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

May 5, 2019

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

1. Introduction

1.1. Motivation
This paper focuses on an intriguing phenomena: 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 and generalizes to new test data. Even more surprising, the train accuracy (orange curve) with respect to the ground truth labels continues to stay around 90% even when 50% of the labels are corrupted. That is, with early stopping overparameterized neural networks even correct the corrupted labels! These plots collectively demonstrate that overparameterized neural networks when combined with early stopping have unique generalization and robustness capabilities. As we detail further in Section D this phenomena holds (albeit less pronounced) for richer data models and architectures.

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

1.2. Models
We now describe the dataset model used in our theoretical results. In this model we assume that the input samples \( x_1, x_2, \ldots, x_n \in \mathbb{R}^d \) come from \( K \) clusters which are located on the unit Euclidian ball in \( \mathbb{R}^d \). We also assume our data set consists of \( K \leq K \) classes where each class can be composed of multiple clusters. We consider a deterministic data set with \( n \) samples with roughly balanced clusters each
consisting on the order of $n/K$ samples.\footnote{This is for ease of exposition rather than a particular challenge arising in the analysis.} Finally, while we allow for multiple classes, in our theoretical model we assume the labels are scalars and take values in $[-1,1]$ interval. We formally define our dataset model below and provide an illustration in Figure 2.

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

$$|x - c_\ell|_{\ell^2} \leq \varepsilon_0,$$

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

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

$$|\alpha_r - \alpha_s| \geq \delta \quad \text{for} \quad r \neq s, \quad (1.1)$$

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

In the data model above $\{c_\ell\}^K_{\ell=1}$ 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 $\delta$. In particular, we can have at most $K \leq \frac{2}{\delta} + 1$ classes. We remark that this model is related to the setup of (4) which focuses on providing polynomial guarantees for learning shallow networks. Finally, note that, we need some sort of separation between the cluster centers to distinguish them. While Definition 1.1 doesn’t specifies such separation explicitly, Definition 2.1 establishes a notion of separation in terms of how well a neural net can distinguish the cluster centers. Next, we introduce our noisy/corrupted dataset model.

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

We note that this definition allows for a fraction $\rho$ of corruptions in each cluster.
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 \( \phi \), input weight matrix \( W \in \mathbb{R}^{k \times d} \) and output weight vector \( \nu \in \mathbb{R}^k \). In this work, we will fix output \( \nu \) to be a unit vector where half the entries are \( 1/\sqrt{k} \) and other half are \( -1/\sqrt{k} \) to simplify exposition.\(^2\) We will only optimize over the weight matrix \( W \) which contains most of the network parameters and will be shown to be sufficient for robust learning. We will also assume \( \phi \) has bounded first and second order derivatives, i.e., \( |\phi'(z)|, |\phi''(z)| \leq \Gamma \) for all \( z \). The network’s prediction at an input sample \( x \) is given by

\[
x \mapsto f(W, x) = \nu^T \phi(Wx),
\]

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

\[
\mathcal{L}(W) = \frac{1}{2} \sum_{i=1}^n (y_i - f(x_i, W))^2.
\]

In particular, we will run gradient descent with a constant learning rate \( \eta \), starting from a random initialization \( W_0 \) via the following updates

\[
W_{\tau+1} = W_{\tau} - \eta \nabla \mathcal{L}(W_{\tau}).
\]

\(^2\)If the number of hidden units is odd we set one entry of \( \nu \) to zero.

2. Main results

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

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

Definition 2.1 Define the matrix of cluster centers

\[
C = [c_1 \ldots c_K]^T \in \mathbb{R}^{K \times d}.
\]

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

\[
\Sigma(C) = (CC^T) \odot \mathbb{E}_g[\phi'(Cg)\phi'(Cg)^T].
\]

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

\[
\kappa(C) = \sqrt{\frac{d}{K} \frac{\|C\|}{\lambda(C)}}.
\]

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

We note that this Theorem is slightly simplified by ignoring logarithmic terms and precise dependencies on \( \Gamma \). See Theorem E.13 for precise statements.

**Theorem 2.2 (Robust learning with early stopping)** Consider an \((s, \varepsilon_0, \delta)\) clusterable corrupted data set of input/label pairs \((\{x_i, y_i\})_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}\) per Definition 1.2 with cluster centers \((c_i)_{i=1}^K\) aggregated as rows of a matrix \( C \in \mathbb{R}^{K \times d} \). Furthermore, let \((\tilde{y}_i)_{i=1}^n\) be the corresponding uncorrupted ground truth labels. Also consider a one-hidden layer neural network of the form (1.2) where the activation \( \phi \) obeys \( |\phi(0)| \leq \Gamma \) and \( |\phi'(z)|, |\phi''(z)| \leq \Gamma \) for all \( z \) and some \( \Gamma \geq 1 \). Furthermore, we set half of the entries of \( \nu \) to \( 1/\sqrt{K} \) and the other half to \( -1/\sqrt{K} \) and train only over \( W \). Starting from an initial weight matrix \( W_0 \) selected at random with i.i.d. \( N(0, 1) \) entries we run Gradient Descent (GD) updates of the form

\[ W_{\tau+1} = W_{\tau} - \eta \nabla C(W_{\tau}) \] on the least-squares loss (1.3) with step size \( \eta = \tilde{c}_1 \frac{K^2}{n \gamma} \) with \( \tilde{c}_1 \). Furthermore, assume the number of parameters obey

\[ kd \geq \tilde{C}_1 \kappa^4(C) \frac{K^4}{d}, \]

with \( \kappa(C) \) the neural net cluster condition number per Definition 2.1. Then as long as \( \varepsilon_0 \leq \tilde{c}_1 / K^2 \) and \( \rho \leq \frac{\delta}{\tau} \) with probability at least \( 1 - 3/\sqrt{K}^{100}, \) after \( \tau_0 = c_1 \kappa^2 \lambda(C) \kappa^2(C) (\log(\frac{1}{\rho})) \) iterations, the neural network \( f(\cdot, W_{\tau_0}) \) found by gradient descent assigns all the input samples \( x_i \) to the correct ground truth labels \( \tilde{y}_i \). That is,

\[ \arg \min \alpha_{i:1 \leq i \leq K} \| f(W_{\tau}, x_i) - \alpha_i = \tilde{y}_i, \] (2.1)

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

\[ \| W_{\tau} - W_0 \| \leq \tilde{C}_1 \left( \sqrt{K} + \frac{K^2}{|C|^2} \tau \varepsilon_0 \right). \]

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

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

\[ \text{Total Number of corrupted labels } \leq \frac{n}{K}. \]

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

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

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

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

---

\[^{\dagger}\text{If } k \text{ is odd we set one entry to zero } 0 \leq \frac{k-1}{2} \text{ to } 1/\sqrt{K} \text{ and } 1 \leq \frac{k-1}{2} \text{ entries to } -1/\sqrt{K}.\]
References


A. Improvements for perfectly cluster-able data

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

**Theorem A.1** (Training with perfectly clustered data)

Consider the setting and assumptions of Theorem E.14 with $\epsilon_0 = 0$. Starting from an initial weight matrix $W_0$ selected at random with i.i.d. $N(0, 1)$ entries we run gradient descent updates of the form $W_{t+1} = W_t - \eta \nabla C(W_t)$ on the least-squares loss (1.3) with step size $\eta \leq \frac{\delta}{2\kappa(C)}$. Furthermore, assume the number of parameters obey

$$kd \geq C T^4 k^2(C) K^2,$$

with $\kappa(C)$ the neural net condition number per Definition 2.1. Then, with probability at least $1 - \frac{1}{K^{100}}$ over randomly initialized $W_0 \sim N(0, 1)$, the iterates $W_T$ obey the following properties:

1. The distance to initial point $W_0$ is upper bounded by

$$\|W_T - W_0\|_F \leq c T \sqrt{K \log K} \lambda(C).$$

2. After $\tau \geq \tau_0 := \frac{K}{\eta n \lambda(C) T^2} \log \left( \frac{\sqrt{n} \log K}{\rho} \right)$ iterations, the entrywise predictions of the learned network with respect to the ground truth labels $(\tilde{y}_i)_{i=1}^n$ satisfy

$$|f(W_T, x_i) - \tilde{y}_i| \leq 4\rho,$$

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

$$\arg \min_{\alpha_i : 1 \leq i \leq K} |f(W_T, x_i) - \alpha_i| = \tilde{y}_i \quad \text{for} \quad i = 1, 2, \ldots, n.$$  

(A.1)

This result shows that in the limit $\epsilon_0 \to 0$ where the data points are perfectly clustered, the required amount of overparameterization can be reduced from $kd \geq K^4$ to $kd \geq K^2$. In this sense this can be thought of a nontrivial analogue of (5) where the number of data points are replaced with the number of clusters and the condition number of the data points is replaced with a cluster condition number. This can be interpreted as ensuring that the network has enough capacity to fit the cluster centers $(c_t)_{t=1}^K$ 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. Indeed, in this case given the clustered nature of the input data the network never overfits to the corrupted data even after many iterations.

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

In this section we wish to provide further insight into why early stopping enables robustness and generalizable solutions. Our main insight is that while a neural network maybe expressive enough to fit a corrupted dataset, the model has to travel a longer distance from the point of initialization as a function of the distance from the cluster centers $\epsilon_0$ 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 $W$ with i.i.d. standard normal entries which ensures that the network is isometric i.e. given input $x$, $\mathbb{E}[|f(W, x)|^2] = O(|x|^2)$.

**Theorem B.1** Let $x_1, x_2 \in \mathbb{R}^d$ be two vectors with unit Euclidean norm obeying $\|x_2 - x_1\|_2 \leq \epsilon_0$. Let $f(W, x) = v^T \phi(W x)$ where $v$ is fixed, $W \in \mathbb{R}^{k \times d}$, and $k \geq cd$ with $c > 0$ a fixed constant. Assume $|\phi'|, |\phi''| \leq \Gamma$. Let $y_1$ and $y_2$ be two scalars satisfying $|y_2 - y_1| \geq \delta$. Suppose $W_0 \sim N(0, 1)$. Then, with probability at least $1 - 2e^{-(k+d) - 2e^{-\frac{\delta}{2}}}$, for any $W \in \mathbb{R}^{k \times d}$ such that $\|W - W_0\|_F \leq c\sqrt{k}$ and

$$f(W, x_1) = y_1 \quad \text{and} \quad f(W, x_2) = y_2,$$

holds, we have

$$\|W - W_0\| \geq \frac{\delta}{CT\epsilon_0} \frac{t}{1000}.$$

In words, this result shows that in order to fit a data set with a single corrupted label, a randomly initialized network has to traverse a distance of at least $\delta/\epsilon_0$. The next lemma clarifies the role of the corruption amount $s$ and shows that more label corruption within a fixed class requires a model with a larger norm in order to fit the labels. For this result we consider a randomized model with $\epsilon_0^2$ input noise variance.
Lemma B.2 Let \( c \in \mathbb{R}^d \) be a cluster center. Consider 2s data points \( \{ x_i \}_{i=1}^s \) and \( \{ \tilde{x}_i \}_{i=1}^s \) in \( \mathbb{R}^d \) generated i.i.d. around \( c \) according to the following distribution
\[
\mathbf{e} + \mathbf{g} \quad \text{with} \quad \mathbf{g} \sim \mathcal{N}(0, \frac{\varpi^2}{d} \mathbf{I}_d).
\]
Assign \( \{ x_i \}_{i=1}^s \) with labels \( y_i = y \) and \( \{ \tilde{x}_i \}_{i=1}^s \) with labels \( \tilde{y}_i = \tilde{y} \) and assume these two labels are \( \delta \) separated i.e. \( |y - \tilde{y}| \geq \delta \). Also suppose \( s \leq d \) and \( |\delta'| \leq 1 \). Then, any \( \mathbf{W} \in \mathbb{R}^{k \times d} \) satisfying
\[
\mathbf{f}(\mathbf{W}, x_i) = y_i \quad \text{and} \quad \mathbf{f}(\mathbf{W}, \tilde{x}_i) = \tilde{y}_i \quad \text{for} \quad i = 1, \ldots, s,
\]
obeys \( \| \mathbf{W}_F \| \geq \sqrt{\frac{s\delta}{4\varpi \varepsilon_0}} \) with probability at least \( 1 - e^{-d/2} \).

