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

Mingchen Li
University of California
Riverside, CA

Mahdi Soltanolkotabi
University of Southern California
Los Angeles, CA

Samet Oymak
University of California
Riverside, CA

Abstract

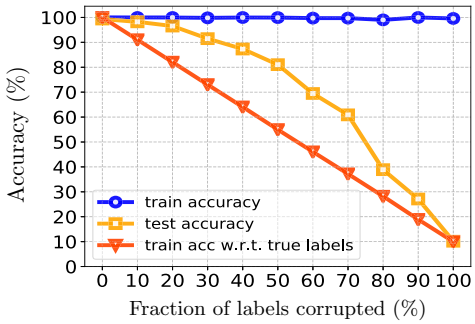
Modern neural networks are typically trained in an over-parameterized regime where the parameters of the model far exceed the size of the training data. Such neural networks in principle have the capacity to (over)fit any set of labels including significantly corrupted ones. Despite this (over)fitting capacity in this paper we demonstrate that such over-parameterized networks have an intriguing robustness capability: they are surprisingly robust to label noise when first order methods with early stopping is used to train them. This paper also takes a step towards demystifying this phenomena. Under a rich dataset model, we show that gradient descent is provably robust to noise/corruption on a constant fraction of the labels. In particular, we prove that: (i) In the first few iterations where the updates are still in the vicinity of the initialization gradient descent only fits to the correct labels essentially ignoring the noisy labels. (ii) To start to overfit to the noisy labels network must stray rather far from the initialization which can only occur after many more iterations. Together, these results show that gradient descent with early stopping is provably robust to label noise and shed light on the empirical robustness of deep networks as well as commonly adopted heuristics to prevent overfitting.

1 Introduction

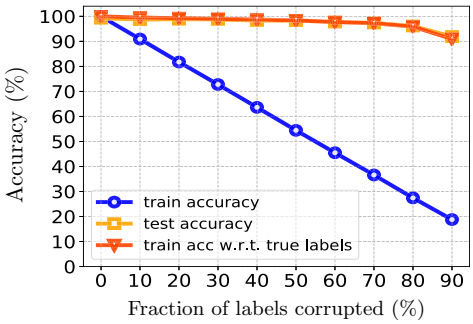
This paper focuses on an intriguing phenomena: over-parameterized neural networks are surprisingly robust to label noise when first order methods with early stopping is used to train them. To observe this phenomena consider Figure 1 where we perform experiments on the MNIST data set. Here, we corrupt 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 but 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 (see also Guan et al. (2018) and Rolnick et al. (2017) for related empirical experiments). That is, with early stopping overparameterized neural networks even correct the corrupted labels! These plots collectively demonstrate that over-parameterized neural networks when combined with early stopping have unique generalization and robustness capabilities. As we detail further in Section 3 this phenomena holds (albeit less pronounced) for richer data models and architectures.

This paper aims 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 over-parameterized learning scenarios. In particular, under a fairly expressive

Proceedings of the 23rdInternational Conference on Artificial Intelligence and Statistics (AISTATS) 2020, Palermo, Italy. PMLR: Volume 108.



(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 MNIST with various amounts of random corruption on the labels. In this architecture the convolution 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 use 50k samples for training, 10k samples for validation, and we test the performance on a 10k test dataset. We depict the training accuracy both w.r.t. the corrupted and uncorrupted labels as well as the 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.

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*.

Our theoretical results further highlight the role of *the distance between final and initial network weights* as a key feature that determines noise robustness vs. overfitting. This is inherently connected to the commonly used early stopping heuristic for DNN training as this heuristic helps avoid models that are too far from the point of initialization. In the presence of label noise, we show that gradient descent *implicitly* ignores the noisy labels as long as the model parameters remain close to the initialization. Hence, our results help explain why early stopping improves robustness and helps prevent overfitting. Under proper normalization, the required distance between the final and initial network and the predictive accuracy of the final network is independent of the size of the network such as number of hidden nodes. Our extensive numerical experiments corroborate our theory and verify the surprising robustness of DNNs to label noise. Finally, we would like to note that while our results show that solutions found by gradient descent are inherently robust to label noise, specialized techniques such as ℓ_1 penalization or sample reweighting are known to further improve robustness. Our theoretical framework may enable more rigorous understandings of the benefits of such heuristics when training overparameterized models.

1.1 Prior Art

Our work is connected to recent advances on theory for deep learning as well as heuristics and theory surrounding outlier robust optimization.

Robustness to label corruption: DNNs have the ability to fit to pure noise Zhang et al. (2016), however they are also empirically observed to be highly resilient to label noise and generalize well despite large corruption Rolnick et al. (2017). In addition to early stopping, several heuristics have been proposed to specifically deal with label noise Reed et al. (2014); Malach and Shalev-Shwartz (2017); Scott et al. (2013); Han et al. (2018); Zhang and Sabuncu (2018); Khetan et al. (2017); Basri et al. (2019); Bartlett et al. (2019). See also Frénay et al. (2014); Shen and Sanghavi (2018); Menon et al. (2018); Ren et al. (2018); Arazo et al. (2019) for additional work on dealing with label noise in classification tasks. Label noise is also connected to outlier robustness in regression which is a traditionally well-studied topic. In the context of robust regression and high-dimensional statistics, much of the focus is on regularization techniques to automatically detect and discard outliers by using tools such as ℓ_1 penalization Chen et al. (2013); Li (2013); Balakrishnan et al. (2017); Liu et al. (2018); Bhatia et al. (2015); Foygel and Mackey (2014); Candès et al. (2011). We would also like to note that there is an interesting line of work that focuses on developing robust algorithms for corruption not only in the labels but also input data Diakonikolas et al. (2018); Prasad et al. (2018); Klivans et al. (2018). Finally, noise robustness is particularly

important in safety critical domains. Noise robustness of neural nets has been empirically investigated by Hinton and coauthors in the context of automated medical diagnosis Guan et al. (2018).

Overparameterized neural networks: Intriguing properties and benefits of overparameterized networks have been the focus of a growing list of publications Zhang et al. (2016); Soltanolkotabi et al. (2018); Brutzkus et al. (2017a); Chizat and Bach (2018); Arora et al. (2018a); Ji and Telgarsky (2018); Venturi et al. (2018); Zhu et al. (2018); Soudry and Carmon (2016); Brutzkus and Globerson (2018); Azizan and Hassibi (2018); Neyshabur et al. (2018). A recent line of work Li and Liang (2018); Allen-Zhu et al. (2018a,b); Du et al. (2018b); Zou et al. (2018); Du et al. (2018a); Oymak and Soltanolkotabi (2019); Papyan (2019) shows that overparameterized neural networks can fit the data with random initialization if the number of hidden nodes are polynomially large in the size of the dataset. This line of work however is not informative about the robustness of the trained network against corrupted labels. Indeed, such theory predicts that (stochastic) gradient descent will eventually fit the corrupted labels. In contrast, our focus here is not in finding a global minima, rather a solution that is robust to label corruption. In particular, we show that with early stopping we fit to the correct labels without overfitting to the corrupted training data. Our result also differs from this line of research in another way. The key property utilized in this research area is that the Jacobian of the neural network is well-conditioned at a random initialization if the dataset is sufficiently diverse (e.g. if the points are well-separated). In contrast, in many practical settings the Jacobian is approximately low-rank. We leverage this low-rankness to prove that gradient descent is robust to label corruptions. We further utilize this to explain why neural nets can work with much smaller amounts of overparameterization where the number of parameters grow with rank rather than the sample size. Furthermore, our numerical experiments verify that the Jacobian matrix of real datasets (such as CIFAR10) indeed exhibit low-rank structure. This is related to the observations on the Hessian of deep networks which is empirically observed to be low-rank Sagun et al. (2017); Chaudhari et al. (2016); Javadi et al. (2019); Ghorbani et al. (2019). Recent papers Su and Yang (2019); Oymak et al. (2019); Rahaman et al. (2018) leverage related phenomena to prove/explain generalization and approximation ability of deep nets. More recently, Hu et al. (2019)[1] shows label noise robustness by utilizing the Rademacher complexity based generalization results of Arora et al. (2019). Also see Arora et al. (2018b); Bartlett et al. (2017); Golowich et al. (2017); Song et al. (2018); Brutzkus et al. (2017b); Belkin et al. (2018a,b); Liang and Rakhlin (2018); Oymak et al. (2019); Cao and Gu (2019); Arora et al. (2019); Ma et al. (2019); Allen-Zhu et al. (2018a) for further recent neural network generalization results. While this work does not tackle generalization in the traditional sense, we do show that solutions found by gradient descent are robust to label noise/corruption which demonstrates their predictive capabilities and in turn suggests better generalization. Finally, related to our work, the role of early-stopping in gradient descent is studied by Yao et al. (2007) in the context of function approximation via kernels.

[1]We note that the first draft of this manuscript appeared earlier than Hu et al. (2019); Su and Yang (2019); Oymak et al. (2019).

1.2 Models

We now describe the dataset model used in our theoretical results. We note that while we mainly focus on this model for exposition purposes our results holds for any data set for which the Jacobian of the network is approximately low-rank with a range that is not too spiky (See Section 4 and the supplementary for further detail). 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 Euclidean ball in $\mathbb{R}^d$. We also assume our dataset consists of $\bar{K} \leq K$ classes where each class can be composed of multiple clusters. We consider a deterministic dataset with n samples with roughly balanced clusters each consisting on the order of n/K samples.[2] Finally, while we allow for multiple classes, in our model we assume the labels are scalars and take values in $[-1, 1]$ interval. Each unit Euclidean norm $\boldsymbol{x}$ is assigned to one of these class labels as described next. We formally define our dataset model below and provide an illustration in Figure 2.

Definition 1.1 ((ε_0, δ) Clusterable dataset) *A clusterable dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ is described as follows. The input samples have unit Euclidean norm and originate from K clusters with the ℓth cluster containing n_ℓ data points where $c_{low}\frac{n}{K} \leq n_\ell \leq c_{up}\frac{n}{K}$ for some positive constants c_{low} and c_{up}. Cluster centers are unit norm vectors denoted by $\{\boldsymbol{c}_\ell\}_{\ell=1}^K \subset \mathbb{R}^d$. An input $\boldsymbol{x}$ that belong to the ℓth cluster obey $\|\boldsymbol{x} - \boldsymbol{c}_\ell\|_{\ell_2} \leq \varepsilon_0$, with ε_0 denoting the input noise level.*

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

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

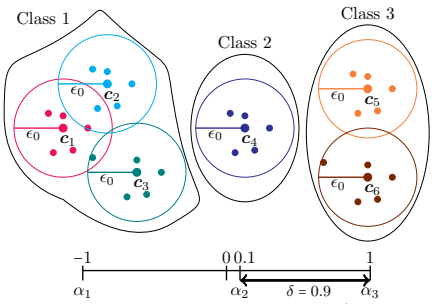


Figure 2: Visualization of the input/label samples and classes according to the clusterable model in Definition 1.1. In the depicted example there are $K = 6$ clusters, $\bar{K} = 3$ classes. In this example the number of data points is $n = 30$ with each cluster containing 5 data points. The labels 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.

$$|\alpha_r - \alpha_s| \geq \delta \quad \textit{for} \quad r \neq s, \tag{1}$$

for some separation $\delta > 0$. Any two clusters ℓ, ℓ' belonging to different classes obey $\|\boldsymbol{c}_\ell - \boldsymbol{c}_{\ell'}\|_{\ell_2} \geq 2\varepsilon_0$.

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, $\bar{K} \leq \frac{2}{\delta} + 1$. Our dataset model is inspired from mixture models and is also related to the setup of Li and Liang (2018) which provides polynomial guarantees for learning shallow networks. Next, we introduce our noisy/corrupted dataset model.

