K^4$ to $kd \gtrsim K^2$. In this sense this can be thought of a nontrivial analogue of (5) where the number of data points are replaced with the number of clusters and the condition number of the data points is replaced with a cluster condition number. This can be interpreted as ensuring that the network has enough capacity to fit the cluster centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$ and the associated true labels. Interestingly, the robustness benefits continue to hold in this case. However, in this perfectly clustered scenario there is no need for early stopping and a robust network is trained as soon as the number of iterations are sufficiently large. Infact, in this case given the clustered nature of the input data the network never overfits to the corrupted data even after many iterations.

E.4. To (over)fit to corrupted labels requires straying far from initialization

In this section we wish to provide further insight into why early stopping enables robustness and generalizable solutions. Our main insight is that while a neural network maybe expressive enough to fit a corrupted dataset, the model has to travel a longer distance from the point of initialization as a function of the distance from the cluster centers ε_0 and the amount of corruption. We formalize this idea as follows. Suppose

1. two input points are close to each other (e.g. they are from the same cluster),
2. but their labels are different, hence the network has to map them to distant outputs.

Then, the network has to be large enough so that it can amplify the small input difference to create a large output difference. Our first result formalizes this for a randomly initialized network. Our random initialization picks $\boldsymbol{W}$ with i.i.d. standard normal entries which ensures that the network is isometric i.e. given input $\boldsymbol{x}$, $\mathbb{E}[f(\boldsymbol{W}, \boldsymbol{x})^2] = \mathcal{O}(\|\boldsymbol{x}\|_{\ell_2}^2)$.

Theorem E.15 *Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$ be two vectors with unit Euclidean norm obeying $\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2} \le \epsilon_0$. Let $f(\boldsymbol{W}, \boldsymbol{x}) = \boldsymbol{v}^T\phi(\boldsymbol{W}\boldsymbol{x})$ where $\boldsymbol{v}$ is fixed, $\boldsymbol{W} \in \mathbb{R}^{k\times d}$, and $k \ge cd$ with $c > 0$ a fixed constant. Assume $|\phi'|, |\phi''| \le \Gamma$. Let y_1 and y_2 be two scalars satisfying $|y_2 - y_1| \ge \delta$. Suppose $\boldsymbol{W}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Then, with probability at least $1 - 2e^{-(k+d)} - 2e^{-\frac{t^2}{2}}$, for any $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ such that $\|\boldsymbol{W} - \boldsymbol{W}_0\|_F \le c\sqrt{k}$ and*

$$f(\boldsymbol{W}, \boldsymbol{x}_1) = y_1 \quad \text{and} \quad f(\boldsymbol{W}, \boldsymbol{x}_2) = y_2,$$

holds, we have

$$\|\boldsymbol{W} - \boldsymbol{W}_0\| \ge \frac{\delta}{C\Gamma\varepsilon_0} - \frac{t}{1000}.$$

In words, this result shows that in order to fit to a data set with a *single corrupted label*, a randomly initialized network has to traverse a distance of at least δ/ε_0. The next lemma clarifies the role of the corruption amount s and shows that more label corruption within a fixed class requires a model with a larger norm in order to fit the labels. For this result we consider a randomized model with ε_0^2 input noise variance.

Lemma E.16 *Let $\boldsymbol{c} \in \mathbb{R}^d$ be a cluster center. Consider $2s$ data points $\{\boldsymbol{x}_i\}_{i=1}^s$ and $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^s$ in $\mathbb{R}^d$ generated i.i.d. around $\boldsymbol{c}$ according to the following distribution*

$$\boldsymbol{c} + \boldsymbol{g} \quad \text{with} \quad \boldsymbol{g} \sim \mathcal{N}(0, \frac{\varepsilon_0^2}{d}\boldsymbol{I}_d).$$

Assign $\{\boldsymbol{x}_i\}_{i=1}^s$ with labels $y_i = y$ and $\{\widetilde{\boldsymbol{x}}_i\}_{i=1}^s$ with labels $\widetilde{y}_i = \widetilde{y}$ and assume these two labels are δ separated i.e. $|y - \widetilde{y}| \ge \delta$. Also suppose $s \le d$ and $|\phi'| \le \Gamma$. Then, any $\boldsymbol{W} \in \mathbb{R}^{k\times d}$ satisfying

$$f(\boldsymbol{W}, \boldsymbol{x}_i) = y_i \quad \text{and} \quad f(\boldsymbol{W}, \widetilde{\boldsymbol{x}}_i) = \widetilde{y}_i \quad \text{for} \quad i = 1, \ldots, s,$$

obeys $\|\boldsymbol{W}\|_F \ge \frac{\sqrt{s}\delta}{5\Gamma\varepsilon_0}$ with probability at least $1 - e^{-d/2}$.