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

The above Theorem implies that the model has to traverse a distance of at least
\[
\| \mathbf{W}_\tau - \mathbf{W}_0 \|_F \geq \sqrt{\frac{\varpi}{K}} \frac{\delta}{\varepsilon_0},
\]
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 \( \tau \) iterates. This demonstrates that there is a gap in the required distance to initialization for fitting enough to generalize and overfitting.

To sum up, our results highlight that, one can find a network with good generalization capabilities and robustness to label corruption within a small neighborhood of the initialization and that the size of this neighborhood is independent of the corruption. However, to fit to the corrupted labels, one has to travel much more, increasing the search space and likely decreasing generalization ability. Thus, early stopping can enable robustness without overfitting by restricting the distance to the initialization.

### C. Technical Approach and General Theory

In this section, we outline our approach to proving robustness of overparameterized neural networks. Towards this goal, we consider a general formulation where we aim to fit a general nonlinear model of the form \( \mathbf{x} \mapsto f(\mathbf{\theta}, \mathbf{x}) \) with \( \mathbf{\theta} \in \mathbb{R}^d \) denoting the parameters of the model. For instance in the case of neural networks \( \mathbf{\theta} \) represents its weights. Given a data set of \( n \) input/label pairs \( \{(x_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R} \), we fit this data by minimizing a nonlinear least-squares loss of the form
\[
\mathcal{L}(\mathbf{\theta}) = \frac{1}{2} \sum_{i=1}^n (y_i - f(\mathbf{\theta}, x_i))^2.
\]
which can also be written in the more compact form
\[
\mathcal{L}(\mathbf{\theta}) = \frac{1}{2} \| f(\mathbf{\theta}) - y \|_{\ell_2}^2 \quad \text{with} \quad f(\mathbf{\theta}) := \begin{bmatrix} f(\mathbf{\theta}, x_1) \\ \vdots \\ f(\mathbf{\theta}, x_n) \end{bmatrix}.
\]

To solve this problem we run gradient descent iterations with a constant learning rate \( \eta \) starting from an initial point \( \mathbf{\theta}_0 \). These iterations take the form
\[
\mathbf{\theta}_{\tau+1} = \mathbf{\theta}_\tau - \eta \nabla \mathcal{L}(\mathbf{\theta}_\tau) \quad \text{with} \quad \nabla \mathcal{L}(\mathbf{\theta}) = \mathbf{J}^T(\mathbf{\theta}) (f(\mathbf{\theta}) - y).
\]

Here, \( \mathbf{J}(\mathbf{\theta}) \) is the \( n \times p \) Jacobian matrix associated with the nonlinear mapping \( f \) defined via
\[
\mathbf{J}(\mathbf{\theta}) = \begin{bmatrix} \frac{\partial f(\mathbf{\theta}, x_1)}{\partial \mathbf{\theta}} & \cdots & \frac{\partial f(\mathbf{\theta}, x_n)}{\partial \mathbf{\theta}} \end{bmatrix}^T.
\]

### C.1. Bimodal jacobian structure

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

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

- **Spectrum over \( S_+ \):** For all \( v \in S_+ \), with unit Euclidian norm we have
  \[
  \alpha \leq \| \mathbf{J}^T(\mathbf{\theta}) v \|_{\ell_2} \leq \beta.
  \]

- **Spectrum over \( S_- \):** For all \( v \in S_- \), with unit Euclidian norm we have
  \[
  \| \mathbf{J}^T(\mathbf{\theta}) v \|_{\ell_2} \leq \epsilon.
  \]

We will refer to \( S_+ \) as the signal subspace and \( S_- \) as the noise subspace.

When \( \epsilon \ll \alpha \) 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 $D \subset \mathbb{R}^p$ containing the initial point $\theta_0$ (i.e. $\theta_0 \in D$). Let $S_c \subset \mathbb{R}^n$ be a subspace and $S_r$ be its complement. For all $\theta \in D$, $v \in S_c$ and $w \in S_r$ with unit Euclidian norm, we have that
\[
\alpha \leq \left\| J^T(\theta)v \right\|_{\ell_2} \leq \beta \quad \text{and} \quad \left\| J^T(\theta)w \right\|_{\ell_2} = 0.
\]

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 $\{e_i\}_{i=1}^{K}$. In this case, the Jacobian will be at most rank $K$ since each row will be in the span of $\{\frac{\partial f(e_i, \theta)}{\partial \theta}\}_{i=1}^{K}$. The subspace $S_r$ is dictated by the membership of each cluster as follows: Let $\Lambda_\ell \subset \{1, ..., n\}$ be the set of coordinates $i$ such that $x_i = \alpha \epsilon$. Then, subspace is characterized by
\[
S_r = \{ v \in \mathbb{R}^n \mid v_{i_1} = v_{i_2} \quad \text{for all} \quad i_1, i_2 \in \Lambda_\ell \quad \text{and} \quad 1 \leq \ell \leq K \}.
\]

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 $S_r$ the spectral norm is small but nonzero.

In Section D, we verify that the Jacobian matrix of real datasets indeed have a bimodal structure i.e. there are few large singular values and the remaining singular values are small which further motivate Assumption 2. This is inline with earlier papers which observed that Hessian matrices of earlier papers which observed that Hessian matrices of nonlinear mappings are perfectly clustered ($\epsilon_0 = 0$) and $y = [y_1 \ldots y_n] \in \mathbb{R}^n$ denoting the corrupted labels. Also let $\tilde{y} = [\tilde{y}_1 \ldots \tilde{y}_n] \in \mathbb{R}^n$ denote the uncorrupted labels and $\epsilon = y - \tilde{y}$ the corruption. Furthermore, suppose the initial residual $f(\theta_0) - \tilde{y}$ with the uncorrupted labels obeys $f(\theta_0) - \tilde{y} \in S_r$. Then, running gradient descent updates of the from (C.1) with a learning rate $\eta \leq \frac{1}{2\epsilon^2} \min \left(1, \frac{\alpha^2}{(1 + (1/\epsilon^2))} \right)$, all iterates obey
\[
\left\| \theta_{\tau} - \theta_0 \right\|_{\ell_2} \leq \frac{4r_0 \epsilon}{\alpha}.
\]

Furthermore, assume $\nu > 0$ is a precision level obeying $\nu \geq \left\| \Pi_{s_r}(\epsilon) \right\|_{\ell_\infty}$. Then, after $\tau \geq \frac{5}{\eta^2 \nu^2} \log \left(\frac{r_0 \epsilon}{\nu} \right)$ iterations, $\theta_{\tau}$ achieves the following error bound with respect to the true labels
\[
\left\| f(\theta_{\tau}) - \tilde{y} \right\|_{\ell_\infty} \leq 2\nu.
\]

Furthermore, if $e$ has at most $s$ nonzeros and $S_r$ is $\gamma$ diffused per Definition C.1, then using $\nu = \left\| \Pi_{s_r}(e) \right\|_{\ell_\infty}$
\[
\left\| f(\theta_{\tau}) - \tilde{y} \right\|_{\ell_\infty} \leq 2\left\| \Pi_{s_r}(e) \right\|_{\ell_\infty} \leq \frac{\gamma \sqrt{s}}{n} \left\| e \right\|_{\ell_2}.
\]

This result shows that when the Jacobian of the nonlinear mapping is low-rank, gradient descent enjoys two intriguing properties. First, gradient descent iterations remain rather close to the initial point. Second, the estimated labels of the algorithm enjoy sample-wise robustness guarantees in the sense that the noise in the estimated labels are gracefully distributed over the dataset and the effects on individual label estimates are negligible. This theorem is the key result that allows us to prove Theorem E.14 when the data points are perfectly clustered ($\epsilon_0 = 0$). Furthermore, this theorem when combined with a perturbation analysis allows us to deal with data that is not perfectly clustered ($\epsilon_0 > 0$) and to conclude that with early stopping neural networks are rather robust to label corruption (Theorem 2.2).
Finally, we note that a few recent publications (7; 9; 13) require the Jacobian to be well-conditioned to fit labels perfectly. In contrast, our low-rank model cannot perfectly fit the corrupted labels. Furthermore, when the Jacobian is bimodal (as seems to be the case for many practical data sets and neural network models) it would take a very long time to perfectly fit the labels and as demonstrated earlier such a model does not generalize and is not robust to corruptions. Instead we focus on proving robustness with early stopping.

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

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

**Theorem C.3** Denote the residual at initialization $\theta_0$ by $r_0 = f(\theta_0) - y$. Define the residual projection over the signal and noise space as

$$E_s = \|\Pi_{S_s}(r_0)\|_{\ell_2} \quad \text{and} \quad E_n = \|\Pi_{S_n}(r_0)\|_{\ell_2}.$$ 

Suppose Assumption 1 holds over an Euclidian ball $\mathcal{D}$ of radius $R < \max\left(\frac{E_s}{\beta}, \frac{E_n}{\epsilon}\right)$ around the initial point $\theta_0$ with $\alpha \geq \epsilon$. Then, over $\mathcal{D}$ there exists no $\theta$ that achieves zero training loss. In particular, if $\mathcal{D} = \mathbb{R}^p$, any parameter $\theta$ achieving zero training loss ($f(\theta) = y$) satisfies the distance bound

$$\|\theta - \theta_0\|_{\ell_2} \geq \max\left(\frac{E_s}{\beta}, \frac{E_n}{\epsilon}\right).$$

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

D. Numerical experiments

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

\[\text{model of Definition 1.2 and verify the robustness capability of gradient descent with early stopping in this model.}\]

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

Next, we study the distribution of the individual sample losses on the CIFAR-10 dataset. We conducted two experiments using Resnet-20 with cross entropy loss\(^3\). In Figure 4 we assess the noise robustness of gradient descent where we used all 50,000 samples with either 30% random corruption

\[\text{loss} \quad \text{achieves similar accuracy.}\]
After training

At initialization

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

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

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

\(^5\)We depict the smaller Jacobian due to the computational cost of calculating the full Jacobian.
capacity of the model. In Figure 6c, we repeat plot after many more iterations at which point the model overfits. This plot shows that the distribution of the two classes overlap demonstrating that the model has overfit the corruption and lacks generalization/robustness.