Definition 1.2 ($(\rho, \varepsilon_0, \delta)$ corrupted dataset) *A $(\rho, \varepsilon_0, \delta)$ noisy/corrupted dataset $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n$ is generated from an (ε_0, δ) clusterable dataset $\{(\boldsymbol{x}_i, \widetilde{y}_i)\}_{i=1}^n$ as follows. For each cluster $1 \leq \ell \leq K$, 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. Next we define the ground truth label function.

Definition 1.3 (Ground truth label function) *Consider the setting of Def. 1.1 with cluster centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^K \subset \mathbb{R}^d$ and class labels $\{\alpha_\ell\}_{\ell=1}^{\bar{K}}$. Define the ground truth label function $\boldsymbol{x} \mapsto \widetilde{y}(\boldsymbol{x})$ as the function that maps a point $\boldsymbol{x} \in \mathbb{R}^d$ to a class label $\{\alpha_1, \alpha_2, \ldots, \alpha_{\bar{K}}\}$ by assigning it the label corresponding to the closest cluster center. Mathematically*

$$\widetilde{y}(\boldsymbol{x}) = \textit{label of } \boldsymbol{c}_{\hat{\ell}} \quad \textit{where} \quad \hat{\ell} = \operatorname*{arg\,min}_{1 \leq \ell \leq K} \|\boldsymbol{x} - \boldsymbol{c}_\ell\|_{\ell_2}.$$

In particular, when applied to the training data it yields the ground truth labels i.e. $\tilde{y}(\boldsymbol{x}_i) = \tilde{y}_i$.

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 the exposition. 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. $|\phi'(z)|, |\phi''(z)| \leq \Gamma$ for some constant $\Gamma > 0$ for all z. The network's prediction at an input sample $\boldsymbol{x}$ is given by

$$\boldsymbol{x} \mapsto f(\boldsymbol{W}, \boldsymbol{x}) = \boldsymbol{v}^T \phi(\boldsymbol{W}\boldsymbol{x}), \tag{2}$$

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

$$\mathcal{L}(\boldsymbol{W}) = \frac{1}{2} \sum_{i=1}^n (y_i - f(\boldsymbol{W}, \boldsymbol{x}_i))^2. \tag{3}$$

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

$$\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta \nabla \mathcal{L}(\boldsymbol{W}_\tau). \tag{4}$$

2 Main Results

Our main result shows that overparameterized neural networks, when trained via gradient descent using early stopping are fairly robust to label noise. Throughout, $\|\cdot\|$ denotes the largest singular value of a given matrix. c, c_0, C, C_0 etc. represent numerical constants. 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 $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$. This definition is induced by the Neural Tangent Kernel (Jacot et al. (2018)) which provides a linearization of the network at random initialization.

Definition 2.1 (Neural Net Cluster Covariance) *Define the matrix of cluster centers*

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

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

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

Here $\odot$ *denotes the elementwise product. Also denote the minimum eigenvalue of* $\boldsymbol{\Sigma}(\boldsymbol{C})$ *by* $\lambda(\boldsymbol{C})$.

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. The more distinct the cluster centers, the larger $\lambda(\boldsymbol{C})$ is. Throughout we shall assume that $\lambda(\boldsymbol{C})$ is strictly positive. Related assumptions are empirically and theoretically studied in earlier works by Allen-Zhu et al. (2018b); Xie et al. (2016); Du et al. (2018b,a). For instance, when the cluster centers are maximally diverse e.g. uniformly at random from the unit sphere $\lambda(\boldsymbol{C})$ scales like a constant (Oymak and Soltanolkotabi (2019)). Additionally, for ReLU activation, if the cluster centers are separated by a distance $\nu > 0$, then $\lambda(\boldsymbol{C}) \geq \frac{\nu}{100K^2}$ (Zou et al. (2018); Oymak and Soltanolkotabi (2019)).

Now that we have a quantitative characterization of distinctiveness/diversity in place we are now ready to state our main result. We note that this theorem is slightly simplified by ignoring logarithmic terms and precise dependencies on Γ. The supplementary provides precise statements.

Theorem 2.2 (Main result) *Consider a dataset* $\{(\boldsymbol{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ *per Def. 1.2. Starting from an initial weight matrix* $\boldsymbol{W}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$ *entries, run gradient updates* $\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau)$ *with properly chosen constant step size* η *and assume*

$$k \gtrsim \frac{K^2\|\boldsymbol{C}\|^4}{\lambda(\boldsymbol{C})^4},$$

If $\varepsilon_0 \lesssim \delta\lambda(\boldsymbol{C})^2/K^2$ *and* $\rho \leq \delta/8$ *with high probability, after* $T \propto \frac{\|\boldsymbol{C}\|^2}{\lambda(\boldsymbol{C})}$ *iterations, the model* $\boldsymbol{W}_T$ *predicts the true label function* $\widetilde{y}(\boldsymbol{x})$ *for all input* $\boldsymbol{x} \in \mathbb{R}^d$ *that lie within* ε_0 *neighborhood of a cluster center* $\{\boldsymbol{c}_k\}_{k=1}^K$. *That is,*

$$\arg\min_{\alpha_\ell:1\leq\ell\leq\bar{K}} |f(\boldsymbol{W}_T, \boldsymbol{x}) - \alpha_\ell| = \widetilde{y}(\boldsymbol{x}). \tag{5}$$

Eq. (5) applies to all training samples. Finally, for all $0 \leq \tau \leq T$, *the distance to initialization obeys*

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

Theorem 2.2 shows that gradient descent with early stopping is robust and predicts the correct labels despite the corruption. See below for further 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 any input data. In particular, when applied to the training data (5) yields $\arg\min_{\alpha_\ell:1\leq\ell\leq\bar{K}} |f(\boldsymbol{W}_\tau, \boldsymbol{x}) - \alpha_\ell| = \widetilde{y}_i$ so that the network labels are identical to the ground truth labels completely removing the corruption on the training data. In our setup, intuitively the label gap obeys $\delta \sim \frac{1}{\bar{K}}$, hence, we prove robustness to

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

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.

Early stopping time. Only a few iterations suffice to find a good model (at most order K iterations typically $\max(1, K/d)$ modulo condition numbers).

Modest overparameterization. Our result applies as soon as the number of hidden units in the network exceeds $K^2\|\boldsymbol{C}\|^4$ which lies between K^2 and K^4 which is independent of the sample size n. This can be interpreted as network having enough capacity to fit the cluster centers $\{\boldsymbol{c}_\ell\}_{\ell=1}^K$ and their true labels. If cluster centers are incoherent (e.g. random) and $K \geq d$, the required number of parameters in the network $(k \times d)$ scales as $dK^2\|\boldsymbol{C}\|^4 \lesssim K^4$.

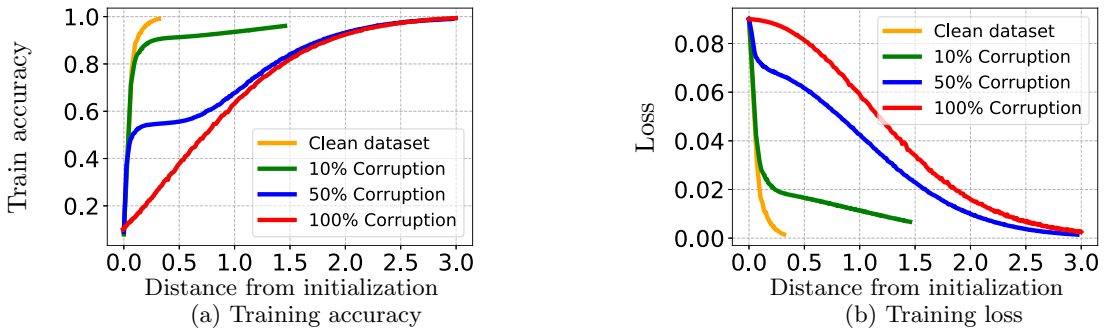
Distance from initialization. Another feature of Theorem 2.2 is that the network weights do not stray far from the initialization as the distance between the initial model and the final model (at most) grows with the square root of the number of clusters ($\sqrt{K}$). Intuitively, more clusters correspond to a richer data distribution, hence we need to travel further away to find a viable model. While our focus in this work is early stopping, the importance of distance to initialization motivates the use of ℓ_2-regularization with respect to the initial point i.e. solving the regularized empirical risk minimization

$$\boldsymbol{W}_{\text{ridge}} = \arg\min_{\boldsymbol{W}} \frac{1}{2}\sum_{i=1}^n (y_i - f(\boldsymbol{W}, \boldsymbol{x}_i))^2 + \lambda\|\boldsymbol{W} - \boldsymbol{W}_0\|_F^2,$$

where $\boldsymbol{W}_0$ is the point of initialization for the gradient based algorithm that will be used to solve above.

2.1 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 generaliz-



(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.

able 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 2.3 *Let $\boldsymbol{x}_1, \boldsymbol{x}_2 \in \mathbb{R}^d$ be two vectors with unit ℓ_2 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 $k \ge cd$ where c, c', c'' are constants and $|\phi'|, |\phi''| \le \Gamma$. 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 $1 - 2e^{-(k+d)} - 2e^{-\frac{t^2}{2}}$, for any $\boldsymbol{W}$ 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 $\|\boldsymbol{W} - \boldsymbol{W}_0\| \ge \frac{c''\delta}{\Gamma\varepsilon_0} - \frac{t}{1000}$.

In words, this result shows that in order to fit to a dataset with a *single corrupted label*, a randomly initialized network has to traverse a distance of at least δ/ε_0. The supplementary 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.

Can we really overfit to corruption? A natural question is whether early stopping is necessary i.e. can we perfectly interpolate to the corrupted dataset model of Definition 1.2. The recent works Du et al. (2018a); Allen-Zhu et al. (2018b); Oymak and Soltanolkotabi (2019) on neural net optimization answers this affirmatively. In particular, as long as no two input samples are identical, sufficiently wide neural networks trained with gradient descent can provably and perfectly interpolate a corrupted dataset.

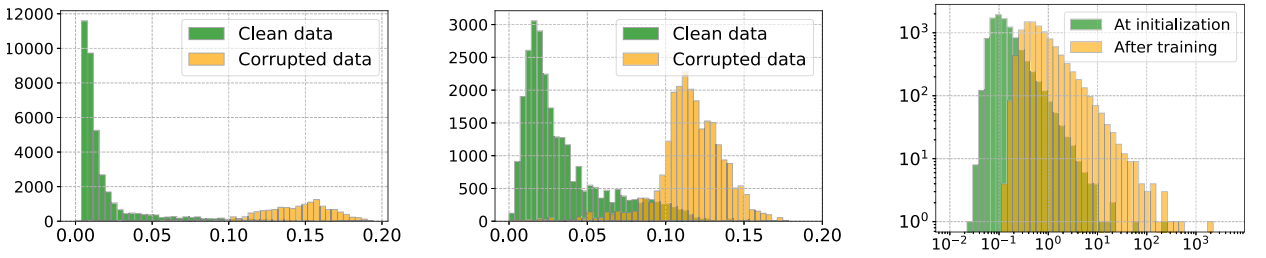
3 Numerical experiments

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[3].

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.