Unlike Theorem E.15 this result lower bounds the network norm in lieu of the distance to the initialization $\boldsymbol{W}_0$. However, using the triangular inequality we can in turn get a guarantee on the distance from initialization $\boldsymbol{W}_0$ via triangle inequality as long as $\|\boldsymbol{W}_0\|_F \lesssim \mathcal{O}(\sqrt{s}\delta/\varepsilon_0)$ (e.g. by choosing a small ε_0).

The above Theorem implies that the model has to traverse a distance of at least

$$\|\boldsymbol{W}_\tau - \boldsymbol{W}_0\|_F \gtrsim \sqrt{\frac{\rho n}{K}}\frac{\delta}{\varepsilon_0},$$

to perfectly fit corrupted labels. In contrast, we note that the conclusions of the upper bound in Theorem 2.2 show that to be able to fit to the uncorrupted true labels the distance to initialization grows at most by $\tau\varepsilon_0$ after τ iterates. This demonstrates that there is a gap in the required distance to initialization for *fitting enough to generalize* and *overfitting*. To sum up, our results highlight that, one can find a network with good generalization capabilities and robustness to label corruption within a small neighborhood of the initialization and that the size of this neighborhood is independent of the corruption. However, to fit to the corrupted labels, one has to travel much more, increasing the search space and likely decreasing generalization ability. Thus, early stopping can enable robustness without overfitting by restricting the distance to the initialization.

F. Proof of Lemma E.16

Create two matrices $\boldsymbol{X} \in \mathbb{R}^{s\times d}$ and $\tilde{\boldsymbol{X}} \in \mathbb{R}^{s\times d}$ by concatenating the input samples. Note that the matrix $\boldsymbol{X} - \tilde{\boldsymbol{X}}$ has i.i.d. $\mathcal{N}(0, 2\varepsilon_0^2/d)$ entries. Thus, using standard results regarding the concentration of the spectral norm with probability at least $1 - e^{-d/2}$, we have

$$\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\| \le \sqrt{2}\left(\sqrt{\frac{s}{d}} + 2\right)\varepsilon_0 \le 5\varepsilon_0.$$

Define the vectors $\boldsymbol{y}, \tilde{\boldsymbol{y}} \in \mathbb{R}^s$ with entries given by y_i and $\tilde{y}_i$, respectively. Suppose $\boldsymbol{W}$ fits these labels perfectly. Using the fact that $\|\boldsymbol{v}\|_{\ell_2} = 1$, we can conclude that

$$\begin{aligned}\sqrt{s}\delta \le \|\boldsymbol{y} - \tilde{\boldsymbol{y}}\|_{\ell_2} &= \|f(\boldsymbol{W}, \boldsymbol{X}) - f(\boldsymbol{W}, \tilde{\boldsymbol{X}})\|_{\ell_2},\\ &= \|\boldsymbol{v}^T(\phi(\boldsymbol{W}\boldsymbol{X}) - \phi(\boldsymbol{W}\tilde{\boldsymbol{X}}))\|_{\ell_2},\\ &\le \Gamma\|\boldsymbol{v}\|_{\ell_2}\|\boldsymbol{W}(\boldsymbol{X} - \tilde{\boldsymbol{X}})\|_F,\\ &\le \Gamma\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\|\|\boldsymbol{W}\|_F \le 5\Gamma\varepsilon_0\|\boldsymbol{W}\|_F.\end{aligned}$$

This implies the desired lower bound on $\|\boldsymbol{W}\|_F$.

G. Single label perturbation

Note that

$$\begin{aligned}&|f(\boldsymbol{W}, \boldsymbol{x}) - f(\boldsymbol{W}, \widetilde{\boldsymbol{x}})|\\ &= \left|\boldsymbol{v}^T\left(\phi(\boldsymbol{W}\boldsymbol{x}) - \phi(\boldsymbol{W}\widetilde{\boldsymbol{x}})\right)\right|\\ &\le \left|\boldsymbol{v}^T\left(\phi(\boldsymbol{W}\boldsymbol{x}) - \phi(\boldsymbol{W}\widetilde{\boldsymbol{x}})\right) - \boldsymbol{v}^T\left(\phi(\boldsymbol{W}_0\boldsymbol{x}) - \phi(\boldsymbol{W}_0\widetilde{\boldsymbol{x}})\right)\right|\\ &\quad + \left|\boldsymbol{v}^T\left(\phi(\boldsymbol{W}_0\boldsymbol{x}) - \phi(\boldsymbol{W}_0\widetilde{\boldsymbol{x}})\right)\right|\end{aligned}$$