E. Proofs

E.1. Proofs for General Theory

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

**Definition E.1 (Average Jacobian)** We define the average Jacobian along the path connecting two points \( x, y \in \mathbb{R}^p \) as
\[
\mathcal{J}(y, x) := \int_0^1 \mathcal{J}(x + \alpha(y - x)) d\alpha.
\] (E.1)

**Lemma E.2 (Linearization of the residual)** Given gradientdescent iterate \( \hat{\theta} = \theta - \eta \nabla \mathcal{L}(\theta) \), define
\[
C(\theta) = \mathcal{J}(\hat{\theta}, \theta) \mathcal{J}(\theta)^T.
\]
The residuals \( \hat{r} = f(\hat{\theta}) - y \), \( r = f(\theta) - y \) obey the following equation
\[
\hat{r} = (I - \eta C(\theta)) r.
\]

**Proof** Following Definition E.1, denoting \( f(\hat{\theta}) - y = \hat{r} \) and \( f(\theta) - y = r \), we find that
\[
\hat{r} = r - f(\theta) + f(\hat{\theta})
\]
\[
\overset{(a)}{=} r + \mathcal{J}(\hat{\theta}, \theta)(\hat{\theta} - \theta)
\]
\[
\overset{(b)}{=} r + \eta \mathcal{J}(\hat{\theta}, \theta) \mathcal{J}(\theta)^T r
\]
\[
= (I - \eta C(\theta)) r.
\] (E.2)
Here (a) uses the fact that Jacobian is the derivative of $f$ and (b) uses the fact that $\nabla L(\theta) = J(\theta)^T r$.

Using Assumption C.1, one can show that sparse vectors have small projection on $S_r$.

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

$$\| \Pi_{S_r}(r) \|_{\ell_2} \leq \frac{\gamma \sqrt{s}}{n} \| r \|_{\ell_2}. \quad (E.3)$$

**Proof** First, we bound the $\ell_2$ projection of $r$ on $S_r$ as follows

$$\| \Pi_{S_r}(r) \|_{\ell_2} = \sup_{v \in S_r} \frac{v^T r}{\|v\|_{\ell_2}} \leq \sqrt{n} \| r \|_{\ell_1} \leq \sqrt{\frac{\gamma n}{s}} \| r \|_{\ell_2}. \quad (E.4)$$

where we used the fact that $\| v \|_{\ell_1} \leq \sqrt{n} \| v \|_{\ell_2} / \sqrt{n}$. Next, we conclude with

$$\| \Pi_{S_r}(r) \|_{\ell_\infty} \leq \sqrt{n} \| \Pi_{S_r}(r) \|_{\ell_2} \leq \frac{\gamma \sqrt{n}}{n} \| r \|_{\ell_2}. \quad (E.5)$$

**E.1.1. Proof of Theorem C.2**

**Proof** The proof will be done inductively over the properties of gradient descent iterates and is inspired from the recent work (13). In particular, (13) requires a well-conditioned Jacobian to fit labels perfectly. In contrast, we have a low-rank Jacobian model which cannot fit the noisy labels (or it would have trouble fitting if the Jacobian was approximately low-rank). Despite this, we wish to prove that gradient descent satisfies desirable properties such as robustness and closeness to initialization. Let us introduce the notation related to the residual. Set $r_\tau = f(\theta_\tau) - y$ and let $r_0 = f(\theta_0) - y$ be the initial residual. We keep track of the growth of the residual by partitioning the residual as $r_\tau = \bar{r}_{\tau} + \tilde{e}_\tau$ where

$$\bar{e}_\tau = \Pi_{S_r}(r_\tau), \quad \bar{r}_{\tau} = \Pi_{S_r}(r_\tau).$$

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

$$\bar{e}_\tau = \bar{e}_0 \quad (E.6)$$

$$\| \bar{r}_{\tau} \|_{\ell_2}^2 \leq \left( 1 - \frac{\eta a^2}{2} \right)^\tau \| \bar{r}_0 \|_{\ell_2}^2, \quad (E.7)$$

$$\frac{1}{4} \| \theta_{\tau+1} - \theta_0 \|_{\ell_2} + \| \bar{r}_{\tau} \|_{\ell_2} \leq \| \bar{r}_0 \|_{\ell_2} \leq \| r_0 \|_{\ell_2}. \quad (E.8)$$

Assuming these conditions hold till some $\tau > 0$, inductively, we focus on iteration $\tau + 1$. First, note that these conditions imply that for all $\tau \geq i \geq 0$, $\theta_i \in D$ where $D$ is the Euclidian ball around $\theta_0$ of radius $\frac{\| r_0 \|_{\ell_2}}{\alpha}$. This directly follows from (E.6) induction hypothesis. Next, we claim that $\theta_{\tau+1}$ is still within the set $D$. This can be seen as follows:

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

**Proof** Since range space of Jacobian is in $S_r$ and $\eta \leq 1/\beta^2$, we begin by noting that

$$\| \theta_{\tau+1} - \theta_\tau \|_{\ell_2} = \eta \| J^T(\theta_\tau)(f(\theta_\tau) - y) \|_{\ell_2} \quad (E.9)$$

$$\leq \| J^T(\theta_\tau) \|_{\ell_2} \| \bar{r}_\tau \|_{\ell_2} \quad (E.10)$$

$$\leq \beta \| \bar{r}_\tau \|_{\ell_2} \quad (E.11)$$

$$\leq \frac{\| \bar{r}_\tau \|_{\ell_2}}{\alpha} \quad (E.12)$$

In the above, (a) follows from the fact that row range space of Jacobian is in $S_r$, via Assumption 2, (b) follows from the definition of $\bar{r}_\tau$, (c) follows from the upper bound on the spectral norm of the Jacobian over $D$ per Assumption 2, (d) from the fact that $\eta \leq \frac{1}{\beta^2}$, (e) from the upper bound on the spectral norm of $\frac{\| \theta_{\tau} - \theta_\tau \|_{\ell_2} \leq \| \bar{r}_\tau \|_{\ell_2}}{\alpha}$

$$\| \theta_{\tau+1} - \theta_0 \|_{\ell_2} \leq \| \theta_{\tau+1} - \theta_\tau \|_{\ell_2} + \| \theta_0 - \theta_\tau \|_{\ell_2} \leq \| \theta_{\tau+1} - \theta_0 \|_{\ell_2} + \| \bar{r}_\tau \|_{\ell_2} \leq \frac{4}{\alpha} \| r_0 \|_{\ell_2},$$

concluding the proof of $\theta_{\tau+1} \in D$.

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

$$r_{\tau+1} = (I - C(\theta_{\tau}))r_\tau.$$ Since both column and row space of $C(\theta_{\tau})$ is subset of $S_r$, we have that

$$e_{\tau+1} = \Pi_{S_r}((I - C(\theta_{\tau}))r_\tau) \quad (E.13)$$

$$= \Pi_{S_r}(r_\tau) \quad (E.14)$$

$$= \bar{e}_\tau. \quad (E.15)$$

This shows the first statement of the induction. Next, over $S_r$, we have

$$\bar{r}_{\tau+1} = \Pi_{S_r}(((I - C(\theta_{\tau}))r_\tau)$$

$$= \Pi_{S_r}((I - C(\theta_{\tau}))\bar{r}_\tau + \Pi_{S_r}((I - C(\theta_{\tau}))\tilde{e}_\tau) \quad (E.16)$$

$$= \Pi_{S_r}((I - C(\theta_{\tau}))\bar{r}_\tau) \quad (E.17)$$

$$= (I - C(\theta_{\tau}))\bar{r}_\tau \quad (E.18)$$

where the second line uses the fact that $\bar{e}_\tau \in S_r$ and last line uses the fact that $\bar{r}_\tau \in S_r$. To proceed, we need to prove that $C(\theta_{\tau})$ has desirable properties over $S_r$, in particular, it contracts this space.
Claim 2 Let \( P_{S^+} \in \mathbb{R}^{n \times n} \) be the projection matrix to \( S^+ \) i.e. it is a positive semi-definite matrix whose eigenvectors over \( S^+ \) is 1 and its complement is 0. Under the induction hypothesis and setup of the theorem, we have that\(^3\)

\[
\beta^2 P_{S^+} \succeq C(\theta_r) \succeq \frac{1}{2}J(\theta_r)J(\theta_r)^T \geq \frac{\alpha^2}{2} P_{S^+}. \tag{E.20}
\]

**Proof** The proof utilizes the upper bound on the learning rate. The argument is similar to the proof of Lemma 9.7 of (13). Suppose Assumption 3 holds. Then, for any \( \theta_1, \theta_2 \in \mathcal{D} \) we have

\[
\|J(\theta_2, \theta_1) - J(\theta_1)\|_2 \leq \int_0^1 \|J(\theta_1 + t(\theta_2 - \theta_1)) - J(\theta_1)\|_2 dt \leq \int_0^1 \|J(\theta_1 + t(\theta_2 - \theta_1)) - J(\theta_1)\|_2 dt \leq \int_0^1 tL \|\theta_2 - \theta_1\|_2 dt \leq \frac{L}{2} \|\theta_2 - \theta_1\|_2^2. \tag{E.21}
\]

Thus, for \( \eta \leq \frac{\alpha}{\|\theta_2 - \theta_1\|_2^2} \),

\[
\|J(\theta_{r+1}, \theta_r) - J(\theta_r)\|_2 \leq \frac{L}{2} \|\theta_{r+1} - \theta_r\|_2 \\
= \eta L \frac{\|J^T(\theta_r)(f(\theta_r) - y)\|_2}{2} \leq \eta \beta L \frac{2}{\|\theta_0\|_2} \leq \eta \frac{\beta L}{\|\theta_0\|_2} \leq (a) \frac{\eta \beta L}{2} \|\theta_0\|_2 \leq (b) \frac{\alpha}{2}. \tag{E.22}
\]

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

\[
C(\theta_r) = J(\theta_{r+1}, \theta_r)J(\theta_r)^T \geq (1/2)J(\theta_r)J(\theta_r)^T.
\]

The \( \beta^2 \) upper bound directly follows from Assumption 2 by again noticing range space of Jacobian is subset of \( S^+ \).

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

\[
AC^T \succeq CC^T \succeq \frac{\alpha^2}{2} P_{S^+}.
\]

**Proof** For \( r \in S^+ \) with unit Euclidian norm, we have

\[
r^T AC^T r = \|C^T r\|_2^2 + r^T (A - C)C^T r \\
\geq \|C^T r\|_2^2 - \|C^T r\|_2 \|r^T (A - C)\|_2 \\
= (\|C^T r\|_2 - \|r^T (A - C)\|_2)\|C^T r\|_2 \\
\geq (\|C^T r\|_2 - \alpha/2)\|C^T r\|_2 \\
\geq \|C^T r\|_2^2/2.
\]

Also, for any \( r \), by range space assumption \( r^T AC^T r = \Pi_{S^+}(r)^T AC^T \Pi_{S^+}(r) \) (same for \( CC^T \)). Combined with above, this concludes the claim.