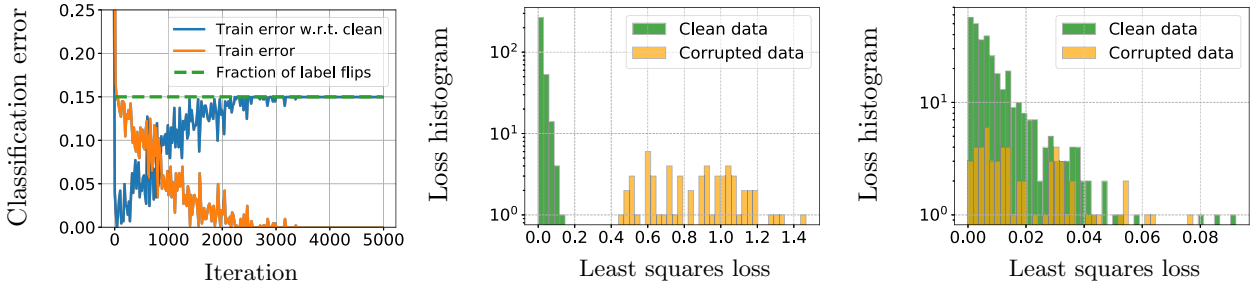
Next, we study the distribution of the individual sample losses on CIFAR-10. We conducted two experiments using Resnet-20 with least square loss. In Figure 4a and 4b we assess the noise robustness of gradient descent where we used all 50,000 samples with either 30% random corruption or 50% random corruption. The supplementary shows that when the corruption

[3] All experiments use least square loss corresponding to our theory, but we have same observation on cross entropy loss and provide figures in appendix.



(a) 30% corruption (b) 50% corruption (c) Class Airplane and Automobile

Figure 4: (a)(b) Are histograms of the least square loss of individual data points based on a model trained on 50,000 samples from CIFAR-10 with early stopping. The loss distribution of clean and corrupted data are separated but gracefully overlap as corruption increases. (c) is histogram of singular values obtained by forming the Jacobian by taking partial derivatives of class Airplane and Automobile on 10000 samples.



(a) Fraction of incorrect predictions (b) Loss histogram at iteration 80 (c) Loss histogram at iteration 4500

Figure 5: 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. In average 15% of labels actually flip which is highlighted by the dashed green line.

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

As we briefly discuss in Section 4 (see proofs in the supplementary for more extensive discussion), our technical framework utilizes the low-rank structure of the Jacobian matrix of the model. We now further investigate this hypothesis. For a binary class task, size of the Jacobian matrix is sample size (n) $\times$ total number of parameters in the model (p). The neural network model we used for CIFAR 10 has around $p = 270,000$ parameters in total. In Figure 4c we illustrate the singular value histogram of binary Jacobian model where the training classes are Airplane and Automobile. We trained the model with all samples and focus on the histogram of all training data ($n = 10,000$) before and after the training. 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. Another intriguing finding is that the distribution of before and after training are fairly close to each other highlighting that even at random initialization, the Jacobian spectrum exhibits bimodal structure.

In Figure 5, 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 5a 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 5b and 5c, we focus on the loss distribution of 5a at iterations 80 and 4500. In this figure, we visualize the loss distribution of clean and corrupted data. Figure 5b 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-capacity of the model. In Figure 5c, 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.

4 Key Insights and Technical Ideas

Our key idea is that semantically meaningful datasets (such as the clusterable dataset model) should have a low-dimensional representation. We use Jacobian mapping of the neural network to capture such structure in data which is represented as follow.

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

The key insight that enable our proofs is that the Jacobian mapping of neural networks typically exhibit (1) low-rank structure with a few large singular values and many small ones and (2) the sparse corruptions are mostly aligned with the small singular directions. We have empirically verified that both properties hold for a variety of neural networks and data sets.

Using these insights we show that the optimization is implicitly decomposed into two stages which corresponds to the column subspaces induced by the large and small singular values of the Jacobian. To make this precise let us denote the overall network prediction by $f(\boldsymbol{W}) = [f(\boldsymbol{W},\boldsymbol{x}_1) \ \ldots \ (\boldsymbol{W},\boldsymbol{x}_n)]$ and note that the gradient mapping takes the form

$$\mathcal{J}^T \underbrace{(f(\boldsymbol{W}_\tau)-\boldsymbol{y})}_{\text{corrupted residual}} = \mathcal{J}^T \underbrace{(f(\boldsymbol{W}_\tau)-\tilde{\boldsymbol{y}})}_{\text{clean residual}} + \underbrace{(\tilde{\boldsymbol{y}}-\boldsymbol{y})}_{\text{label corruption}}$$

We prove that the clean residual is aligned with the top singular direction whereas label noise is aligned with the small singular directions. The latter is a consequence of the fact that the top singular vectors are diffused and the label noise is sparse (constant fraction of corruption). As a result, gradient descent learns the useful information (clean residual) in few iterations whereas it takes much longer to overfit to noise justifying the use of early stopping.

The following meta theorem focuses on the first stage of the optimization and shows that in a general nonlinear learning problem if the Jacobian is low-rank and has a diffused range then the label noise is effectively suppressed in the first few iterations. This in turn provides a sharp control on the impact of noise on the final model for each input example. Formally, we assume that the range space $\mathcal{S} = \text{range}(\mathcal{J}(\boldsymbol{\theta}))$ is diffused in the sense that any unit length $\boldsymbol{v} \in \mathcal{S}$ satisfies $\|\boldsymbol{v}\|_{\ell_\infty} \le \sqrt{\gamma/n}$ for a small γ (e.g. $\boldsymbol{v}$ is scaled all ones vector). We note that for this diffuseness property to hold it is sufficient for the Jacobian to be approximately low-rank and the prominent directions to be diffused.

Theorem 4.1 (Robustness via diffuseness) *Consider a nonlinear least squares problem of the form $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2}\|f(\boldsymbol{\theta})-\boldsymbol{y}\|_{\ell_2}^2$. Suppose $f(\boldsymbol{\theta}_0) = 0$ and assume that $\mathcal{J}(\boldsymbol{\theta})$ is sufficiently smooth function of $\boldsymbol{\theta}$ (see Assumption 3 in supplementary) and $\mathcal{S} = \text{range}(\mathcal{J}(\boldsymbol{\theta}))$ is γ-diffused as above. Let $\widetilde{\boldsymbol{y}} = [\widetilde{y}_1 \ \ldots \ \widetilde{y}_n] \in \mathcal{S}$ denote the uncorrupted labels and $\boldsymbol{e} = \boldsymbol{y} - \widetilde{\boldsymbol{y}}$ denote the label corruption. Also assume $\boldsymbol{e}$ is ρn-sparse and its entries are bounded by 1 in absolute value. Then, running gradient descent with a constant learning rate, after polynomially many iterations, we have*

$$\|f(\boldsymbol{\theta}_\tau) - \widetilde{\boldsymbol{y}}\|_{\ell_\infty} \le \gamma\rho.$$

In words, more diffused subspace and sparser vector leads to smaller entrywise prediction error. Note that as long as $\gamma\rho < \delta/2$ (where δ is class label separation), network accurately classifies all examples. For our proofs surrounding the clusterable dataset model, we show that $\mathcal{S}$ is indeed very diffused (essentially constant γ) to obtain such tight entrywise error control.

5 Conclusions

In this paper, we studied the robustness of overparameterized neural networks to label corruption from a theoretical lens. We provided robustness guarantees for training networks with gradient descent when early stopping is used and complemented these guarantees with lower bounds. Our results point to the distance between final and initial network weights as a key feature to determine robustness vs. overfitting which is inline with weight decay and early stopping heuristics. We also carried out extensive numerical experiments to verify the theoretical predictions as well as technical assumptions. While our results shed light on the intriguing properties of overparameterized neural network optimization, it would be appealing (i) to extend our results to deeper network architecture, (ii) to more complex data models, and also (iii) to explore other heuristics that can further boost the robustness of gradient descent methods.

6 Acknowledgements

Samet Oymak is supported by NSF-CNS award #1932254. Mahdi Soltanolkotabi is supported by the Packard Fellowship in Science and Engineering, a Sloan Research Fellowship in Mathematics, an NSF-CAREER under award #1846369, the Air Force Office of Scientific Research Young Investigator Program (AFOSR-YIP) under award #FA9550-18-1-0078, and an NSF-CIF award #1813877.

References

Allen-Zhu, Z., Li, Y., and Liang, Y. (2018a). Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918*.

Allen-Zhu, Z., Li, Y., and Song, Z. (2018b). A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962*.

Arazo, E., Ortego, D., Albert, P., O'Connor, N. E., and McGuinness, K. (2019). Unsupervised label noise modeling and loss correction. *arXiv preprint arXiv:1904.11238*.

Arora, S., Cohen, N., and Hazan, E. (2018a). On the optimization of deep networks: Implicit acceleration by overparameterization. *arXiv preprint arXiv:1802.06509*.

Arora, S., Du, S. S., Hu, W., Li, Z., and Wang, R. (2019). Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584*.

Arora, S., Ge, R., Neyshabur, B., and Zhang, Y. (2018b). Stronger generalization bounds for deep nets via a compression approach.

Azizan, N. and Hassibi, B. (2018). Stochastic gradient/mirror descent: Minimax optimality and implicit regularization. *arXiv preprint arXiv:1806.00952*.

Balakrishnan, S., Du, S. S., Li, J., and Singh, A. (2017). Computationally efficient robust sparse estimation in high dimensions. In *Conference on Learning Theory*, pages 169–212.

Bartlett, P., Foster, D. J., and Telgarsky, M. (2017). Spectrally-normalized margin bounds for neural networks.

Bartlett, P. L., Long, P. M., Lugosi, G., and Tsigler, A. (2019). Benign overfitting in linear regression. *arXiv preprint arXiv:1906.11300*.

Basri, R., Jacobs, D., Kasten, Y., and Kritchman, S. (2019). The convergence rate of neural networks for learned functions of different frequencies. *arXiv preprint arXiv:1906.00425*.

Belkin, M., Hsu, D., and Mitra, P. (2018a). Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate.

Belkin, M., Rakhlin, A., and Tsybakov, A. B. (2018b). Does data interpolation contradict statistical optimality?

Bhatia, K., Jain, P., and Kar, P. (2015). Robust regression via hard thresholding. In *Advances in Neural Information Processing Systems*, pages 721–729.

Brutzkus, A. and Globerson, A. (2018). Over-parameterization improves generalization in the xor detection problem.

Brutzkus, A., Globerson, A., Malach, E., and Shalev-Shwartz, S. (2017a). Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174*.

Brutzkus, A., Globerson, A., Malach, E., and Shalev-Shwartz, S. (2017b). Sgd learns over-parameterized networks that provably generalize on linearly separable data.

Candès, E. J., Li, X., Ma, Y., and Wright, J. (2011). Robust principal component analysis? *Journal of the ACM (JACM)*, 58(3):11.

Cao, Y. and Gu, Q. (2019). A generalization theory of gradient descent for learning over-parameterized deep relu networks. *arXiv preprint arXiv:1902.01384*.

Chaudhari, P., Choromanska, A., Soatto, S., LeCun, Y., Baldassi, C., Borgs, C., Chayes, J., Sagun, L., and Zecchina, R. (2016). Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838*.

Chen, Y., Caramanis, C., and Mannor, S. (2013). Robust sparse regression under adversarial corruption. In *International Conference on Machine Learning*, pages 774–782.

Chizat, L. and Bach, F. (2018). On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545*.

Diakonikolas, I., Kamath, G., Kane, D. M., Li, J., Steinhardt, J., and Stewart, A. (2018). Sever: A robust meta-algorithm for stochastic optimization. *arXiv preprint arXiv:1803.02815*.

Du, S. S., Lee, J. D., Li, H., Wang, L., and Zhai, X. (2018a). Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804*.

Du, S. S., Zhai, X., Poczos, B., and Singh, A. (2018b). Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054*.

Foygel, R. and Mackey, L. (2014). Corrupted sensing: Novel guarantees for separating structured signals. *IEEE Transactions on Information Theory*, 60(2):1223–1247.

Frénay, B., Kabán, A., et al. (2014). A comprehensive introduction to label noise. In *ESANN*.

Ghorbani, B., Krishnan, S., and Xiao, Y. (2019). An investigation into neural net optimization

via hessian eigenvalue density. *arXiv preprint arXiv:1901.10159*.