To continue note that by the general mean value theorem there exists a point $(\overline{\boldsymbol{W}}, \overline{\boldsymbol{x}})$ in the square $(\boldsymbol{W}_0, \boldsymbol{x}), (\boldsymbol{W}_0, \widetilde{\boldsymbol{x}}), (\boldsymbol{W}, \boldsymbol{x})$, and $(\boldsymbol{W}, \widetilde{\boldsymbol{x}})$ such that

$$\begin{aligned}&\left(f(\boldsymbol{W}, \boldsymbol{x}) - f(\boldsymbol{W}_0, \boldsymbol{x})\right) - \left(f(\boldsymbol{W}, \widetilde{\boldsymbol{x}}) - f(\boldsymbol{W}_0, \widetilde{\boldsymbol{x}})\right)\\ &\quad = (\boldsymbol{x} - \widetilde{\boldsymbol{x}})^T\overline{\boldsymbol{W}}^T\operatorname{diag}(\boldsymbol{v})\operatorname{diag}\left(\phi''(\overline{\boldsymbol{W}}\overline{\boldsymbol{x}})\right)(\boldsymbol{W} - \boldsymbol{W}_0)\overline{\boldsymbol{x}}\\ &\qquad + \boldsymbol{v}^T\operatorname{diag}\left(\phi'\left(\overline{\boldsymbol{W}}\overline{\boldsymbol{x}}\right)\right)(\boldsymbol{W} - \boldsymbol{W}_0)(\boldsymbol{x} - \widetilde{\boldsymbol{x}})\end{aligned}$$

Using the above we have that

$$\begin{aligned}&\left|\left(f(\boldsymbol{W}, \boldsymbol{x}) - f(\boldsymbol{W}_0, \boldsymbol{x})\right) - \left(f(\boldsymbol{W}, \widetilde{\boldsymbol{x}}) - f(\boldsymbol{W}_0, \widetilde{\boldsymbol{x}})\right)\right|\\ &\overset{(a)}{\le} \left|(\boldsymbol{x} - \widetilde{\boldsymbol{x}})^T\overline{\boldsymbol{W}}^T\operatorname{diag}(\boldsymbol{v})\operatorname{diag}\left(\phi''(\overline{\boldsymbol{W}}\overline{\boldsymbol{x}})\right)(\boldsymbol{W} - \boldsymbol{W}_0)\overline{\boldsymbol{x}}\right|\\ &\quad + \left|\boldsymbol{v}^T\operatorname{diag}\left(\phi'\left(\overline{\boldsymbol{W}}\overline{\boldsymbol{x}}\right)\right)(\boldsymbol{W} - \boldsymbol{W}_0)(\boldsymbol{x} - \widetilde{\boldsymbol{x}})\right|\\ &\overset{(b)}{\le} \left(\|\boldsymbol{v}\|_{\ell_\infty}\|\overline{\boldsymbol{x}}\|_{\ell_2}\left\|\overline{\boldsymbol{W}}\right\| + \|\boldsymbol{v}\|_{\ell_2}\right)\Gamma\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W} - \boldsymbol{W}_0\|\\ &\overset{(c)}{\le} \left(\frac{1}{\sqrt{k}}\|\overline{\boldsymbol{x}}\|_{\ell_2}\left\|\overline{\boldsymbol{W}}\right\| + 1\right)\Gamma\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W} - \boldsymbol{W}_0\|\\ &\overset{(d)}{\le} \left(\frac{1}{\sqrt{k}}\left\|\overline{\boldsymbol{W}}\right\| + 1\right)\Gamma\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W} - \boldsymbol{W}_0\|\\ &\overset{(e)}{\le} \Gamma\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W} - \boldsymbol{W}_0\|\\ &\qquad \cdot\left(\frac{1}{\sqrt{k}}\|\boldsymbol{W}_0\| + \frac{1}{\sqrt{k}}\left\|\overline{\boldsymbol{W}} - \boldsymbol{W}_0\right\| + 1\right)\\ &\overset{(f)}{\le} \Gamma\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W} - \boldsymbol{W}_0\|\\ &\qquad \cdot\left(\frac{1}{\sqrt{k}}\|\boldsymbol{W}_0\| + \frac{1}{\sqrt{k}}\left\|\overline{\boldsymbol{W}} - \boldsymbol{W}_0\right\|_F + 1\right)\\ &\overset{(g)}{\le} \Gamma\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W} - \boldsymbol{W}_0\|\\ &\qquad \cdot\left(\frac{1}{\sqrt{k}}\left\|\overline{\boldsymbol{W}} - \boldsymbol{W}_0\right\|_F + 3 + 2\sqrt{\frac{d}{k}}\right)\\ &\overset{(h)}{\le} C\Gamma\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W} - \boldsymbol{W}_0\| \qquad \text{(G.1)}\end{aligned}$$