What remains is proving the final two statements of the induction (E.6). Note that, using the claim above and recalling (E.19) and using the fact that \( \|J(\theta_{r+1}, \theta_r)\| \leq \beta \), the residual satisfies

\[
\|\tilde{r}_{r+1}\|_2 = \|(I - \eta C(\theta_r))\tilde{r}_r\|_2 \\
= \|\tilde{r}_r\|_2 - 2\eta \tilde{r}_r^T C(\theta_r)\tilde{r}_r + \eta^2 \tilde{r}_r^T C(\theta_r) C(\theta_r)^T \tilde{r}_r \\
\leq \|\tilde{r}_r\|_2 - \eta \tilde{r}_r^T J(\theta_r)J(\theta_r)^T \tilde{r}_r \\
+ \eta^2 \beta^2 \tilde{r}_r^T J(\theta_r)J(\theta_r)^T \tilde{r}_r \\
\leq \|\tilde{r}_r\|_2 - \eta \beta \|J(\theta_r)\|_2 \|\tilde{r}_r\|_2 \\
\leq \|\tilde{r}_r\|_2 - \eta \frac{\beta}{2} \|J(\theta_r)\|_2 \|\tilde{r}_r\|_2, \tag{E.23}
\]

where we used the fact that \( \eta \leq \frac{1}{2\beta^2} \). Now, using the fact that \( J(\theta_r)J(\theta_r)^T \succeq \alpha^2 P_{S^+} \), we have

\[
\|\tilde{r}_r\|_2 - \eta \frac{\alpha^2}{2} \|J(\theta_r)\|_2 \|\tilde{r}_r\|_2 \leq \frac{1 - \eta \frac{\alpha^2}{2}}{2} \|\tilde{r}_r\|_2^2 \\
\leq \frac{1 - \eta \frac{\alpha^2}{2}}{2\beta^2} \|\tilde{r}_r\|_2^2,
\]

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

\[
\|\tilde{r}_{r+1}\|_2 \leq \sqrt{\|\tilde{r}_r\|_2^2 - \eta \frac{\alpha^2}{2} \|J(\theta_r)\|_2^2} \\
\leq \|\tilde{r}_r\|_2 - \eta \frac{\alpha^2}{4} \|J(\theta_r)\|_2^2. \tag{E.24}
\]

On the other hand, the distance to initial point satisfies

\[
\|\theta_{r+1} - \theta_0\|_2 \leq \|\theta_{r+1} - \theta_r\|_2 + \|\theta_r - \theta_0\|_2 \\
\leq \|\theta_r - \theta_0\|_2 + \eta \|J(\theta_r)\|_2 \tilde{r}_r \|_2.
\]
Combining the last two lines (by scaling the second line by \(1/\alpha\)) and using induction hypothesis (E.6), we find that

\[
\frac{1}{4}\alpha \left| \theta_{t+1} - \theta_0 \right|_{\ell_2} + \left| \tilde{r}_{t+1} \right|_{\ell_2} \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \eta \left\| J(\theta_t) \tilde{r}_t \right\|_{\ell_2} \right) + \left| \tilde{r}_t \right|_{\ell_2} \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \eta \left( J(\theta_t)^T \tilde{r}_t \right) \right)_{\ell_2} + \left| \tilde{r}_t \right|_{\ell_2} \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \left| \tilde{r}_t \right|_{\ell_2} \right) + \frac{\eta}{4} \left [ \alpha \left( J(\theta_t)^T \tilde{r}_t \right)_{\ell_2} \right ] \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \left| \tilde{r}_t \right|_{\ell_2} \right) \\
+ \frac{\eta}{4} \left [ \alpha \left( J(\theta_t)^T \tilde{r}_t \right)_{\ell_2} \right ] \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \left| \tilde{r}_t \right|_{\ell_2} \right) \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \left| \tilde{r}_t \right|_{\ell_2} \right) \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \left| \tilde{r}_t \right|_{\ell_2} \right) \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \left| \tilde{r}_t \right|_{\ell_2} \right) \\
\leq \frac{1}{4}\alpha \left( \left| \theta_{t} - \theta_0 \right|_{\ell_2} + \left| \tilde{r}_t \right|_{\ell_2} \right)
\]

Consequently, if \(\tau \geq \frac{5}{\eta \alpha^2} \log\left( \frac{|r_0|_{\ell_2}}{\nu} \right)\), we find that \(\left| \tilde{r}_t \right|_{\ell_\infty} \leq \left| \tilde{r}_t \right|_{\ell_2} \leq \nu\), which guarantees

\[
\left| r_t - e \right|_{\ell_\infty} \leq 2\nu.
\]

which is the advertised result. If \(e\) is \(s\) sparse and \(S_+\) is diffused, applying Lemma C.1 we have

\[
\left| \Pi_{S_+}(e) \right|_{\ell_\infty} \leq \frac{\sqrt{s}}{n} \left| e \right|_{\ell_2}.
\]

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

Proof Suppose \(\theta \in D\) satisfies \(y = f(\theta)\). Define \(J_r = J((1 - \tau)\theta + \tau \theta_0)\) and \(J = J(\theta, \theta_0) = \int_0^1 J_r \, d\tau\). Since Jacobian is derivative of \(f\), we have that

\[
f(\theta) - f(\theta_0) = \int_0^1 J_r(\theta - \theta_0) \, d\tau = J(\theta - \theta_0).
\]

Now, define the matrices \(J_+ = \Pi_{S_+}(J)\) and \(J_- = \Pi_{S_-}(J)\). Using Assumption 1, we bound the spectral norms via

\[
\left| J_+ \right| = \sup_{\left| r \right|_{\ell_2} \leq 1} \left| J^T v \right|_{\ell_2} \leq \beta
\]

\[
\left| J_- \right| = \sup_{\left| r \right|_{\ell_2} \leq 1} \left| J^T v \right|_{\ell_2} \leq \epsilon.
\]

To proceed, projecting the residual on \(S_+\), we find for any \(\theta\) with \(f(\theta) = y\)

\[
\Pi_{S_+}(f(\theta) - f(\theta_0)) = \Pi_{S_+}(J(\theta - \theta_0)) \implies \left| \theta - \theta_0 \right|_{\ell_2} \geq \left| \Pi_{S_+}(f(\theta) - f(\theta_0)) \right|_{\ell_2} \geq \frac{E_+}{\beta}.
\]

The identical argument for \(S_-\) yields \(\left| \theta - \theta_0 \right|_{\ell_2} \geq \frac{E_-}{\epsilon}\). Together this implies

\[
\left| \theta - \theta_0 \right|_{\ell_2} \geq \max\left( \frac{E_+}{\beta}, \frac{E_-}{\epsilon} \right).
\]

If \(R\) is strictly smaller than right hand side, we reach a contradiction as \(\theta \not \in D\). If \(D = \mathbb{R}^p\), we still find (E.30).

This shows that if \(\epsilon\) 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 \(\ell_2\) regularized problem

\[
\min_{\theta} L(\theta) \quad \left| \theta - \theta_0 \right|_{\ell_2} \leq R.
\]

If we set \(R = \frac{E_+}{\beta}\), we can hope that solution will learn only the signal and does not overfit to the noise. The next section builds on this intuition and formalizes our algorithmic guarantees.
E.2. Proofs for Neural Networks

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

Definition E.5 (Support subspace) Let $\{x_i\}_{i=1}^n$ be an input dataset generated according to Definition 1.1. Also let $\{\bar{x}_i\}_{i=1}^n$ be the associated cluster centers, that is, $\bar{x}_i = c_\ell$ iff $x_i$ is from the $\ell$th cluster. We define the support subspace $S_+$ as a subspace of dimension $K$, dictated by the cluster membership as follows. Let $\Lambda_\ell \subseteq \{1, \ldots, n\}$ be the set of coordinates $i$ such that $\bar{x}_i = c_\ell$. Then, $S_+$ is characterized by

$$S_+ = \{v \in \mathbb{R}^n \mid v_{i_1} = v_{i_2} \; \text{ for all } \; i_1, i_2 \in \Lambda_\ell, \; 1 \leq \ell \leq K\}.$$ 

Definition E.6 (Neural Net Jacobian) Given input samples $(x_i)_{i=1}^n$, form the input matrix $X = [x_1 \ldots x_n]^T \in \mathbb{R}^{n \times d}$. The Jacobian of the learning problem (1.3), at a matrix $W$ is denoted by $J(W, X) \in \mathbb{R}^{n \times k \times d}$ and is given by

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

Here $*$ denotes the Khatri-Rao product.

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

Theorem E.7 (Jacobian Properties at Cluster Center) Suppose $X = [x_1 \ldots x_n]^T \in \mathbb{R}^{n \times d}$ be an input dataset satisfying $\lambda(X) > 0$. Suppose $|\phi'|, |\phi''| \leq \Gamma$. The Jacobian mapping with respect to the input-to-hidden weights obey the following properties.

- Smoothness is bounded by
  $$\|J(\tilde{W}, X) - J(W, X)\| \leq \Gamma \sqrt{\frac{\text{sup}}{kK}} \|W - \tilde{W}\|_F$$ 
  for all $W, \tilde{W} \in \mathbb{R}^{K \times d}$.

- Top singular value is bounded by
  $$\|J(W, X)\| \leq \sqrt{\frac{\text{sup}}{K}} \|X\|.$$ 

- Let $C > 0$ be an absolute constant. As long as
  $$k \geq \frac{C T^2 \log n \|X\|^2}{\lambda(X)}.$$ 
  At random Gaussian initialization $W_0 \sim \mathcal{N}(0, 1)^{k \times d}$, with probability at least $1 - 1/K^{100}$, we have
  $$\sigma_{\min}(J(W_0, X)) \geq \sqrt{\frac{\lambda(X)}{2K}}.$$ 

  - The range space obeys $\text{range}(J(W_0, \hat{X})) \subseteq S_+$ where $S_+$ is given by Definition E.5.

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

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

$$\sigma_{\min}(X, S) = \sup_{\|v\|_2=1, U^TU=PS} \|v^T U^TX\|_{\ell_2}.$$ 

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

Theorem E.8 (Jacobian Properties at Clusterable Dataset) Let input samples $(x_i)_{i=1}^n$ be generated according to $(\varepsilon_0, \delta)$ clusterable dataset model of Definition 1.1 and define $X = [x_1 \ldots x_n]^T$. Let $S_+$ be the support space and $(\bar{x}_i)_{i=1}^n$ be the associated clean dataset as described by Definition E.5. Set $\tilde{X} = [\bar{x}_1 \ldots \bar{x}_n]^T$. Assume $|\phi'|, |\phi''| \leq \Gamma$ and $\lambda(C) > 0$. The Jacobian mapping at $\tilde{X}$ with respect to the input-to-hidden weights obey the following properties.

- Smoothness is bounded by
  $$\|J(\tilde{W}, \tilde{X}) - J(W, X)\| \leq \Gamma \sqrt{\frac{\text{sup}}{kK}} \|C\| \|\tilde{W} - W\|_F$$ 
  for all $W, \tilde{W} \in \mathbb{R}^{K \times d}$.

- Top singular value is bounded by
  $$\|J(W, \tilde{X})\| \leq \sqrt{\frac{\text{sup}}{K}} \|C\|.$$ 

- As long as
  $$k \geq \frac{C T^2 \log n \|C\|^2}{\lambda(C)}.$$ 
  At random Gaussian initialization $W_0 \sim \mathcal{N}(0, 1)^{k \times d}$, with probability at least $1 - 1/K^{100}$, we have
  $$\sigma_{\min}(J(W_0, \tilde{X}, S_+)) \geq \sqrt{\frac{\lambda(C)}{2K}}.$$ 

- The range space obeys $\text{range}(J(W_0, \tilde{X})) \subseteq S_+$ where $S_+$ is given by Definition E.5.
Note that $\tilde{X}$ is obtained by duplicating the rows of the cluster center matrix $C$. This implies that $\mathcal{J}(W, \tilde{X})$ 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 $c_{\text{low}} n / K$ and $c_{\text{up}} n / K$.