Golowich, N., Rakhlin, A., and Shamir, O. (2017). Size-independent sample complexity of neural networks.

Guan, M. Y., Gulshan, V., Dai, A. M., and Hinton, G. E. (2018). Who said what: Modeling individual labelers improves classification. In *Thirty-Second AAAI Conference on Artificial Intelligence*.

Han, B., Yao, Q., Yu, X., Niu, G., Xu, M., Hu, W., Tsang, I., and Sugiyama, M. (2018). Co-teaching: Robust training of deep neural networks with extremely noisy labels. In *Advances in Neural Information Processing Systems*, pages 8536–8546.

Hochreiter, S. and Schmidhuber, J. (1997). Flat minima. *Neural Computation*, 9(1):1–42.

Hu, W., Li, Z., and Yu, D. (2019). Understanding generalization of deep neural networks trained with noisy labels. *arXiv preprint arXiv:1905.11368*.

Jacot, A., Gabriel, F., and Hongler, C. (2018). Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580.

Javadi, H., Balestriero, R., and Baraniuk, R. (2019). A hessian based complexity measure for deep networks. *arXiv preprint arXiv:1905.11639*.

Ji, Z. and Telgarsky, M. (2018). Gradient descent aligns the layers of deep linear networks.

Khetan, A., Lipton, Z. C., and Anandkumar, A. (2017). Learning from noisy singly-labeled data. *arXiv preprint arXiv:1712.04577*.

Klivans, A., Kothari, P. K., and Meka, R. (2018). Efficient algorithms for outlier-robust regression. *arXiv preprint arXiv:1803.03241*.

Li, X. (2013). Compressed sensing and matrix completion with constant proportion of corruptions. *Constructive Approximation*, 37(1):73–99.

Li, Y. and Liang, Y. (2018). Learning overparameterized neural networks via stochastic gradient descent on structured data. In *Advances in Neural Information Processing Systems*, pages 8168–8177.

Liang, T. and Rakhlin, A. (2018). Just interpolate: Kernel "ridgeless" regression can generalize.

Liu, L., Shen, Y., Li, T., and Caramanis, C. (2018). High dimensional robust sparse regression. *arXiv preprint arXiv:1805.11643*.

Ma, C., Wu, L., et al. (2019). A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics. *arXiv preprint arXiv:1904.04326*.

Malach, E. and Shalev-Shwartz, S. (2017). Decoupling" when to update" from" how to update". In *Advances in Neural Information Processing Systems*, pages 960–970.

Menon, A. K., van Rooyen, B., and Natarajan, N. (2018). Learning from binary labels with instance-dependent noise. *Machine Learning*, pages 1–35.

Neyshabur, B., Li, Z., Bhojanapalli, S., LeCun, Y., and Srebro, N. (2018). Towards understanding the role of over-parametrization in generalization of neural networks. *arXiv preprint arXiv:1805.12076*.

Oymak, S., Fabian, Z., Li, M., and Soltanolkotabi, M. (2019). Generalization guarantees for neural networks via harnessing the low-rank structure of the jacobian. *arXiv preprint arXiv:1906.05392*.

Oymak, S. and Soltanolkotabi, M. (2018). Overparameterized nonlinear learning: Gradient descent takes the shortest path? *arXiv preprint arXiv:1812.10004*.

Oymak, S. and Soltanolkotabi, M. (2019). Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674*.

Papyan, V. (2019). Measuring the spectrum of deepnet hessians.

Prasad, A., Suggala, A. S., Balakrishnan, S., and Ravikumar, P. (2018). Robust estimation via robust gradient estimation. *arXiv preprint arXiv:1802.06485*.

Rahaman, N., Baratin, A., Arpit, D., Draxler, F., Lin, M., Hamprecht, F. A., Bengio, Y., and Courville, A. (2018). On the spectral bias of neural networks. *arXiv preprint arXiv:1806.08734*.

Reed, S., Lee, H., Anguelov, D., Szegedy, C., Erhan, D., and Rabinovich, A. (2014). Training deep neural networks on noisy labels with bootstrapping. *arXiv preprint arXiv:1412.6596*.

Ren, M., Zeng, W., Yang, B., and Urtasun, R. (2018). Learning to reweight examples for robust deep learning. *arXiv preprint arXiv:1803.09050*.

Rolnick, D., Veit, A., Belongie, S., and Shavit, N. (2017). Deep learning is robust to massive label noise. *arXiv preprint arXiv:1705.10694*.

Sagun, L., Evci, U., Guney, V. U., Dauphin, Y., and Bottou, L. (2017). Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454*.

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

Scott, C., Blanchard, G., and Handy, G. (2013). Classification with asymmetric label noise: Consistency and maximal denoising. In *Conference On Learning Theory*, pages 489–511.

Shen, Y. and Sanghavi, S. (2018). Iteratively learning from the best. *arXiv preprint arXiv:1810.11874*.

Soltanolkotabi, M., Javanmard, A., and Lee, J. D. (2018). Theoretical insights into the optimization landscape of over-parameterized shallow neural networks. *IEEE Transactions on Information Theory*.

Song, M., Montanari, A., and Nguyen, P. (2018). A mean field view of the landscape of two-layers neural networks. In *Proceedings of the National Academy of Sciences*, volume 115, pages E7665–E7671.

Soudry, D. and Carmon, Y. (2016). No bad local minima: Data independent training error guarantees for multilayer neural networks.

Su, L. and Yang, P. (2019). On learning over-parameterized neural networks: A functional approximation prospective. *arXiv preprint arXiv:1905.10826*.

Talagrand, M. (2006). *The generic chaining: upper and lower bounds of stochastic processes*. Springer Science & Business Media.

Venturi, L., Bandeira, A., and Bruna, J. (2018). Spurious valleys in two-layer neural network optimization landscapes. *arXiv preprint arXiv:1802.06384*.

Xie, B., Liang, Y., and Song, L. (2016). Diverse neural network learns true target functions. *arXiv preprint arXiv:1611.03131*.

Yao, Y., Rosasco, L., and Caponnetto, A. (2007). On early stopping in gradient descent learning. *Constructive Approximation*, 26(2):289–315.

Zhang, C., Bengio, S., Hardt, M., Recht, B., and Vinyals, O. (2016). Understanding deep learning requires rethinking generalization. *International Conference on Learning Representations*.

Zhang, Z. and Sabuncu, M. R. (2018). Generalized cross entropy loss for training deep neural networks with noisy labels. *arXiv preprint arXiv:1805.07836*.

Zhu, Z., Soudry, D., Eldar, Y. C., and Wakin, M. B. (2018). The global optimization geometry of shallow linear neural networks.

Zou, D., Cao, Y., Zhou, D., and Gu, Q. (2018). Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888*.

7 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 7.1 (Training with perfectly clustered data) *Consier the setting and assumptions of Theorem 7.1 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 (3) with step size $\eta \le \frac{K}{2c_{up}n\Gamma^2\|\boldsymbol{C}\|^2}$. Furthermore, assume the number of hidden nodes obey*

$$k \ge C\Gamma^4 \frac{K\log(K)\|\boldsymbol{C}\|^2}{\lambda(\boldsymbol{C})^2},$$

with $\lambda(\boldsymbol{C})$ is the minimum cluster 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.

$$\arg\min_{\alpha_\ell : 1\le\ell\le K} |f(\boldsymbol{W}_\tau, \boldsymbol{x}_i) - \alpha_\ell| = \widetilde{y}_i \quad \text{for} \quad i = 1, 2, \ldots, n. \tag{6}$$

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 Oymak and Soltanolkotabi (2019) 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. In fact, in this case given the clustered nature of the input data the network never overfits to the corrupted data even after many iterations.

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

Lemma 8.1 *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 2.3 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.

9 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

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{\theta}_\tau) \quad \text{with} \quad \nabla\mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}^T(\boldsymbol{\theta})\left(f(\boldsymbol{\theta}) - \boldsymbol{y}\right). \tag{7}$$

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

$$\mathcal{J}(\boldsymbol{\theta}) = \begin{bmatrix} \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_1)}{\partial \boldsymbol{\theta}} & \ldots & \frac{\partial f(\boldsymbol{\theta}, \boldsymbol{x}_n)}{\partial \boldsymbol{\theta}} \end{bmatrix}^T. \tag{8}$$

9.1 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*

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

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 3, 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) Sagun et al. (2017) and is related to various results demonstrating that there are flat directions in the loss landscape Hochreiter and Schmidhuber (1997).

9.2 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.

Definition 9.1 (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 $\gamma > 0$.

[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).

The following theorem is the formal version of Theorem 4.1 and 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 7.1 for the perfectly clustered data ($\epsilon_0 = 0$) is obtained by combining this result with specific estimates developed for neural networks.

Theorem 9.2 (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 $\boldsymbol{y} = [y_1 \ \ldots \ y_n] \in \mathbb{R}^n$ denoting the corrupted labels. Also let $\widetilde{\boldsymbol{y}} = [\widetilde{y}_1 \ \ldots \ \widetilde{y}_n] \in \mathbb{R}^n$ 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 (7) with a learning rate $\eta \le \frac{1}{2\beta^2}\min\left(1, \frac{\alpha\beta}{L\|\boldsymbol{r}_0\|_{\ell_2}}\right)$, all iterates obey*

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

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

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

Furthermore, if $\boldsymbol{e}$ has at most s nonzeros and $\mathcal{S}_+$ is γ diffused per Definition 9.1, then using $\nu = \|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}$

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

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 7.1 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 Oymak and Soltanolkotabi (2018); Allen-Zhu et al. (2018b); Du et al. (2018b) 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.

9.3 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 9.3 *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 $\alpha \ge \epsilon$. 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.

10 Proofs

10.1 Proofs for General Theory

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

Definition 10.1 (Average Jacobian) *We define the average Jacobian along the path connecting two points $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^p$ as*

$$\mathcal{J}(\boldsymbol{y}, \boldsymbol{x}) := \int_0^1 \mathcal{J}(\boldsymbol{x} + \alpha(\boldsymbol{y} - \boldsymbol{x}))d\alpha. \tag{9}$$

Lemma 10.2 (Linearization of the residual) *Given gradient descent iterate $\hat{\boldsymbol{\theta}} = \boldsymbol{\theta} - \eta\nabla\mathcal{L}(\boldsymbol{\theta})$, define*

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

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 10.1, denoting $f(\hat{\boldsymbol{\theta}}) - \boldsymbol{y} = \hat{\boldsymbol{r}}$ and $f(\boldsymbol{\theta}) - \boldsymbol{y} = \boldsymbol{r}$, we find that

$$\begin{aligned}
\hat{\boldsymbol{r}} &= \boldsymbol{r} - f(\boldsymbol{\theta}) + f(\hat{\boldsymbol{\theta}}) \\
&\overset{(a)}{=} \boldsymbol{r} + \mathcal{J}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}) \\
&\overset{(b)}{=} \boldsymbol{r} - \eta\mathcal{J}(\hat{\boldsymbol{\theta}}, \boldsymbol{\theta})\mathcal{J}(\boldsymbol{\theta})^T\boldsymbol{r} \\
&= (\boldsymbol{I} - \eta\boldsymbol{C}(\boldsymbol{\theta}))\boldsymbol{r}.
\end{aligned} \tag{10}$$

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

Lemma 10.3 *Suppose Assumption 9.1 holds. If $\boldsymbol{r} \in \mathbb{R}^n$ is a vector with s nonzero entries, we have that*

$$\|\Pi_{\mathcal{S}_+}(\boldsymbol{r})\|_{\ell_\infty} \leq \frac{\gamma\sqrt{s}}{n}\|\boldsymbol{r}\|_{\ell_2}. \tag{11}$$

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

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

■

10.1.1 Proof of Theorem 9.2

Proof The proof will be done inductively over the properties of gradient descent iterates and is inspired from the recent work Oymak and Soltanolkotabi (2018). In particular, Oymak and Soltanolkotabi (2018) 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.

$$\bar{\boldsymbol{e}}_\tau = \bar{\boldsymbol{e}}_0 \tag{12}$$

$$\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 \leq \left(1 - \frac{\eta\alpha^2}{2}\right)^\tau \|\bar{\boldsymbol{r}}_0\|_{\ell_2}^2, \tag{13}$$

$$\frac{1}{4}\alpha\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} + \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} \leq \|\bar{\boldsymbol{r}}_0\|_{\ell_2} \leq \|\boldsymbol{r}_0\|_{\ell_2}. \tag{14}$$

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 (14) 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* (12), $\boldsymbol{\theta}_{\tau+1} \in \mathcal{D}$.