Here, (a) follows from the triangle inequality, (b) from simple algebraic manipulations along with the fact that $|\phi'(z)| \le \Gamma$ and $|\phi''(z)| \le \Gamma$, (c) from the fact that $\boldsymbol{v}_\ell = \pm\frac{1}{\sqrt{k}}$, (d) from $\|\boldsymbol{x}\|_{\ell_2} = \|\widetilde{\boldsymbol{x}}\|_{\ell_2} = 1$ which implies $\|\overline{\boldsymbol{x}}\|_{\ell_2} \le 1$, (e) from triangular inequality, (f) from the fact that Frobenius norm dominates the spectral norm, (g) from the fact that with probability at least $1 - 2e^{-(d+k)}$, $\|\boldsymbol{W}_0\| \le 2(\sqrt{k} + \sqrt{d})$, and

(h) from the fact that $\left\|\overline{\boldsymbol{W}} - \boldsymbol{W}_0\right\| \le \|\boldsymbol{W} - \boldsymbol{W}_0\|_F \le \widetilde{c}\sqrt{k}$ and $k \ge cd$.

Next we note that for a Gaussian random vector $\boldsymbol{g} \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{I}_d)$ we have

$$\begin{aligned}
&\|\phi(\boldsymbol{g}^T\boldsymbol{x}) - \phi(\boldsymbol{g}^T\widetilde{\boldsymbol{x}})\|_{\psi_2} \\
&\qquad = \|\phi(\boldsymbol{g}^T\boldsymbol{x}) - \phi(\boldsymbol{g}^T\widetilde{\boldsymbol{x}})\|_{\psi_2} \\
&\qquad = \|\phi'\left(t\boldsymbol{g}^T\boldsymbol{x} + (1-t)\boldsymbol{g}^T\widetilde{\boldsymbol{x}}\right)\boldsymbol{g}^T(\boldsymbol{x} - \widetilde{\boldsymbol{x}})\|_{\psi_2} \\
&\qquad \le \Gamma\|\boldsymbol{g}^T(\boldsymbol{x} - \widetilde{\boldsymbol{x}})\|_{\psi_2} \\
&\qquad \le c\Gamma\left\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\right\|_{\ell_2}. \qquad \text{(G.2)}
\end{aligned}$$

Also note that

$$\begin{aligned}
f(\boldsymbol{W}_0, \boldsymbol{x}) - f(\boldsymbol{W}_0, \widetilde{\boldsymbol{x}}) =& \boldsymbol{v}^T\left(\phi\left(\boldsymbol{W}_0\boldsymbol{x}\right) - \phi\left(\boldsymbol{W}_0\widetilde{\boldsymbol{x}}\right)\right) \\
&\sim \sum_{\ell=1}^{k} \boldsymbol{v}_\ell\left(\phi(\boldsymbol{g}_\ell^T\boldsymbol{x}) - \phi(\boldsymbol{g}_\ell^T\widetilde{\boldsymbol{x}})\right)
\end{aligned}$$

where $\boldsymbol{g}_1, \boldsymbol{g}_2, \ldots, \boldsymbol{g}_k$ are i.i.d. vectors with $\mathcal{N}(0, \boldsymbol{I}_d)$ distribution. Also for $\boldsymbol{v}$ obeying $\boldsymbol{1}^T\boldsymbol{v} = 0$ this random variable has mean zero. Hence, using the fact that weighted sum of subGaussian random variables are subgaussian combined with (G.2) we conclude that $f(\boldsymbol{W}_0, \boldsymbol{x}) - f(\boldsymbol{W}_0, \widetilde{\boldsymbol{x}})$ is also subGaussian obeying $\|f(\boldsymbol{W}_0, \boldsymbol{x}) - f(\boldsymbol{W}_0, \widetilde{\boldsymbol{x}})\|_{\psi_2} \le c\Gamma\|\boldsymbol{v}\|_{\ell_2}\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2}$. Thus

$$\begin{aligned}
|f(\boldsymbol{W}_0, \boldsymbol{x}) - f(\boldsymbol{W}_0, \widetilde{\boldsymbol{x}})| \le& ct\Gamma\|\boldsymbol{v}\|_{\ell_2}\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\|_{\ell_2} \\
=& ct\Gamma\left\|\boldsymbol{x} - \widetilde{\boldsymbol{x}}\right\|_{\ell_2}, \qquad \text{(G.3)}
\end{aligned}$$

with probability at least $1 - e^{-\frac{t^2}{2}}$. Thus, using $t = 2\sqrt{\log n}$ for n data points

$$|f(\boldsymbol{W}_0, \boldsymbol{x}_i) - f(\boldsymbol{W}_0, \widetilde{\boldsymbol{x}}_i)| \le 2c\Gamma\sqrt{\log n}\left\|\boldsymbol{x}_i - \widetilde{\boldsymbol{x}}_i\right\|_{\ell_2},$$

holds for all $i = 1, 2, \ldots, n$ with probability at least

$$1 - ne^{-\frac{t^2}{2}} \ge 1 - \frac{1}{n^{100}}.$$