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

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

This implies $\tilde{p} \in S_*$ hence range($\mathcal{J}(W, \tilde{X})$) $\subset$ $S_*$. Furthermore, the entries of $\tilde{p}$ repeats the entries of $p$ somewhere between $c_{\text{low}} n / K$ and $c_{\text{up}} n / K$. This implies that,$$
abla \frac{c_{\text{up}} n \parallel p \parallel_2}{K} \geq |\tilde{p}|_{\ell_2} \geq \sqrt{c_{\text{low}} n \parallel p \parallel_2}$$

and establishes the upper and lower bounds on the singular values of $\mathcal{J}(W, \tilde{X})$ over $S_*$ in terms of the singular values of $\mathcal{J}(W, C)$. Finally, the smoothness can be established similarly. Given matrices $W, \tilde{W}$, the rows of the difference $||\mathcal{J}(\tilde{W}, \tilde{X}) - \mathcal{J}(W, \tilde{X})||$ is obtained by duplicating the rows of $||\mathcal{J}(\tilde{W}, C) - \mathcal{J}(W, C)||$ by at most $c_{\text{up}} n / K$ times. Hence the spectral norm is scaled by at most $\sqrt{c_{\text{up}} n / K}$. ■

Lemma E.9 (Upper bound on initial misfit) Consider a one-hidden layer neural network model of the form $x \mapsto v^T \phi(W_0 x)$ where the activation $\phi$ has bounded derivatives obeying $|\phi(0)|, |\phi'(z)| \leq \Gamma$. Suppose entries of $v \in \mathbb{R}^K$ are half $1/\sqrt{K}$ and half $-1/\sqrt{K}$ so that $|v|_{\ell_2} = 1$. Also assume we have $n$ data points $x_1, x_2, \ldots, x_n \in \mathbb{R}^d$ with unit euclidean norm $(|x_i|_{\ell_2} = 1)$ aggregated as rows of a matrix $X \in \mathbb{R}^{n \times d}$ and the corresponding labels given by $y \in \mathbb{R}^n$ generated according to $(\rho, \varepsilon_0 = 0, \delta)$ noisy dataset (Definition 1.2). Then for $W_0 \in \mathbb{R}^{K \times d}$ with i.i.d. $\mathcal{N}(0, 1)$ entries

$$\|v^T \phi(W_0 X^T) - y\|_{\ell_2} \leq \mathcal{O}(\Gamma \sqrt{n \log K}),$$

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

Proof This lemma is based on a fairly straightforward union bound. First, by construction $|y|_{\ell_2} \leq \sqrt{n}$. What remains is bounding $\|v^T \phi(W_0 X^T)\|_{\ell_2}$. 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 $w$ be a row of $W_0$ and $x$ be a row of $X$. Since $\phi$ is $\Gamma$ Lipschitz and $|\phi(0)| \leq \Gamma$, each entry of $\phi(X w)$ is $\mathcal{O}(\Gamma)$-subgaussian. Hence $v^T \phi(W_0 x)$ is weighted average of $k$ i.i.d. subgaussians which are entries of $\phi(W_0 x)$. Additionally it is zero mean since $\Sigma_{i=1}^n v_i = 0$. Thus $v^T \phi(W_0 x)$ is also $\mathcal{O}(\Gamma)$ subgaussian and obeys

$$\mathbb{P}(\|v^T \phi(W_0 x)\| \geq c \Gamma \sqrt{\log K}) \leq K^{-101},$$

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

E.2.1. PROOF OF THEOREM E.14

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

Lemma E.10 Let $\{(x_i, y_i)\}_{i=1}^n$ 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 $\mathcal{J}(W, C)$ be the Jacobian at the cluster center matrix which is rank $K$ and $S_*$ be its column space. Then, the difference between noiseless and noisy labels satisfy the bound

$$\|\Pi_{S_*}(y - \tilde{y})\|_{\ell_\infty} \leq 2 \rho.$$

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

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

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

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

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

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

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

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

Since number of errors within cluster $\ell$ is at most $\rho \|e_\ell\|_{\ell_1}$, we find that

$$\|P e_\ell\|_{\ell_\infty} = \frac{1}{n_\ell} \|1(\ell) 1(\ell)^T e_\ell\|_{\ell_\infty} \leq \frac{\|e_\ell\|_{\ell_1}}{n_\ell} \leq 2 \rho.$$
The final line yields the bound
\[ \| \mathcal{P}_S(y - \hat{y}) \|_\infty = \| \mathcal{P}_S(e) \|_\infty = \| Pe \|_\infty \leq 2 \rho. \]

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

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

Using Corollary E.8, Assumption 3 holds with \( L = \sqrt{\frac{\text{cap}^n}{\epsilon K}} \) where \( L \) is the Lipschitz constant of Jacobian spectrum. Denote \( r_\tau = f(W_\tau) - y \). Using Lemma E.9 with probability \( 1 - K^{-100} \), we have that \( \| r_0 \|_{\ell_2} = \| y - f(W_0) \|_{\ell_2} \leq \Gamma \sqrt{c_0 n \log K / 128} \) for some \( c_0 > 0 \). Corollary E.8 guarantees a uniform bound for \( \beta \), hence in Assumption 2, we pick
\[ \beta \leq \sqrt{\frac{\text{cap}^n}{K}} \Gamma \| C \|. \]

We shall also pick the minimum singular value over \( S_\tau \) to be
\[ \alpha = \frac{\alpha_0}{2} \quad \text{where} \quad \alpha_0 = \sqrt{\frac{c_{\text{low}} n \lambda(C)}{2 K}}. \]

We wish to verify Assumption 2 over the radius of
\[ R = \frac{4}{\alpha} \| f(W_\tau) - y \|_{\ell_2}, \]
\[ \leq \Gamma \sqrt{c_0 n \log K / 8}, \]
\[ = \sqrt{\frac{c_0 n \log K / 2}{c_{\text{low}} n \lambda(C)}}, \]
\[ = \sqrt{\frac{c_0 K \log K}{c_{\text{low}} n \lambda(C)}}, \]

neighborhood of \( W_0 \). What remains is ensuring that Jacobian over \( S_\tau \) is lower bounded by \( \alpha \). Our choice of \( k \) guarantees that at the initialization, with probability \( 1 - K^{-100} \), we have
\[ \sigma_{\text{min}}(J(W_0, X), S_\tau) \geq \alpha_0. \]

Suppose \( LR \leq \alpha = \alpha_0 / 2 \). Using triangle inequality on Jacobian spectrum, for any \( W \in \mathcal{D} \), using \( \| W - W_0 \|_F \leq R \), we would have
\[ \sigma_{\text{min}}(J(W, X), S_\tau) \geq \sigma_{\text{min}}(J(W_0, X), S_\tau) - LR \geq \alpha_0 - \alpha = \alpha. \]

Now, observe that
\[ LR = \Gamma \sqrt{\text{cap}^n / K} \| C \| \sqrt{c_0 K \log(K)} \]
\[ = \Gamma^2 \sqrt{c_0 n \log K / c_{\text{low}} K \lambda(C)}, \]
\[ \leq \frac{\alpha_0}{2}, \]
\[ = \sqrt{\frac{c_{\text{low}} n \lambda(C)}{8 K}}, \]
as \( k \) satisfies
\[ k \geq O\left( 4 \sqrt{c_0 n \log(K)} \right) \geq O\left( \frac{4 \Gamma^4 K \log(K)}{\lambda(C)^2} \right). \]

Finally, since \( LR = 4L \| r_0 \|_{\ell_2} / \alpha \leq \alpha \), the learning rate is
\[ \eta \leq \frac{5}{2} \min \left( 1, \frac{\alpha \beta}{2 LR} \right) = \frac{K}{2} \min \left( 1, \frac{\alpha \beta}{2 LR} \right), \]
\[ \leq \tau, \]
\[ \| \Pi_{S_\tau}(y - y) \|_{\ell_\infty} \leq 2 \rho. \]

The final step is the properties of the label corruption. Using Lemma E.10, we find that
\[ \| \Pi_{S_\tau}(y - y) \|_{\ell_\infty} \leq 2 \rho. \]

Substituting the values corresponding to \( \alpha, \beta, L \) yields that, for all gradient iterations with
\[ \frac{5}{\eta \alpha^2} \log \left( \frac{\| r_0 \|_{\ell_2}}{2 \rho} \right) \leq \frac{5}{\eta \alpha^2} \log \left( \frac{\Gamma \sqrt{c_0 n \log K / 32}}{2 \rho} \right) \]
\[ = O\left( \frac{K}{\eta n \lambda(C)} \log \left( \frac{\Gamma \sqrt{c_0 n \log K}}{\rho} \right) \right) \leq \tau, \]
denoting the clean labels by \( y \) and applying Theorem C.2, we have that, the infinity norm of the residual obeys (using \( \| \Pi_{S_\tau}(e) \|_{\ell_\infty} \leq 2 \rho \))
\[ \| f(W) - \hat{y} \|_{\ell_\infty} \leq 4 \rho. \]

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

**E.2.2. Proof of Theorem E.15**

Consider
\[ f(W, x) = v^T \phi(W x) \]
and note that
\[
\nabla_x f(W, x) = W^T \text{diag}(\phi'(Wx)) v
\]

Thus
\[
\frac{\partial}{\partial x} f(W, x) u = v^T \text{diag}(\phi'(Wx)) W u = \sum_{\ell=1}^k v_\ell \phi'((w_\ell, x)) w_\ell^T u
\]

Thus, denoting vectorization of a matrix by vect(·)
\[
\operatorname{vect}(U)^T \left( \frac{\partial}{\partial \operatorname{vect}(W)} \frac{\partial}{\partial x} f(W, x) \right) u
\]
\[
= \sum_{\ell=1}^k v_\ell \left( \phi''(w_\ell^T x)(w_\ell^T u)(u_\ell^T x) + \phi'(w_\ell^T x)(u_\ell^T u) \right)
\]
\[
= u^T W^T \text{diag}(v) \text{diag}(\phi''(Wx)) Ux + v^T \text{diag}(\phi'(Wx)) U u \phi(g^T x_2) - \phi(g^T x_1) |_{\psi_2} 
\]
\[
= \|\phi(g^T x_2) - \phi(g^T x_1) |_{\psi_2}
\]
\[
= \|\phi'(tg^T x_2 + (1-t)g^T x_1) g^T(x_2 - x_1) |_{\psi_2}
\]
\[
\leq \Gamma \|g^T(x_2 - x_1) |_{\ell_2}
\]
\[
\leq c \Gamma \|x_2 - x_1 |_{\ell_2}
\]

Thus by the general mean value theorem there exists a point 
\((\tilde{W}, \tilde{x})\) in the square \((W_0, x_1), (W_0, x_2), (W, x_1)\) and 
\((W, x_2)\) such that

Using the above we have that
\[
(a) \left| f(W, x_2) - f(W_0, x_2) - (f(W, x_1) - f(W_0, x_1)) \right|
\]
\[
\leq \|x_2 - x_1 |_{\ell_2} \|\bar{W} - \bar{W}_0 \|_F \|x_2 - x_1 |_{\ell_2}
\]
\[
\leq \left( \frac{1}{\sqrt{k}} \|\bar{W} - \bar{W}_0 \|_F \right) \|x_2 - x_1 |_{\ell_2}
\]
\[
\leq CT \|x_2 - x_1 |_{\ell_2}
\]

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

Also note that
\[
(f(W_0, x_2) - f(W_0, x_1)) = v^T (\phi(W_0 x_2) - \phi(W_0 x_1))
\]
\[
= \sum_{\ell=1}^k v_\ell (\phi(g^T x_2) - \phi(g^T x_1) |_{\psi_2}
\]
\[
= \|\phi(tg^T x_2 + (1-t)g^T x_1) g^T(x_2 - x_1) |_{\psi_2}
\]
\[
\leq \Gamma \|g^T(x_2 - x_1) |_{\ell_2}
\]
\[
\leq c \Gamma \|x_2 - x_1 |_{\ell_2}
\]

where \( g_1, g_2, \ldots, g_k \) are i.i.d. vectors with \( \mathcal{N}(0, I_d) \) distribution. Also for \( v \) obeying \( 1^T v = 0 \) this random variable
Thus, using the fact that weighted sum of subGaussian random variables are subgaussian combined with (G.2) we conclude that $f(W_0, x_2) - f(W_0, x_1)$ is also subGaussian obeying $\|f(W_0, x_2) - f(W_0, x_1)\|_{\psi_2} \leq c\Gamma\|v\|_{\ell_2} \|x_2 - x_1\|_{\ell_2}$. Thus

$$\|f(W_0, x_2) - f(W_0, x_1)\| \leq c\Gamma\|v\|_{\ell_2} \|x_2 - x_1\|_{\ell_2} = c\Gamma\|x_2 - x_1\|_{\ell_2},$$

(E.34)

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

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

$$\delta \leq \|y_2 - y_2\| = \|f(W, x_1) - f(W, x_2)\| \leq \|v^T (\phi(Wx_2) - \phi(Wx_1))\| \leq \|v^T (f(W, x_2) - f(W, x_1)) - (f(W, x_1) - f(W, x_1))\| + \|v^T (\phi(Wx_2) - \phi(Wx_1))\| \leq c\Gamma \|x_2 - x_1\|_{\ell_2} \|W - W_0\|_{\ell_2} + c\Gamma \|x_2 - x_1\|_{\ell_2} \leq c\Gamma \|W - W_0\|_{\ell_2} + \frac{c\Gamma \|x_2 - x_1\|_{\ell_2}}{1000}$$

Thus

$$\|W - W_0\| \geq \frac{\delta}{c\Gamma \|x_2 - x_1\|_{\ell_2}} - \frac{t}{1000},$$

with high probability.