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

$$\|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_\tau\|_{\ell_2} = \eta\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\left(f(\boldsymbol{\theta}_\tau) - \boldsymbol{y}\right)\|_{\ell_2} \tag{15}$$

$$\overset{(a)}{=} \eta\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\left(\Pi_{\mathcal{S}_+}(f(\boldsymbol{\theta}_\tau) - \boldsymbol{y})\right)\|_{\ell_2} \tag{16}$$

$$\overset{(b)}{=} \eta\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2} \tag{17}$$

$$\overset{(c)}{\leq} \eta\beta\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} \tag{18}$$

$$\overset{(d)}{\leq} \frac{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}{\beta} \tag{19}$$

$$\overset{(e)}{\leq} \frac{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}{\alpha} \tag{20}$$

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 (14) yields (after scaling (14) by $4/\alpha$)

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

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

To proceed, we shall verify that (14) holds for $\tau + 1$ as well. Note that, following Lemma 10.2, 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

$$\bar{\boldsymbol{e}}_{\tau+1} = \Pi_{\mathcal{S}_-}((\boldsymbol{I} - \boldsymbol{C}(\boldsymbol{\theta}_\tau))\boldsymbol{r}_\tau) \tag{21}$$

$$= \Pi_{\mathcal{S}_-}(\boldsymbol{r}_\tau) \tag{22}$$

$$= \bar{\boldsymbol{e}}_\tau, \tag{23}$$

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

$$\begin{aligned}\bar{\boldsymbol{r}}_{\tau+1} &= \Pi_{\mathcal{S}_+}((\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{\theta}_\tau))\boldsymbol{r}_\tau) && (24)\\ &= \Pi_{\mathcal{S}_+}((\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{r}}_\tau) + \Pi_{\mathcal{S}_+}((\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{e}}_\tau) && (25)\\ &= \Pi_{\mathcal{S}_+}((\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{r}}_\tau) && (26)\\ &= (\boldsymbol{I}-\boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{r}}_\tau && (27)\end{aligned}$$

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*[5]

$$\beta^2 \boldsymbol{P}_{\mathcal{S}_+} \succeq \boldsymbol{C}(\boldsymbol{\theta}_\tau) \succeq \frac{1}{2}\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T \succeq \frac{\alpha^2}{2}\boldsymbol{P}_{\mathcal{S}_+}. \tag{28}$$

Proof The proof utilizes the upper bound on the learning rate. The argument is similar to the proof of Lemma 9.7 of Oymak and Soltanolkotabi (2018). Suppose Assumption 3 holds. Then, for any $\boldsymbol{\theta}_1, \boldsymbol{\theta}_2 \in \mathcal{D}$ we have

$$\begin{aligned}\|\mathcal{J}(\boldsymbol{\theta}_2,\boldsymbol{\theta}_1) - \mathcal{J}(\boldsymbol{\theta}_1)\| &= \left\| \int_0^1 \left(\mathcal{J}\left(\boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)\right) - \mathcal{J}(\boldsymbol{\theta}_1)\right) dt \right\|,\\ &\le \int_0^1 \left\|\mathcal{J}\left(\boldsymbol{\theta}_1 + t(\boldsymbol{\theta}_2-\boldsymbol{\theta}_1)\right) - \mathcal{J}(\boldsymbol{\theta}_1)\right\| dt,\\ &\le \int_0^1 tL\|\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\|_{\ell_2} dt \le \frac{L}{2}\|\boldsymbol{\theta}_2-\boldsymbol{\theta}_1\|_{\ell_2}. \end{aligned} \tag{29}$$

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

$$\begin{aligned}\|\mathcal{J}(\boldsymbol{\theta}_{\tau+1},\boldsymbol{\theta}_\tau) - \mathcal{J}(\boldsymbol{\theta}_\tau)\| &\le \frac{L}{2}\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_\tau\|_{\ell_2} && (30)\\ &= \frac{\eta L}{2}\left\|\mathcal{J}^T(\boldsymbol{\theta}_\tau)\left(f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}\right)\right\|_{\ell_2} \le \frac{\eta\beta L}{2}\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} && (31)\\ &\overset{(a)}{\le} \frac{\eta\beta L}{2}\|\bar{\boldsymbol{r}}_0\|_{\ell_2} \overset{(b)}{\le} \frac{\alpha}{2}. && (32)\end{aligned}$$

where for (a) we utilized the induction hypothesis (14) and (b) follows from the upper bound on η. Now that (32) 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 10.4 (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,*

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

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

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

[5] 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 $\boldsymbol{v}$, $\boldsymbol{v}^T(\boldsymbol{A}-\boldsymbol{B})\boldsymbol{v} \ge 0$.

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 (14). Note that, using the claim above and recalling (27) and using the fact that $\|\mathcal{J}(\boldsymbol{\theta}_{\tau+1},\boldsymbol{\theta}_\tau)\| \le \beta$, the residual satisfies

$$\begin{aligned}
\|\bar{\boldsymbol{r}}_{\tau+1}\|_{\ell_2}^2 = \|(\boldsymbol{I}-\eta\boldsymbol{C}(\boldsymbol{\theta}_\tau))\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 &= \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - 2\eta\bar{\boldsymbol{r}}_\tau^T\boldsymbol{C}_\tau\bar{\boldsymbol{r}}_\tau + \eta^2\bar{\boldsymbol{r}}_\tau^T\boldsymbol{C}_\tau^T\boldsymbol{C}_\tau\bar{\boldsymbol{r}}_\tau && (33)\\
&\le \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - \eta\bar{\boldsymbol{r}}_\tau^T\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau + \eta^2\beta^2\bar{\boldsymbol{r}}_\tau^T\mathcal{J}(\boldsymbol{\theta}_\tau)\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau && (34)\\
&\le \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - (\eta-\eta^2\beta^2)\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 && (35)\\
&\le \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2 - \frac{\eta}{2}\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2. && (36)
\end{aligned}$$

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

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

which establishes the second statement of the induction (14). What remains is obtaining the last statement of (14). To address this, completing squares, observe that

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

On the other hand, the distance to initial point satisfies

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

Combining the last two lines (by scaling the second line by $\frac{1}{4}\alpha$) and using induction hypothesis (14), we find that

$$\begin{aligned}
\frac{1}{4}\alpha\|\boldsymbol{\theta}_{\tau+1}-\boldsymbol{\theta}_0\|_{\ell_2} + \|\bar{\boldsymbol{r}}_{\tau+1}\|_{\ell_2} &\le \frac{1}{4}\alpha(\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2} + \eta\|\mathcal{J}(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2}) + \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} - \frac{\eta}{4}\frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2}{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}} && (37)\\
&\le \left[\frac{1}{4}\alpha\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2} + \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}\right] + \frac{\eta}{4}\left[\alpha\|\mathcal{J}(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2} - \frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}^2}{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}\right] && (38)\\
&\le \left[\frac{1}{4}\alpha\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2} + \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}\right] + \frac{\eta}{4}\|\mathcal{J}(\boldsymbol{\theta}_\tau)\bar{\boldsymbol{r}}_\tau\|_{\ell_2}\left[\alpha - \frac{\|\mathcal{J}(\boldsymbol{\theta}_\tau)^T\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}{\|\bar{\boldsymbol{r}}_\tau\|_{\ell_2}}\right] && (39)\\
&\le \frac{1}{4}\alpha\|\boldsymbol{\theta}_\tau-\boldsymbol{\theta}_0\|_{\ell_2} + \|\bar{\boldsymbol{r}}_\tau\|_{\ell_2} && (40)\\
&\le \|\bar{\boldsymbol{r}}_0\|_{\ell_2} \le \|\boldsymbol{r}_0\|_{\ell_2}. && (41)
\end{aligned}$$

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

$$\begin{aligned}
\|f(\boldsymbol{\theta}_\tau)-\boldsymbol{y}-\boldsymbol{e}\|_{\ell_\infty} &= \|\boldsymbol{r}_\tau-\boldsymbol{e}\|_{\ell_\infty} && (42)\\
&\le \|\bar{\boldsymbol{r}}_\tau\|_{\ell_\infty} + \|\boldsymbol{e}-\bar{\boldsymbol{e}}_\tau\|_{\ell_\infty} && (43)\\
&= \|\bar{\boldsymbol{r}}_\tau\|_{\ell_\infty} + \|\boldsymbol{e}-\Pi_{\mathcal{S}_-}(\boldsymbol{e})\|_{\ell_\infty} && (44)\\
&= \|\bar{\boldsymbol{r}}_\tau\|_{\ell_\infty} + \|\Pi_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}. && (45)
\end{aligned}$$

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 (14), we have that

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

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

$$(1 - \frac{\eta\alpha^2}{4})^\tau \|\boldsymbol{r}_0\|_{\ell_2} \leq \nu \iff \tau \log(1 - \frac{\eta\alpha^2}{4}) \leq \log(\frac{\nu}{\|\boldsymbol{r}_0\|_{\ell_2}}) \tag{46}$$

$$\iff \tau \log(\frac{1}{1 - \frac{\eta\alpha^2}{4}}) \geq \log(\frac{\|\boldsymbol{r}_0\|_{\ell_2}}{\nu}) \tag{47}$$

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 9.1 we have

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

■

10.1.2 Proof of Generic Lower Bound – Theorem 9.3

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

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

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

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

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

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \geq \max(\frac{E_-}{\epsilon}, \frac{E_+}{\beta}). \tag{48}$$

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 (48). ■

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.

10.2 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 10.5 (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 $\Lambda_\ell \subset \{1,\ldots,n\}$ be the set of coordinates i such that $\tilde{\boldsymbol{x}}_i = \boldsymbol{c}_\ell$. Then, $\mathcal{S}_+$ is characterized by*

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

Definition 10.6 (Neural Net Jacobian) *Given input samples $(\boldsymbol{x}_i)_{i=1}^n$, form the input matrix $\boldsymbol{X} = [\boldsymbol{x}_1 \;\ldots\; \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$. The Jacobian of the learning problem (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 Oymak and Soltanolkotabi (2019) 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 10.7 (Jacobian Properties at Cluster Center) *Suppose $\boldsymbol{X} = [\boldsymbol{x}_1 \;\ldots\; \boldsymbol{x}_n]^T \in \mathbb{R}^{n\times d}$ 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 10.8 (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 $\boldsymbol{X} = [\boldsymbol{x}_1 \;\ldots\; \boldsymbol{x}_n]^T$. Let $\mathcal{S}_+$ be the support space and $(\tilde{\boldsymbol{x}}_i)_{i=1}^n$ be the associated clean dataset as described by Definition 10.5. Set $\tilde{\boldsymbol{X}} = [\tilde{\boldsymbol{x}}_1 \;\ldots\; \tilde{\boldsymbol{x}}_n]^T$. 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*

$$\left\|\mathcal{J}(\widetilde{\boldsymbol{W}},\tilde{\boldsymbol{X}})-\mathcal{J}(\boldsymbol{W},\tilde{\boldsymbol{X}})\right\|\leq\Gamma\sqrt{\frac{c_{up}n}{kK}}\left\|\boldsymbol{C}\right\|\left\|\widetilde{\boldsymbol{W}}-\boldsymbol{W}\right\|_F\quad\text{for all}\quad\widetilde{\boldsymbol{W}},\boldsymbol{W}\in\mathbb{R}^{k\times d}.$$

- *Top singular value is bounded by*

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

- *As long as*

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

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*

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

- *The range space obeys* $\text{range}(\mathcal{J}(\boldsymbol{W}_0,\tilde{\boldsymbol{X}}))\subset\mathcal{S}_+$ *where* $\mathcal{S}_+$ *is given by Definition 10.5.*

Proof Let $\mathcal{J}(\boldsymbol{W},\boldsymbol{C})$ be the Jacobian at the cluster center matrix. Applying Theorem 10.7, 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 10.5. 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 10.9 (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 $\|\boldsymbol{y}\|_{\ell_2} \le \sqrt{n}$. 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 $|\phi(0)| \le \Gamma$, 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

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

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

10.2.1 Proof of Theorem 7.1

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

Lemma 10.10 *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*

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

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. $\mathcal{I}_\ell \subset \{1, \dots, n\}$ 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 $1 \le i \le n$. 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

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

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

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

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

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

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 7.1.