E.3. Perturbation analysis for perfectly clustered data

(Proof of Theorem 2.2)

Denote average neural net Jacobian at data X via

$$J(W_1, W_2, X) = \int_0^1 J(\alpha W_1 + (1 - \alpha) W_2, X) \, d\alpha.$$ 

Lemma E.11 (Perturbed Jacobian Distance) Let $X = [x_1 \ldots x_n]^T$ be the input matrix obtained from Definition 1.1. Let $\bar{x}$ be the noiseless inputs where $\bar{x}_i$ is the cluster center corresponding to $x_i$. Given weight matrices $W_1, W_2, \tilde{W}_1, \tilde{W}_2$, we have that

$$\|J(W_1, W_2, X) - J(\tilde{W}_1, \tilde{W}_2, \bar{x})\| \leq \Gamma \sqrt{n} \left( \frac{\|W_1 - W_1\|_F + \|\tilde{W}_2 - \tilde{W}_2\|_F}{2\sqrt{k}} + \varepsilon_0 \right).$$

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

$$\|J(W, X) - J(\tilde{W}, \bar{x})\| \leq \|J(W, X) - J(\tilde{W}, \bar{x})\|$$

We first bound

$$\|J(W, X) - J(\tilde{W}, X)\| = \|\text{diag}(v)\phi'(W X^T) \ast X^T - \text{diag}(v)\phi'(\tilde{W} X^T) \ast X^T\| = \frac{1}{\sqrt{k}} \|\phi'(W X^T) - \phi'(\tilde{W} X^T)\ast X^T\|$$

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

$$\|A \ast B\| = \sqrt{\|A^T A\| \ast \|B^T B\|} \leq \sqrt{\|A^T A\|} = \|A\|.$$ 

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

$$\|\phi'(W X^T) - \phi'(\tilde{W} X^T)\ast X^T\| \leq \|\phi'(W X^T) - \phi'(\tilde{W} X^T)\| \leq \|\phi'(W X^T)\ast X^T\| \leq \|X^T\|^{1/2} \|\phi'(W X^T)\| \leq \sqrt{n} \|X^T\|^{1/2} \|\phi'(W X^T)\|$$

Secondly,

$$\|J(\bar{W}, X) - J(\tilde{W}, \bar{x})\| = \frac{1}{\sqrt{k}} \|\phi'(\tilde{W} X^T) \ast (X - \bar{x})\|$$

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

$$\|\phi'(W X^T) \ast (X - \bar{x})\| \leq \Gamma \sqrt{n} \|X - \bar{x}\| \leq \Gamma \sqrt{k n \varepsilon_0}.$$

Combining both estimates yields

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

To get the result on $\|J(W_1, W_2, X) - J(\tilde{W}_1, \tilde{W}_2, X)\|$, we integrate

$$\|J(W_1, W_2, X) - J(\tilde{W}_1, \tilde{W}_2, X)\| \leq \int_0^1 \Gamma \sqrt{n} \left[ \alpha \|W_1 - W_1\|_F + (1 - \alpha) \|\tilde{W}_1 - W_1\|_F \right] d\alpha$$

$$+ \int_0^1 \Gamma \sqrt{n \varepsilon_0} d\alpha$$

$$\leq \Gamma \sqrt{n} \left( \frac{\|W_1 - W_1\|_F + \|\tilde{W}_2 - W_2\|_F}{2\sqrt{k}} + \varepsilon_0 \right).$$
Theorem E.12 (Robustness of gradient path) Generate samples \((x_i, y_i)_{i=1}^n\) according to \((\rho, \varepsilon_0, \delta)\) noisy dataset model and form the concatenated input/labels \(X \in \mathbb{R}^{d \times n}, y \in \mathbb{R}^n\). Let \(X\) be the clean input sample matrix obtained by mapping \(x_i\) to its associated cluster center. Set learning rate \(\eta \leq \frac{K}{\varepsilon_0 n^2 \log^2 C_{\lambda}}\) and maximum iterations \(\tau_0\) satisfying

\[
\eta \tau_0 = \frac{C_1 K}{\lambda(C)} \log(\frac{\sqrt{n \log K}}{\rho}).
\]

where \(C_1 \geq 1\) is a constant of our choice. Suppose input noise level \(\varepsilon_0\) and number of hidden nodes obey

\[
\varepsilon_0 \leq O\left(\frac{\lambda(C)}{\Gamma^2 K \log(\frac{\sqrt{n \log K}}{\rho})}\right)
\]

and

\[
k \geq O\left(\frac{\Gamma^2 K^2 |C|^4}{\lambda(C)^4} \log(\frac{\sqrt{n \log K}}{\rho})^6\right).
\]

Set \(W_0, v_{0,t} \sim N(0, 1).\) Starting from \(W_0 = \bar{W}_0\) consider the gradient descent iterations over the losses

\[
W_{t+1} = W_t - \eta \nabla L(W_t)
\]

where

\[
L(W) = \frac{1}{n} L(W) = \frac{1}{2} \sum_{i=1}^{n} (y_i - f(W, \bar{x}_i))^2
\]

\[
\bar{W}_{t+1} = W_t - \nabla \hat{L}(W_t)
\]

Then, for all gradient descent iterations satisfying \(\tau \leq \tau_0\), we have that

\[
\|f(W_t, X) - f(\bar{W}_t, \bar{X})\|_{\ell_2} \leq c_0 \eta \varepsilon_{0} \Gamma \sqrt{n \log K},
\]

and

\[
\|W_t - \bar{W}_t\|_F \leq \Omega(\tau \eta \varepsilon_{0} \Gamma^2 K n \log(\frac{\sqrt{n \log K}}{\rho})^2).
\]

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

\[
r_i = f(W_i, X) - y, \quad \hat{r}_i = f(\bar{W}_i, \bar{x}_i) - y,
\]

\[
\bar{J}_i = J(W_i, X), \quad \bar{J}_{i+1, i} = J(W_{i+1}, W_i, X), \quad \bar{J}_i = J(W_i, \bar{X}), \quad \bar{J}_{i+1, i} = J(W_{i+1}, \bar{W}_i, \bar{X}),
\]

\[
d_i = \|W_i - \bar{W}_i\|_F, \quad p_i = \|r_i - \hat{r}_i\|_F,
\]

\[
\beta = \sqrt{C_{\lambda} \sqrt{c_{\lambda} n / K}}, \quad L = \Gamma \sqrt{C_{\lambda} \sqrt{c_{\lambda} n / K k}}.
\]

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

Applying Lemma E.11, note that

\[
\|J(W_t, X) - J(\bar{W}_t, \bar{X})\| \leq L \|\bar{W} - W\|_F + \Gamma \sqrt{n \varepsilon_0} \leq L \Delta_t + \Gamma \sqrt{n \varepsilon_0}
\]

\[
\|J(W_{t+1}, W_t, X) - J(\bar{W}_{t+1}, \bar{W}_t, \bar{X})\| \leq L (\Delta_t + d_{t+1})^2 + \Gamma \sqrt{n \varepsilon_0}.
\]

Following this and using that noiseless residual is non-increasing and satisfies \(\|r_t\|_{\ell_2} \leq \|\hat{r}_0\|_{\ell_2}\), note that parameter satisfies

\[
W_{t+1} = W_t - \eta \bar{J}_t r_t, \quad \hat{W}_{t+1} = \hat{W}_t - \eta \bar{J}_t \hat{r}_t
\]

\[
\|W_{t+1} - \hat{W}_{t+1}\|_F \leq \|W_t - \hat{W}_t\|_F + \eta \|J_t - \bar{J}_t\|_F \|\hat{r}_t\|_{\ell_2}
\]

\[
+ \eta \bar{J}_t \|r_t - \hat{r}_t\|_{\ell_2}
\]

\[
d_{t+1} \leq d_t + \eta ((L \Delta_t + \Gamma \sqrt{n \varepsilon_0})) \|\hat{r}_0\|_{\ell_2} + \beta p_t,
\]

and residual satisfies (using \(I \geq \bar{J}_{i+1, i} \bar{J}_i^T / \beta^2 \geq 0\))

\[
r_{i+1} = r_i - \eta \bar{J}_{i+1, i} \bar{J}_i^T r_i \implies \]

\[
r_{i+1} - \hat{r}_{i+1} = (r_i - \hat{r}_i) - \eta (\bar{J}_{i+1, i} - \bar{J}_{i+1, i}) \bar{J}_i^T r_i
\]

\[
- \eta \bar{J}_{i+1, i} \bar{J}_i^T (r_i - \hat{r}_i).
\]

\[
r_{i+1} - \hat{r}_{i+1} = (I - \eta \bar{J}_{i+1, i} \bar{J}_i^T)(r_i - \hat{r}_i)
\]

\[
- \eta \bar{J}_{i+1, i} \bar{J}_i^T r_i
\]

\[
- \eta \bar{J}_{i+1, i} \bar{J}_i^T (r_i - \hat{r}_i).
\]

\[
\|r_{i+1} - \hat{r}_{i+1}\|_{\ell_2} \leq \|r_i - \hat{r}_i\|_{\ell_2}
\]

\[
+ \eta \beta \|r_i\|_{\ell_2} (L(3d_t + d_{t+1})/2 + 2 \Gamma \sqrt{n \varepsilon_0}).
\]

\[
r_{i+1} - \hat{r}_{i+1} \leq \|r_i - \hat{r}_i\|_{\ell_2}
\]

\[
+ \eta \beta L/2 (\|\hat{r}_0\|_{\ell_2} + p_i)(3d_t + d_{t+1})
\]

\[
+ 2 \Gamma \sqrt{n \eta \beta} (\|\hat{r}_0\|_{\ell_2} + p_i) \varepsilon_0.
\]

where we used \(\|r_i\|_{\ell_2} \leq p_i + \|\hat{r}_0\|_{\ell_2}\) and \(\|J_t - \bar{J}_t\|_F \|v\|_{\ell_2} \leq \|v\|_{\ell_2}\) which follows from (E.23). This implies

\[
p_{i+1} \leq p_i + \eta \beta (\|\hat{r}_0\|_{\ell_2} + p_i)(L(3d_t + d_{t+1})/2 + 2 \Gamma \sqrt{n \varepsilon_0}).
\]

Finalizing proof: Next, using Lemma E.9, we have \(\|\hat{r}_0\|_{\ell_2} \leq \Theta := C_0 \sqrt{n \log K}.\) We claim that if

\[
\varepsilon_0 \leq \frac{1}{8 \tau \eta \beta \Gamma \sqrt{n}} \leq \frac{1}{8 \tau \eta \beta \Gamma \sqrt{n}}
\]

\[
2 \leq \frac{1}{5 \tau \eta \beta \Theta (1 + 8 \tau \theta \beta^2)} \leq 30 (\tau \eta \beta \Theta)
\]

(where we used \(\eta \theta \beta^2 \geq 1\), for all \(t \leq \tau_0\), we have that

\[
p_t \leq 8 \eta \Gamma \sqrt{n \varepsilon_0} \Theta \beta \leq \Theta
\]

\[
d_t \leq 2 \eta \Gamma \sqrt{n \varepsilon_0} \Theta (1 + 8 \tau \eta \beta^2).
\]
The proof is by induction. Suppose it holds until $t \leq \tau_0 - 1$.
At $t + 1$, via (E.37) we have that
$$\frac{d_{t+1} - d_t}{\eta} \leq Ld_t \Theta + \Gamma \sqrt{n\varepsilon_0 \Theta} + 8\tau_0 \eta \beta^2 \Gamma \sqrt{n\varepsilon_0 \Theta}$$
$$\leq 2\Gamma \sqrt{n\varepsilon_0 \Theta} (1 + 8\eta \tau_0 \beta^2).$$

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

Next, we show the induction on $p_t$. Observe that $3d_t + d_{t+1} \leq 10\tau_0 \eta \Gamma \sqrt{n\varepsilon_0 \Theta} (1 + 8\eta \tau_0 \beta^2)$. Following (E.39) and using $p_t \leq \Theta$, we need
$$\frac{p_{t+1} - p_t}{\eta} \leq \beta \Theta (L(3d_t + d_{t+1}) + 4\Gamma \sqrt{n\varepsilon_0 \Theta})$$
$$\leq 8\Gamma \sqrt{n\varepsilon_0 \Theta} \beta \leftarrow$$
$$L(3d_t + d_{t+1}) + 4\Gamma \sqrt{n\varepsilon_0 \Theta} \leftarrow$$
$$L(3d_t + d_{t+1}) \leq 4\Gamma \sqrt{n\varepsilon_0 \Theta} \leftarrow$$
$$10L\tau_0 \eta (1 + 8\eta \tau_0 \beta^2) \Theta \leq 4 \leftarrow$$
$$L \leq \frac{2}{5\tau_0 \eta (1 + 8\eta \tau_0 \beta^2) \Theta}.$$

Concluding the induction since $L$ satisfies the final line. Consequently, for all $0 \leq t \leq \tau_0$, we have that
$$p_t \leq 8\tau_0 \Gamma \sqrt{n\varepsilon_0 \Theta} \beta = c_0 \tau_0 \eta \Gamma^3 \sqrt{n} \sqrt[3]{\sqrt{\log K}}.$$

Next, note that, condition on $L$ is implied by
$$k \geq 1000 \Gamma^2 n (\tau_0 \eta \beta)^4 \Theta^2$$
$$\leq \mathcal{O}(\Gamma^4 n \frac{K^4}{n^4 \lambda(C)^4} \log(\frac{\sqrt{n} \log K}{\rho}))^4$$
$$\cdot \left(\frac{1}{\sqrt{n}} \Gamma \sqrt{\eta} \right)^4 \left(\frac{1}{\sqrt{n}} \log K \right)^2$$
$$\leq \mathcal{O}(\Gamma^4 \frac{K^4}{\lambda(C)^4} \log(\frac{\sqrt{n} \log K}{\rho}))^4 \log^2(K))$$
which is implied by $k \geq \mathcal{O}(\Gamma^4 \frac{K^4}{\lambda(C)^4} \log(\frac{\sqrt{n} \log K}{\rho}))^6$.

Finally, following (E.40), distance satisfies
$$d_t \leq 10 \eta \tau_0 \Theta \left(\frac{\Gamma^4 K^4}{\lambda(C)^4} \log(\frac{\sqrt{n} \log K}{\rho})\right)^2$$
$$\leq \mathcal{O}(\rho \tau_0 \Theta \left(\frac{\Gamma^4 \sqrt{n} \log K}{\lambda(C)^4} \log(\frac{\sqrt{n} \log K}{\rho})\right)^2).$$

### Theorem E.13 (Training neural nets with corrupted labels)
Let $\{(x_i, y_i)\}_{i=1}^n$ 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 $|\phi(0)|, |\phi'|, |\phi''| \leq \Gamma$ for some $\Gamma \geq 1$, input noise and the number of hidden nodes satisfy
$$\varepsilon_0 \leq \mathcal{O}\left(\frac{\lambda(C)}{\Gamma^2 K \log(\frac{\sqrt{n} \log K}{\rho})}\right)$$
$$\leq \mathcal{O}\left(\frac{\Gamma^4 K^2 |C|^4}{\lambda(C)^4} \log(\frac{\sqrt{n} \log K}{\rho})^6\right).$$

where $C \in \mathbb{R}^{K \times d}$ is the matrix of cluster centers. Set learning rate $\eta \leq \frac{K}{2c_{app} n \Gamma^2 |C|^2}$ and randomly initialize $W_0 \sim N(0, 1)$. With probability $1 - 3/K^{100}$, after $\tau = \mathcal{O}\left(\frac{K}{\rho n \lambda(C)}\right) \log(\frac{\sqrt{n} \log K}{\rho})$ iterations, for all $1 \leq i \leq n$, we have that
- The per sample normalized $\ell_2$ norm bound satisfies
  $$\|f(W, X) - \tilde{y}_i\|_{\ell_2} \leq \mathcal{O}(\frac{\varepsilon_0 K^3 \sqrt{n} \log K}{\lambda(C)} \log(\frac{\sqrt{n} \log K}{\rho})^4 + 4 \rho).$$

- Suppose $\rho \leq \delta/8$. Denote the total number of prediction errors with respect to true labels (i.e. not satisfying (E.46)) by $\text{err}(\mathbf{W})$. With same probability, $\text{err}(\mathbf{W})_\tau$ obeys
  $$\text{err}(\mathbf{W})_\tau \leq \mathcal{O}(\frac{\text{err}(\mathbf{W})_0}{\rho} + c_0 K \Gamma^3 \sqrt{n} \log K \log(\frac{\sqrt{n} \log K}{\rho})).$$

- Suppose $\rho \leq \delta/8$ and $\varepsilon_0 \leq c' \frac{\delta \lambda(C)^2}{\Gamma^4 K^2 n \log(\frac{\sqrt{n} \log K}{\rho})^3}$, then, $W_\tau$ assigns all input samples $x_i$ to correct ground truth labels $y_i$, i.e. (E.46) holds for all $1 \leq i \leq n$.

- Finally, for any iteration count $0 \leq t \leq \tau$ the total distance to initialization is bounded as
  $$\|W - W_0\|_F \leq \mathcal{O}(\sqrt{\frac{\sqrt{n} \log K}{\lambda(C)}} + t \eta \varepsilon_0 \frac{K^4 n}{\lambda(C)} \log(\frac{\sqrt{n} \log K}{\rho})^2).$$

#### Proof
Note that proposed number of iterations $\tau$ is set so that it is large enough for Theorem E.14 to achieve small error in the clean input model ($\varepsilon_0 = 0$) and it is small enough so that Theorem E.12 is applicable. In light of Theorems E.12 and E.14 consider two gradient descent iterations starting from $W_0$ where one uses clean dataset (as if input vectors are perfectly cluster centers) $\mathbf{X}$ and other uses the
original dataset $X$. Denote the prediction residual vectors of the noiseless and original problems at time $\tau$ with respect true ground truth labels $\hat{y}$ by $\hat{r}_\tau = f(W_\tau, X) - \hat{y}$ and $r_\tau = f(W_\tau, X) - \hat{y}$ respectively. Applying Theorems E.12 and E.14, under the stated conditions, we have that
\[
|\hat{r}_\tau|_{\ell_\infty} \leq 4\rho \quad \text{and} \quad (E.42)
\]
\[
|r_\tau - \hat{r}_\tau|_{\ell_2} \leq \varepsilon_0 \frac{K}{n\lambda(C)} \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho}) I^{3/2} \sqrt{\log K} 
\]
\[
= c \varepsilon_0 I^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)} \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho}) 
\]
\[
= \varepsilon_0 I^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)} \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho}) 
\]
\[
(E.43)
\]
\[
(E.44)
\]

**First statement:** The latter two results imply the $\ell_2$ error bounds on $r_\tau = f(W_\tau, X) - \hat{y}$.

**Second statement:** To assess the classification rate we count the number of entries of $r_\tau = f(W_\tau, X) - \hat{y}$ that is larger than the class margin $\delta/2$ in absolute value. Suppose $\rho \leq \delta/8$. Let $\mathcal{Z}$ be the set of entries obeying this. For $i \in \mathcal{Z}$ using $|\hat{r}_\tau|_{\ell_\infty} \leq 4\rho \leq \delta/4$, we have
\[
|r_{\tau,i}| \geq \delta/2 \implies |r_{\tau,i}| + |r_{\tau,i} - \hat{r}_{\tau,i}| \geq \delta/2 
\]
\[
\implies |r_{\tau,i} - \hat{r}_{\tau,i}| \geq \delta/4.
\]

Consequently, we find that
\[
|r_\tau - \hat{r}_\tau|_{\ell_1} \geq |\mathcal{Z}|\delta/4.
\]

Converting $\ell_2$ upper bound on the left hand side to $\ell_1$, we obtain
\[
|\mathcal{Z}| \leq \varepsilon_0 I^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)} \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho}) \geq |\mathcal{Z}|\delta/4.
\]

Hence, the total number of errors is at most
\[
|\mathcal{Z}| \leq \varepsilon_0 I^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)} \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho})
\]
\[
(E.45)
\]

**Third statement – Showing zero error:** Pick an input sample $x$ from dataset and its clean version $\bar{x}$. We will argue that $f(W_\tau, x) - f(W_\tau, \bar{x})$ is smaller than $\delta/4$ when $\varepsilon_0$ is small enough. We again write
\[
|f(W_\tau, x) - f(W_\tau, \bar{x})| \leq |f(W_\tau, x) - f(W_\tau, \bar{x})|
\]
\[
+ |f(W_\tau, x) - f(W_\tau, \bar{x})|
\]
\[
= |f(W_\tau, x) - f(W_\tau, \bar{x})|
\]
\[
\leq \Gamma \|W_\tau - \bar{W}_\tau\|_{F}
\]
\[
\leq O(\varepsilon_0 I^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)} \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho})^3)
\]

Next, we need to bound
\[
|f(\bar{W}_\tau, x) - f(W_\tau, \bar{x})| \leq \|v^T \phi(W_\tau, x) - v^T \phi(W_\tau, \bar{x})
\]
\[
= \|\tilde{v}^T \phi(W_\tau, x) - \phi(W_\tau, x)\|_{\ell_2}
\]
\[
\leq \Gamma \|W_\tau - \bar{W}_\tau\|_{F}
\]
\[
\leq O(\varepsilon_0 I^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)} \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho})^3)
\]