Proof The proof is based on the meta Theorem 9.2, hence we need to verify its Assumptions 2 and 3 with proper values and apply Lemma 10.10 to get $\|\mathcal{P}_{\mathcal{S}_+}(\boldsymbol{e})\|_{\ell_\infty}$. We will also make significant use of Corollary 10.8.

Using Corollary 10.8, 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 10.9 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 10.8 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

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

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

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

Now, observe that

$$LR = \Gamma\sqrt{\frac{c_{up}n}{kK}}\|\boldsymbol{C}\|\Gamma\sqrt{\frac{c_0 K \log(K)}{c_{low}\lambda(\boldsymbol{C})}} = \Gamma^2\|\boldsymbol{C}\|\sqrt{\frac{c_{up}c_0 n \log K}{c_{low}k\lambda(\boldsymbol{C})}} \le \frac{\alpha_0}{2} = \sqrt{\frac{c_{low}n\lambda(\boldsymbol{C})}{8K}}, \tag{49}$$

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 9.2 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 10.10, 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

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

denoting the clean labels by $\tilde{\boldsymbol{y}}$ and applying Theorem 9.2, 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. ∎

10.2.2 Proof of Theorem 2.3

Consider

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

and note that

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

Thus

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

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 $\mathrm{vect}(\cdot)$

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

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

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

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|\\ &\quad\overset{(a)}{\le} \left|(\boldsymbol{x}_2-\boldsymbol{x}_1)^T\widetilde{\boldsymbol{W}}^T\mathrm{diag}\left(\boldsymbol{v}\right)\mathrm{diag}\left(\phi''(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{x}})\right)(\boldsymbol{W}-\boldsymbol{W}_0)\widetilde{\boldsymbol{x}}\right|\\ &\qquad + \left|\boldsymbol{v}^T\mathrm{diag}\left(\phi'\left(\widetilde{\boldsymbol{W}}\widetilde{\boldsymbol{x}}\right)\right)(\boldsymbol{W}-\boldsymbol{W}_0)(\boldsymbol{x}_2-\boldsymbol{x}_1)\right|\\ &\quad\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\|\\ &\quad\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\|\\ &\quad\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\|\\ &\quad\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)\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\quad\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)\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\quad\overset{(g)}{\le} \left(\frac{1}{\sqrt{k}}\left\|\widetilde{\boldsymbol{W}}-\boldsymbol{W}_0\right\|_F + 3 + 2\sqrt{\frac{d}{k}}\right)\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\|\\ &\quad\overset{(h)}{\le} C\Gamma\|\boldsymbol{x}_2-\boldsymbol{x}_1\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\| \qquad (50)\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

$$\begin{aligned}\|\phi(\boldsymbol{g}^T\boldsymbol{x}_2) - \phi(\boldsymbol{g}^T\boldsymbol{x}_1)\|_{\psi_2} =& \|\phi(\boldsymbol{g}^T\boldsymbol{x}_2) - \phi(\boldsymbol{g}^T\boldsymbol{x}_1)\|_{\psi_2} \\ =& \|\phi'\left(t\boldsymbol{g}^T\boldsymbol{x}_2 + (1-t)\boldsymbol{g}^T\boldsymbol{x}_1\right)\boldsymbol{g}^T(\boldsymbol{x}_2 - \boldsymbol{x}_1)\|_{\psi_2} \\ \le& \Gamma\|\boldsymbol{g}^T(\boldsymbol{x}_2 - \boldsymbol{x}_1)\|_{\psi_2} \\ \le& c\Gamma\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2}. \end{aligned} \tag{51}$$

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 (97) we conclude that $f(\boldsymbol{W}_0, \boldsymbol{x}_2) - f(\boldsymbol{W}_0, \boldsymbol{x}_1)$ is also subGaussian obeying $\|f(\boldsymbol{W}_0, \boldsymbol{x}_2) - f(\boldsymbol{W}_0, \boldsymbol{x}_1)\|_{\psi_2} \le c\Gamma\|\boldsymbol{v}\|_{\ell_2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2}$. Thus

$$|f(\boldsymbol{W}_0, \boldsymbol{x}_2) - f(\boldsymbol{W}_0, \boldsymbol{x}_1)| \le ct\Gamma\|\boldsymbol{v}\|_{\ell_2}\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2} = ct\Gamma\|\boldsymbol{x}_2 - \boldsymbol{x}_1\|_{\ell_2}, \tag{52}$$

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

Now combining (95) and (52) we have

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

Thus

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

with high probability.

10.3 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 10.11 (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*

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

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

$$\|\mathcal{J}(\boldsymbol{W},\boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}},\tilde{\boldsymbol{X}})\| \le \|\mathcal{J}(\boldsymbol{W},\boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}},\boldsymbol{X})\| + \|\mathcal{J}(\tilde{\boldsymbol{W}},\boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}},\tilde{\boldsymbol{X}})\|.$$

We first bound

$$\|\mathcal{J}(\boldsymbol{W},\boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}},\boldsymbol{X})\| = \|\mathrm{diag}(\boldsymbol{v})\phi'(\boldsymbol{W}\boldsymbol{X}^T) * \boldsymbol{X}^T - \mathrm{diag}(\boldsymbol{v})\phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T) * \boldsymbol{X}^T\| \tag{53}$$

$$= \frac{1}{\sqrt{k}}\|(\phi'(\boldsymbol{W}\boldsymbol{X}^T) - \phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T)) * \boldsymbol{X}^T\| \tag{54}$$

To proceed, we use the results on the spectrum of Hadamard product of matrices due to Schur Schur (1911). 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

$$\|\boldsymbol{A} * \boldsymbol{B}\| = \sqrt{\|(\boldsymbol{A} * \boldsymbol{B})^T(\boldsymbol{A} * \boldsymbol{B})\|} = \sqrt{\|(\boldsymbol{A}^T\boldsymbol{A}) \odot (\boldsymbol{B}^T\boldsymbol{B})\|} \le \sqrt{\|\boldsymbol{A}^T\boldsymbol{A}\|} = \|\boldsymbol{A}\|.$$

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

$$\|(\phi'(\boldsymbol{W}\boldsymbol{X}^T) - \phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T)) * \boldsymbol{X}^T\| \le \|\phi'(\boldsymbol{W}\boldsymbol{X}^T) - \phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T)\| \le \Gamma\|(\tilde{\boldsymbol{W}} - \boldsymbol{W})\boldsymbol{X}^T\|_F \le \Gamma\sqrt{n}\|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F.$$

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 $|\phi'| \le \Gamma$

$$\|\phi'(\tilde{\boldsymbol{W}}\boldsymbol{X}^T) * (\boldsymbol{X} - \tilde{\boldsymbol{X}})\| \le \Gamma\sqrt{k}\|\boldsymbol{X} - \tilde{\boldsymbol{X}}\| \le \Gamma\sqrt{kn}\varepsilon_0.$$

Combining both estimates yields

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

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

$$\|\mathcal{J}(\boldsymbol{W}_1,\boldsymbol{W}_2,\boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}_1,\tilde{\boldsymbol{W}}_2,\tilde{\boldsymbol{X}})\| \le \int_0^1 \Gamma\sqrt{n}\Big(\frac{\|\alpha(\tilde{\boldsymbol{W}}_1 - \boldsymbol{W}_1) + (1-\alpha)(\tilde{\boldsymbol{W}}_1 - \boldsymbol{W}_1)\|_F}{\sqrt{k}} + \varepsilon_0\Big)d\alpha \tag{55}$$

$$\le \Gamma\sqrt{n}\Big(\frac{\|\tilde{\boldsymbol{W}}_1 - \boldsymbol{W}_1\|_F + \|\tilde{\boldsymbol{W}}_2 - \boldsymbol{W}_2\|_F}{2\sqrt{k}} + \varepsilon_0\Big). \tag{56}$$

■

Theorem 10.12 (Robustness of gradient path to perturbation) *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*

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

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

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

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

$$\boldsymbol{W}_{\tau+1} = \boldsymbol{W}_\tau - \eta\nabla\mathcal{L}(\boldsymbol{W}_\tau) \quad \text{where} \quad \mathcal{L}(\boldsymbol{W}) = \frac{1}{2}\sum_{i=1}^n (y_i - f(\boldsymbol{W},\tilde{\boldsymbol{x}}_i))^2 \tag{57}$$

$$\tilde{\boldsymbol{W}}_{\tau+1} = \tilde{\boldsymbol{W}}_\tau - \nabla\tilde{\mathcal{L}}(\tilde{\boldsymbol{W}}_\tau) \quad \text{where} \quad \tilde{\mathcal{L}}(\tilde{\boldsymbol{W}}) = \frac{1}{2}\sum_{i=1}^n (y_i - f(\tilde{\boldsymbol{W}},\tilde{\boldsymbol{x}}_i))^2 \tag{58}$$

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

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

Proof Since $\tilde{\boldsymbol{W}}_\tau$ are the noiseless iterations, with probability $1 - 2K^{-100}$, the statements of Theorem 7.1 hold on $\tilde{\boldsymbol{W}}_\tau$. To proceed with proof, we first introduce short hand notations. We use

$$\boldsymbol{r}_i = f(\boldsymbol{W}_i, \boldsymbol{X}) - \boldsymbol{y}, \;\; \tilde{\boldsymbol{r}}_i = f(\tilde{\boldsymbol{W}}_i, \tilde{\boldsymbol{X}}_i) - \boldsymbol{y} \tag{59}$$

$$\mathcal{J}_i = \mathcal{J}(\boldsymbol{W}_i, \boldsymbol{X}), \;\; \mathcal{J}_{i+1,i} = \mathcal{J}(\boldsymbol{W}_{i+1}, \boldsymbol{W}_i, \boldsymbol{X}), \;\; \tilde{\mathcal{J}}_i = \mathcal{J}(\tilde{\boldsymbol{W}}_i, \tilde{\boldsymbol{X}}), \;\; \tilde{\mathcal{J}}_{i+1,i} = \mathcal{J}(\tilde{\boldsymbol{W}}_{i+1}, \tilde{\boldsymbol{W}}_i, \tilde{\boldsymbol{X}}) \tag{60}$$

$$d_i = \|\boldsymbol{W}_i - \tilde{\boldsymbol{W}}_i\|_F, \;\; p_i = \|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i\|_F, \;\; \beta = \Gamma\|\boldsymbol{C}\|\sqrt{c_{up} n/K}, \;\; L = \Gamma\|\boldsymbol{C}\|\sqrt{c_{up} n/Kk}. \tag{61}$$

Here β is the upper bound on the Jacobian spectrum and L is the spectral norm Lipschitz constant as in Theorem 10.8. Applying Lemma 10.11, note that

$$\|\mathcal{J}(\boldsymbol{W}_\tau, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{X}})\| \le L\|\tilde{\boldsymbol{W}} - \boldsymbol{W}\|_F + \Gamma\sqrt{n}\varepsilon_0 \le L d_\tau + \Gamma\sqrt{n}\varepsilon_0 \tag{62}$$

$$\|\mathcal{J}(\boldsymbol{W}_{\tau+1}, \boldsymbol{W}_\tau, \boldsymbol{X}) - \mathcal{J}(\tilde{\boldsymbol{W}}_{\tau+1}, \tilde{\boldsymbol{W}}_\tau, \tilde{\boldsymbol{X}})\| \le L(d_\tau + d_{\tau+1})/2 + \Gamma\sqrt{n}\varepsilon_0. \tag{63}$$

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

$$\boldsymbol{W}_{i+1} = \boldsymbol{W}_i - \eta \mathcal{J}_i \boldsymbol{r}_i \quad , \quad \tilde{\boldsymbol{W}}_{i+1} = \tilde{\boldsymbol{W}}_i - \eta \tilde{\mathcal{J}}_i^T \tilde{\boldsymbol{r}}_i \tag{64}$$

$$\|\boldsymbol{W}_{i+1} - \tilde{\boldsymbol{W}}_{i+1}\|_F \le \|\boldsymbol{W}_i - \tilde{\boldsymbol{W}}_i\|_F + \eta\|\mathcal{J}_i - \tilde{\mathcal{J}}_i\|\|\tilde{\boldsymbol{r}}_i\|_{\ell_2} + \eta\|\mathcal{J}_i\|\|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i\|_{\ell_2} \tag{65}$$

$$d_{i+1} \le d_i + \eta((L d_i + \Gamma\sqrt{n}\varepsilon_0)\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} + \beta p_i), \tag{66}$$

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

$$\boldsymbol{r}_{i+1} = \boldsymbol{r}_i - \eta \mathcal{J}_{i+1,i}\mathcal{J}_i^T \boldsymbol{r}_i \implies \tag{67}$$

$$\boldsymbol{r}_{i+1} - \tilde{\boldsymbol{r}}_{i+1} = (\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i) - \eta(\mathcal{J}_{i+1,i} - \tilde{\mathcal{J}}_{i+1,i})\mathcal{J}_i^T \boldsymbol{r}_i - \eta\tilde{\mathcal{J}}_{i+1,i}(\mathcal{J}_i^T - \tilde{\mathcal{J}}_i^T)\boldsymbol{r}_i - \eta\tilde{\mathcal{J}}_{i+1,i}\tilde{\mathcal{J}}_i^T(\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i). \tag{68}$$

$$\boldsymbol{r}_{i+1} - \tilde{\boldsymbol{r}}_{i+1} = (\boldsymbol{I} - \eta\tilde{\mathcal{J}}_{i+1,i}\tilde{\mathcal{J}}_i^T)(\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i) - \eta(\mathcal{J}_{i+1,i} - \tilde{\mathcal{J}}_{i+1,i})\mathcal{J}_i^T \boldsymbol{r}_i - \eta\tilde{\mathcal{J}}_{i+1,i}(\mathcal{J}_i^T - \tilde{\mathcal{J}}_i^T)\boldsymbol{r}_i. \tag{69}$$

$$\|\boldsymbol{r}_{i+1} - \tilde{\boldsymbol{r}}_{i+1}\|_{\ell_2} \le \|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i\|_{\ell_2} + \eta\beta\|\boldsymbol{r}_i\|_{\ell_2}(L(3d_\tau + d_{\tau+1})/2 + 2\Gamma\sqrt{n}\varepsilon_0). \tag{70}$$

$$\|\boldsymbol{r}_{i+1} - \tilde{\boldsymbol{r}}_{i+1}\|_{\ell_2} \le \|\boldsymbol{r}_i - \tilde{\boldsymbol{r}}_i\|_{\ell_2} + \eta\beta(\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} + p_i)(L(3d_\tau + d_{\tau+1})/2 + 2\Gamma\sqrt{n}\varepsilon_0). \tag{71}$$

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 (36). This implies

$$p_{i+1} \le p_i + \eta\beta(\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} + p_i)(L(3d_\tau + d_{\tau+1})/2 + 2\Gamma\sqrt{n}\varepsilon_0). \tag{72}$$

Finalizing proof: Next, using Lemma 10.9, we have $\|\tilde{\boldsymbol{r}}_0\|_{\ell_2} \le \Theta := C_0\Gamma\sqrt{n \log K}$. We claim that if

$$\boxed{\varepsilon_0 \le \mathcal{O}(\frac{1}{\tau_0 \eta \Gamma^2 n}) \le \frac{1}{8\tau_0\eta\beta\Gamma\sqrt{n}} \quad \text{and} \quad L \le \frac{2}{5\tau_0\eta\Theta(1 + 8\eta\tau_0\beta^2)} \le \frac{1}{30(\tau_0\eta\beta)^2\Theta},} \tag{73}$$

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

$$p_t \le 8t\eta\Gamma\sqrt{n}\varepsilon_0\Theta\beta \le \Theta \quad , \quad d_t \le 2t\eta\Gamma\sqrt{n}\varepsilon_0\Theta(1 + 8\eta\tau_0\beta^2). \tag{74}$$

The proof is by induction. Suppose it holds until $t \le \tau_0 - 1$. At $t+1$, via (66) we have that

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

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 (72) and using $p_t \le \Theta$, we need

$$\frac{p_{t+1}-p_t}{\eta} \le \beta\Theta(L(3d_\tau + d_{\tau+1}) + 4\Gamma\sqrt{n}\varepsilon_0) \overset{?}{\le} 8\Gamma\sqrt{n}\varepsilon_0\Theta\beta \iff \tag{75}$$

$$L(3d_\tau + d_{\tau+1}) + 4\Gamma\sqrt{n}\varepsilon_0 \overset{?}{\le} 8\Gamma\sqrt{n}\varepsilon_0 \iff \tag{76}$$

$$L(3d_\tau + d_{\tau+1}) \overset{?}{\le} 4\Gamma\sqrt{n}\varepsilon_0 \iff \tag{77}$$

$$10L\tau_0\eta(1+8\eta\tau_0\beta^2)\Theta \overset{?}{\le} 4 \iff \tag{78}$$

$$L \overset{?}{\le} \frac{2}{5\tau_0\eta(1+8\eta\tau_0\beta^2)\Theta}. \tag{79}$$

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

$$k \ge 1000\Gamma^2 n(\tau_0\eta\beta)^4\Theta^2 \tag{80}$$

$$= \mathcal{O}(\Gamma^4 n\frac{K^4}{n^4\lambda(\boldsymbol{C})^4}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^4(\|\boldsymbol{C}\|\Gamma\sqrt{n/K})^4(\Gamma\sqrt{n\log K})^2) \tag{81}$$

$$= \mathcal{O}(\Gamma^{10}\frac{K^2\|\boldsymbol{C}\|^4}{\lambda(\boldsymbol{C})^4}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^4\log^2(K)) \tag{82}$$

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 (74), distance satisfies

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

■

10.3.1 Completing the Proof of Theorem 2.2

The formal statement of Theorem 2.2 is provided below. 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 10.13 (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*

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

where $\boldsymbol{C} \in \mathbb{R}^{K\times d}$ is the matrix of cluster centers. Fix half of the entries of $\boldsymbol{v}$ to $1/\sqrt{k}$ and the other half to $-1/\sqrt{k}$ and train only over $\boldsymbol{W}$. 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 $1 - 3/K^{100} - K\exp(-100d)$, after $\tau = \mathcal{O}(\frac{K}{\eta n\lambda(\boldsymbol{C})})\log(\frac{\Gamma\sqrt{n\log K}}{\rho})$ iterations, we have that

- *The per sample normalized ℓ_2 norm bound satisfies*

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

- *Suppose $\rho \le \delta/8$. Denote the total number of prediction errors with respect to true labels (i.e. not satisfying (6)) 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 $\varepsilon_0 \le c'\delta\min(\frac{\lambda(\boldsymbol{C})^2}{\Gamma^5 K^2\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3}, \frac{1}{\Gamma\sqrt{d}})$, then, $\boldsymbol{W}_\tau$ assigns all inputs in the ε_0 neighborhood of cluster centers to the correct labels i.e. for any cluster center $\boldsymbol{c}_\ell$ and $\boldsymbol{x}$ obeying $\|\boldsymbol{x}-\boldsymbol{c}_\ell\|_{\ell_2} \le \varepsilon_0$, $\boldsymbol{x}$ receives the ground truth label of $\boldsymbol{c}_\ell$.*

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

$$\|\boldsymbol{W}_t - \boldsymbol{W}_0\|_F \le \mathcal{O}(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}} + t\eta\varepsilon_0\frac{\Gamma^4 Kn}{\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^2). \tag{83}$$

Proof Note that proposed number of iterations τ is set so that it is large enough for Theorem 7.1 to achieve small error in the clean input model ($\varepsilon_0 = 0$) and it is small enough so that Theorem 10.12 is applicable. In light of Theorems 10.12 and 7.1 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 10.12 and 7.1, under the stated conditions, we have that

$$\|\tilde{\boldsymbol{r}}_\tau\|_{\ell_\infty} \le 4\rho \quad \text{and} \tag{84}$$

$$\|\boldsymbol{r}_\tau - \tilde{\boldsymbol{r}}_\tau\|_{\ell_2} \le c\varepsilon_0\frac{K}{n\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})\Gamma^3 n^{3/2}\sqrt{\log K} \tag{85}$$

$$= c\frac{\varepsilon_0\Gamma^3 K\sqrt{n\log K}}{\lambda(\boldsymbol{C})}\log(\frac{\Gamma\sqrt{n\log K}}{\rho}) \tag{86}$$

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

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

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 $\boldsymbol{x}$ within ε_0 neighborhood of one of the cluster centers $\boldsymbol{c} \in (\boldsymbol{c}_\ell)_{\ell=1}^K$. We will argue that $f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{c})$ is smaller than $\delta/4$ when ε_0 is small enough. We again write

$$|f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{c})| \le |f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x})| + |f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{c})|$$

The first term can be bounded via

$$|f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x})| = |\boldsymbol{v}^T\phi(\boldsymbol{W}_\tau\boldsymbol{x}) - \boldsymbol{v}^T\phi(\tilde{\boldsymbol{W}}_\tau\boldsymbol{x})| \le \|\boldsymbol{v}\|_{\ell_2}\|\phi(\boldsymbol{W}_\tau\boldsymbol{x}) - \phi(\tilde{\boldsymbol{W}}_\tau\boldsymbol{x})\|_{\ell_2} \tag{87}$$

$$\le \Gamma\|\boldsymbol{W}_\tau - \tilde{\boldsymbol{W}}_\tau\|_F \tag{88}$$

$$\le \mathcal{O}(\varepsilon_0\frac{\Gamma^5 K^2}{\lambda(\boldsymbol{C})^2}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3) \tag{89}$$

Next, we need to bound

$$|f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{c})| \le |\boldsymbol{v}^T \phi(\tilde{\boldsymbol{W}}_\tau \boldsymbol{x}) - \boldsymbol{v}^T \phi(\tilde{\boldsymbol{W}}_\tau \boldsymbol{c})| \tag{90}$$

where $\|\tilde{\boldsymbol{W}}_\tau - \boldsymbol{W}_0\|_F \le \mathcal{O}(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}})$, $\|\boldsymbol{x} - \boldsymbol{c}\|_{\ell_2} \le \varepsilon_0$ 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 Theorem 12.1 (which is a variation of Theorem 2.3), with probability at $1 - K\exp(-100d)$, for all inputs $\boldsymbol{x}$ lying ε_0 neighborhood of cluster centers, we find that

$$|f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{c})| \le C'\Gamma\varepsilon_0(\|\tilde{\boldsymbol{W}}_\tau - \boldsymbol{W}_0\|_F + \sqrt{d}) \tag{91}$$

$$C\Gamma\varepsilon_0(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}} + \sqrt{d}). \tag{92}$$

Combining the two bounds above we get

$$|f(\boldsymbol{W}_\tau, \boldsymbol{x}) - f(\tilde{\boldsymbol{W}}_\tau, \boldsymbol{c})| \le \varepsilon_0\mathcal{O}(\frac{\Gamma^5 K^2}{\lambda(\boldsymbol{C})^2}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3 + \Gamma(\Gamma\sqrt{\frac{K\log K}{\lambda(\boldsymbol{C})}} + \sqrt{d})) \tag{93}$$

$$\le \varepsilon_0\mathcal{O}(\frac{\Gamma^5 K^2}{\lambda(\boldsymbol{C})^2}\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3). \tag{94}$$

Hence, if $\varepsilon_0 \le c'\delta\min(\frac{\lambda(\boldsymbol{C})^2}{\Gamma^5 K^2\log(\frac{\Gamma\sqrt{n\log K}}{\rho})^3}, \frac{1}{\Gamma\sqrt{d}})$, we obtain that, for all $\boldsymbol{x}$, the associated cluster $\boldsymbol{c}$ and true label assigned to cluster $\tilde{y} = \tilde{y}(\boldsymbol{c})$, we have that

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

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

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

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 10.12 and 7.1 respectively. This implies (83). ■

11 Proof of Lemma 8.1

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 $\|\boldsymbol{W}\|_F$.

12 Uniform guarantee for minimum distance

Theorem 12.1 *Assume $|\phi'|, |\phi''| \le \Gamma$ and $k \gtrsim d$. Suppose $\boldsymbol{W}_0 \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$. Let $\boldsymbol{c}_1, \dots, \boldsymbol{c}_K$ be cluster centers. Then, with probability at least $1-2e^{-(k+d)}-Ke^{-100d}$ over $\boldsymbol{W}_0$, any matrix $\boldsymbol{W}$ satisfying $\|\boldsymbol{W}-\boldsymbol{W}_0\|_F \lesssim \sqrt{k}$ satisfies the following. For all $1 \le i \le K$,*

$$\sup_{\|\boldsymbol{x}-\boldsymbol{c}_i\|_{\ell_2}, \|\tilde{\boldsymbol{x}}-\boldsymbol{c}_i\|_{\ell_2} \le \varepsilon_0} |f(\boldsymbol{W},\boldsymbol{x}) - f(\boldsymbol{W},\tilde{\boldsymbol{x}})| \le C\Gamma\varepsilon_0(\|\boldsymbol{W}-\boldsymbol{W}_0\| + \sqrt{d}).$$

Proof Note that

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

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

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

Using the above we have that

$$\begin{aligned}
&\left|(f(\boldsymbol{W},\boldsymbol{x}) - f(\boldsymbol{W}_0,\boldsymbol{x})) - (f(\boldsymbol{W},\widetilde{\boldsymbol{x}}) - f(\boldsymbol{W}_0,\widetilde{\boldsymbol{x}}))\right| \qquad (95)\\
&\quad \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| \\
&\qquad + \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| \\
&\quad \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\| \\
&\quad \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\| \\
&\quad \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\| \\
&\quad \overset{(e)}{\le} \left(\frac{1}{\sqrt{k}}\|\boldsymbol{W}_0\| + \frac{1}{\sqrt{k}}\left\|\overline{\boldsymbol{W}}-\boldsymbol{W}_0\right\| + 1\right)\Gamma\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\| \\
&\quad \overset{(f)}{\le} \left(\frac{1}{\sqrt{k}}\|\boldsymbol{W}_0\| + \frac{1}{\sqrt{k}}\left\|\overline{\boldsymbol{W}}-\boldsymbol{W}_0\right\|_F + 1\right)\Gamma\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\| \\
&\quad \overset{(g)}{\le} \left(\frac{1}{\sqrt{k}}\left\|\overline{\boldsymbol{W}}-\boldsymbol{W}_0\right\|_F + 3 + 2\sqrt{\frac{d}{k}}\right)\Gamma\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\| \\
&\quad \overset{(h)}{\le} C\Gamma\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_{\ell_2}\|\boldsymbol{W}-\boldsymbol{W}_0\| \qquad (96)
\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 $\left\|\overline{\boldsymbol{W}}-\boldsymbol{W}_0\right\| \le \left\|\boldsymbol{W}-\boldsymbol{W}_0\right\|_F \le \widetilde{c}\sqrt{k}$ 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

$$\begin{aligned}
\|\phi(\boldsymbol{g}^T\boldsymbol{x}) - \phi(\boldsymbol{g}^T\widetilde{\boldsymbol{x}})\|_{\psi_2} &= \|\phi(\boldsymbol{g}^T\boldsymbol{x}) - \phi(\boldsymbol{g}^T\widetilde{\boldsymbol{x}})\|_{\psi_2} \\
&= \|\phi'\left(t\boldsymbol{g}^T\boldsymbol{x} + (1-t)\boldsymbol{g}^T\widetilde{\boldsymbol{x}}\right)\boldsymbol{g}^T(\boldsymbol{x}-\widetilde{\boldsymbol{x}})\|_{\psi_2} \\
&\le \Gamma\|\boldsymbol{g}^T(\boldsymbol{x}-\widetilde{\boldsymbol{x}})\|_{\psi_2} \\
&\le c\Gamma\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_{\ell_2}. \qquad (97)
\end{aligned}$$

Also note that

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

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 $\mathbf{1}^T\boldsymbol{v}=0$ this random variable has mean zero. Hence, using the fact that weighted sum of subGaussian random variables are subgaussian combined with (97) we conclude that $f(\boldsymbol{W}_0,\boldsymbol{x})-f(\boldsymbol{W}_0,\widetilde{\boldsymbol{x}})$ is also subGaussian with Orlicz norm obeying $\|f(\boldsymbol{W}_0,\boldsymbol{x})-f(\boldsymbol{W}_0,\widetilde{\boldsymbol{x}})\|_{\psi_2}\le c\Gamma\|\boldsymbol{v}\|_{\ell_2}\|\boldsymbol{x}-\widetilde{\boldsymbol{x}}\|_{\ell_2}$. Now, suppose $\boldsymbol{x},\tilde{\boldsymbol{x}}$ be within ε_0 neighborhood of a cluster center $\boldsymbol{c}$. We write

$$|f(\boldsymbol{W}_0,\widetilde{\boldsymbol{x}})-f(\boldsymbol{W}_0,\boldsymbol{x})|\le|f(\boldsymbol{W}_0,\boldsymbol{c})-f(\boldsymbol{W}_0,\boldsymbol{x})|+|f(\boldsymbol{W}_0,\widetilde{\boldsymbol{x}})-f(\boldsymbol{W}_0,\boldsymbol{c})|$$

To proceed, since $X_{\boldsymbol{x}}=f(\boldsymbol{W}_0,\boldsymbol{x})$ is a Gaussian process, applying standard chaining bounds Talagrand (2006), we find

$$\sup_{\|\boldsymbol{x}-\boldsymbol{c}\|_{\ell_2}\le\varepsilon_0}|f(\boldsymbol{W}_0,\boldsymbol{c})-f(\boldsymbol{W}_0,\boldsymbol{x})|\le c'\Gamma\varepsilon_0\sqrt{d} \tag{98}$$

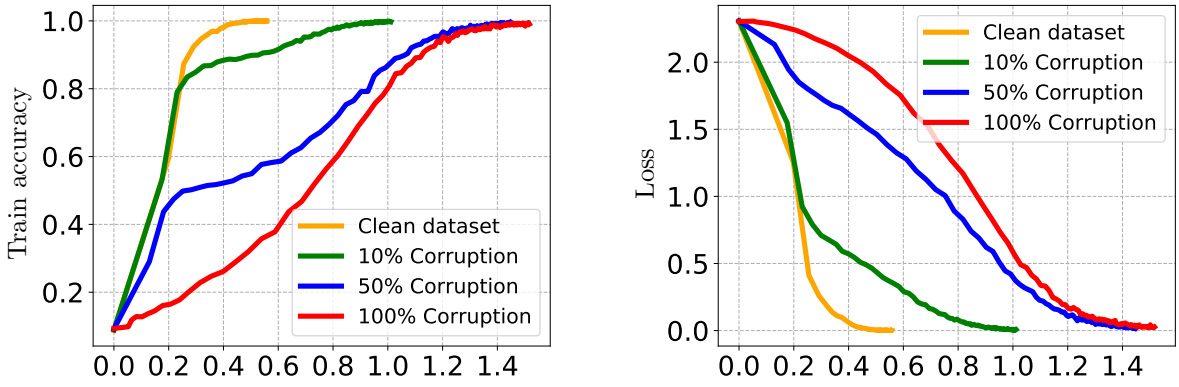
with probability $1-\exp(-100d)$. Here $\varepsilon_0\sqrt{d}$ comes from the γ_2 functional of the scaled ball around the cluster. Applying a union bound over all clusters $\boldsymbol{c}_1$ to $\boldsymbol{c}_K$, we find that, with $1-\exp(-d)$ probability, (98) holds uniformly which implies that for all $\boldsymbol{x},\tilde{\boldsymbol{x}}$ pairs of interest

$$\sup_{\boldsymbol{x},\tilde{\boldsymbol{x}}\text{ within cluster}}|f(\boldsymbol{W}_0,\widetilde{\boldsymbol{x}})-f(\boldsymbol{W}_0,\boldsymbol{x})|\le 2c'\Gamma\varepsilon_0\sqrt{d}.$$

Combining this with (96), we conclude with the advertised bound. ■

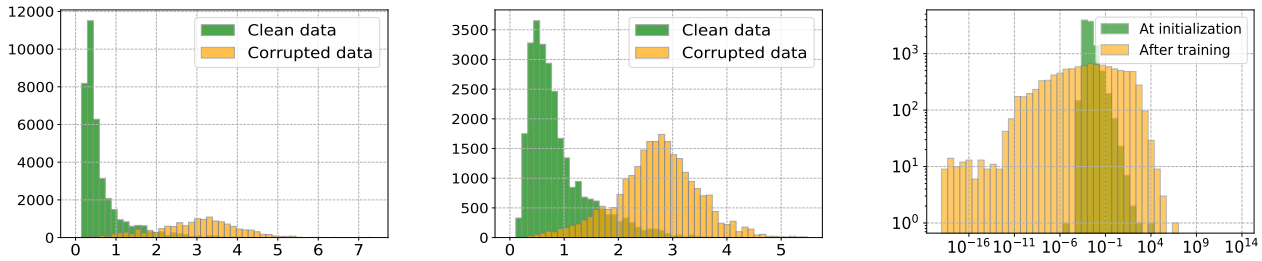
13 Experiments on cross entropy loss

We also do simulations on cross entropy loss which has same configuration as 3 and 4. The same observation happens when optimizing cross entropy loss.



(a) Training accuracy (b) Training loss

Figure 6: 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 target loss is cross entropy, the x-axis keeps track of the distance between the current and initial weights of all layers combined.



(a) 30% corruption (b) 50% corruption (c) Class Airplane and Automobile

Figure 7: (a)(b) Are histograms of the cross entropy loss of individual data points based on a model trained on 50,000 samples from CIFAR-10 with early stopping. The loss distribution of clean and corrupted data are separated but gracefully overlap as corruption increases. (c) is histogram of singular values obtained by forming the Jacobian by taking partial derivatives of class Airplane and Automobile on 10000 samples.