Finally, we have
\[
|\bar{W}_\tau - W_\tau|_{F} \leq |\bar{W}_\tau - \bar{W}_\tau|_{F} + |\bar{W}_\tau - W_\tau|_{F}
\]
\[
(E.46)
\]

and applying an argument similar to Theorem E.15 (detailed in Appendix G), with probability at least $1 - 1/n^{100}$, we find that
\[
|f(W_\tau, x) - f(W_\tau, \bar{x})| \leq C\Gamma \varepsilon_0 \left(\|\bar{W}_\tau - W_\tau\|_{F} + \sqrt{\log n}\right),
\]
\[
\leq C\Gamma \varepsilon_0 \left(\|\bar{W}_\tau - W_\tau\|_{F} + \sqrt{\log n}\right).
\]

Combining the two bounds above we get
\[
|f(W_\tau, x) - f(W_\tau, \bar{x})| \leq \varepsilon_0 O(\Gamma^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)})^3 \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho})
\]
\[
+ \Gamma \left(\|\bar{W}_\tau - W_\tau\|_{F} + \sqrt{\log n}\right)
\]
\[
\leq \varepsilon_0 O(\Gamma^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)})^3 \log(\frac{\Gamma n \log \frac{K}{\lambda(C)}}{\rho})^3)
\]

Hence, if $\varepsilon_0 \leq C\Gamma \varepsilon_0 \left(\|\bar{W}_\tau - W_\tau\|_{F} + \sqrt{\log n}\right)^3$, we obtain that, for all $1 \leq i \leq n$,
\[
|f(W_\tau, x_i) - \bar{y}_i| \leq |f(W_\tau, \bar{x}_i) - f(W_\tau, x_i)|
\]
\[
+ |f(W_\tau, x_i) - f(W_\tau, \bar{x}_i)|
\]
\[
\leq 4\rho + \frac{\delta}{4}.
\]

If $\rho \leq \delta/8$, we obtain
\[
|f(W_\tau, x_i) - \bar{y}_i| \leq \delta/2
\]

hence, $W_\tau$ outputs the correct decision for all samples.

**Fourth statement – Distance:** This follows from the triangle inequality
\[
|\bar{W}_\tau - W_\tau|_{F} \leq |\bar{W}_\tau - \bar{W}_\tau|_{F} + |\bar{W}_\tau - W_\tau|_{F}
\]
\[
(E.47)
\]

We have that right hand side terms are at most $O(\Gamma^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)})^3$ and $O(\Gamma^{3/2} \frac{K \sqrt{n \log K}}{\lambda(C)})^3$ from Theorems E.12 and E.14 respectively. This implies (E.41).
**Definition 2.1.** Then, with probability at least $1 - 2/K^{100}$ over randomly initialized $W_0 \sim N(0,1)$, the iterates $W_\tau$ obey the following properties.

- The distance to initial point $W_0$ is upper bounded by
  \[ ||W_\tau - W_0||_F \leq c_1 \sqrt{K \log K \lambda(C)}. \]

- After $\tau \geq \tau_0 := \frac{\log \sqrt{n} \log K}{\rho \log K}$ iterations, the entrywise predictions of the learned network with respect to the ground truth labels $\{y_i\}_{i=1}^n$ satisfy
  \[ |f(W_\tau, x_i) - \tilde{y}_i| \leq 4\rho, \]
  for all $1 \leq i \leq n$. Furthermore, if the noise level $\rho$ obeys $\rho \leq s/8$ the network predicts the correct label for all samples i.e.
  \[ \arg \min_{\alpha \in \mathbb{R}^d} |f(W_\tau, x_i) - \alpha_d| = \tilde{y}_i \quad \text{for} \quad i = 1, 2, \ldots, n. \quad \text{(E.46)} \]

This result shows that in the limit $\epsilon_0 \to 0$ where the data points are perfectly clustered, the required amount of over-parameterization can be reduced from $kd \geq K^4$ to $kd \geq K^2$.

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

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

In this section we wish to provide further insight into why early stopping enables robustness and generalizability solutions. Our main insight is that while a neural network may be 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 $\epsilon_0$ 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 $W$ with i.i.d. standard normal entries which ensures that the network is isometric i.e. given input $x$, $\mathbb{E}[f(W,x)^2] = O(||x||^2_F)$.

**Theorem E.15** Let $x_1, x_2 \in \mathbb{R}^d$ be two vectors with unit Euclidean norm obeying $||x_2 - x_1||_F \leq \epsilon_0$. Let $f(W,x) = \nu^T \phi(Wx)$ where $\nu$ is fixed, $W \in \mathbb{R}^{k \times d}$, and $k \geq cd$ with $c > 0$ a fixed constant. Assume $|\phi'|, |\phi''| \leq \Gamma$. Let $y_1$ and $y_2$ be two scalars satisfying $|y_2 - y_1| \geq \delta$. Suppose $W_0 \sim N(0,1)$, $W_0 \sim N(0,1)$. Then, with probability at least $1 - 2e^{-(k+cd)} - 2e^{-5\tau_0^2}$, for any $W \in \mathbb{R}^{k \times d}$ such that $|W - W_0|_F \leq c\sqrt{k}$ and
  \[ f(W, x_1) = y_1 \quad \text{and} \quad f(W, x_2) = y_2, \]
holds, we have
  \[ ||W - W_0||_F \geq \frac{\delta}{CT\epsilon_0} - \frac{t}{1000}. \]

In words, this result shows that in order to fit to a data set with a single corrupted label, a randomly initialized network has to traverse a distance of at least $\delta/\epsilon_0$. The next lemma clarifies the role of the corruption amount $s$ and shows that more label corruption within a fixed class requires a model with a larger norm in order to fit the labels. For this result we consider a randomized model with $\epsilon_0^2$ input noise variance.

**Lemma E.16** Let $c \in \mathbb{R}^d$ be a cluster center. Consider $2s$ data points $\{x_i\}_{i=1}^s$ and $\{\bar{x}_i\}_{i=1}^s$ in $\mathbb{R}^d$ generated i.i.d. around $c$ according to the following distribution
  \[ c + g \quad \text{with} \quad g \sim N(0, \frac{\epsilon_0^2}{d} I_d). \]
Assign $\{x_i\}_{i=1}^s$ with labels $y_i = y$ and $\{\bar{x}_i\}_{i=1}^s$ with labels $\bar{y}_i = \bar{y}$ and assume these two labels are $\delta$ separated i.e. $|y - \bar{y}| \geq \delta$. Also suppose $s \leq d$ and $|\phi'| \leq \Gamma$. Then, any $W \in \mathbb{R}^{k \times d}$ satisfying
  \[ f(W, x_i) = y_i \quad \text{and} \quad f(W, \bar{x}_i) = \bar{y}_i \quad \text{for} \quad i = 1, \ldots, s, \]
obes $||W||_F \geq \frac{s\delta}{\delta \epsilon_0}$ with probability at least $1 - e^{-d/2}$.
Unlike Theorem E.15 this result lower bounds the network norm in lieu of the distance to the initialization $W_0$. However, using the triangular inequality we can in turn get a guarantee on the distance from initialization $W_0$ via triangular inequality as long as $\|W_0\|_F \leq O(\sqrt{\delta}/\varepsilon_0)$ (e.g. by choosing a small $\varepsilon_0$).

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

$$\|W_\tau - W_0\|_F \geq \sqrt{\frac{m}{K}} \frac{\delta}{\varepsilon_0},$$

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 $\tau$ iterates. This demonstrates that there is a gap in the required distance to initialization for fitting enough to generalize and overfitting.

To sum up, our results highlight that, one can find a network with good generalization capabilities and robustness to label corruption within a small neighborhood of the initialization and that the size of this neighborhood is independent of the corruption. However, to fit to the corrupted labels, one has to travel much more, increasing the search space and likely decreasing generalization ability. Thus, early stopping can enable robustness without overfitting by restricting the distance to the initialization.

**F. Proof of Lemma E.16**

Create two matrices $X \in \mathbb{R}^{n \times d}$ and $\tilde{X} \in \mathbb{R}^{n \times d}$ by concatenating the input samples. Note that the matrix $X - \tilde{X}$ 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

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

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

$$\sqrt{\delta} \leq \|y - \tilde{y}\|_{\ell_2} = \|f(W, X) - f(W, \tilde{X})\|_{\ell_2},$$

$$= \|v^T(\phi(WX) - \phi(W\tilde{X}))\|_{\ell_2},$$

$$\leq \Gamma \|v\|_{\ell_2} \|X - \tilde{X}\|_F,$$

$$\leq \Gamma \|X - \tilde{X}\| \|W\|_F \leq 5\Gamma \varepsilon_0 \|W\|_F.$$

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

**G. Single label perturbation**

Note that

$$|f(W, x) - f(W, \tilde{x})|$$

$$= |v^T(\phi(Wx) - \phi(W\tilde{x}))|$$

$$\leq |v^T(\phi(Wx) - \phi(W\tilde{x}))| - |v^T(\phi(W_0x) - \phi(W_0\tilde{x}))|$$

$$+ |v^T(\phi(W_0x) - \phi(W_0\tilde{x}))|,$$

To continue note that by the general mean value theorem there exists a point $(\tilde{W}, \tilde{x})$ in the square $(W_0, x), (W_0, \tilde{x}), (W, x), \text{and} (W, \tilde{x})$ such that

$$|f(W, x) - f(W_0, x)| = |(W - W_0)|_{\ell_2} \left|\frac{\partial}{\partial \tilde{x}} \phi'(\tilde{W}\tilde{x})\right|_{\ell_2}$$

Using the above we have that

$$\left|\left|f(W, x) - f(W_0, x)\right| - (f(W, \tilde{x}) - f(W_0, \tilde{x})\right|$$

$$\leq \left|\left|\phi'(\tilde{W}\tilde{x})\right| - (f(W_0, \tilde{x}) - f(W_0, \tilde{x})\right|$$

$$\leq \left|\left|\phi'(\tilde{W}\tilde{x})\right| - \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right| - \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right| - \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right|$$

$$\leq \left|\phi'(\tilde{W}\tilde{x})\right| + \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right| - \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right| + \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right| + \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right|$$

$$\leq \left|\phi'(\tilde{W}\tilde{x})\right| + \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right| + \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right| + \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right| + \Gamma \left|\left|\phi'(\tilde{W}\tilde{x})\right|$$

Here, (a) follows from the triangle inequality, (b) from simple algebraic manipulations along with the fact that $|\phi'(z)| \leq \Gamma$ and $|\phi''(z)| \leq \Gamma$, (c) from the fact that $v_\ell = \frac{1}{\sqrt{\ell}}$, (d) from $\|x\|_{\ell_2} = \|\tilde{x}\|_{\ell_2} = 1$ which implies $\|\tilde{X}\|_{\ell_2} \leq 1$, (e) from triangular inequality, (f) from the fact that Frobenius norm dominates the spectral norm, (g) from the fact that with probability at least $1 - 2e^{-(d+\ell)}$, $\|W_0\| \leq 2(\sqrt{\ell} + \sqrt{d})$, and...
(h) from the fact that $\|\mathbf{W} - \mathbf{W}_0\| \leq \|\mathbf{W} - \mathbf{W}_0\|_F \leq \bar{c} \sqrt{k}$
and $k \geq cd$.

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

$$\|\phi(g^T x) - \phi(g^T \bar{x})\|_2^2 = \|\phi(g^T x) - \phi(g^T \bar{x})\|_2^2 = \|\phi'(t g^T x + (1-t)g^T \bar{x}) g^T (x - \bar{x})\|_2^2 \leq \Gamma \|g^T (x - \bar{x})\|_2 \leq c \Gamma \|x - \bar{x}\|_{L_2}.$$  \hfill (G.2)

Also note that $f(W_0, x) - f(W_0, \bar{x}) = v^T (\phi(W_0 x) - \phi(W_0 \bar{x}))$

$$\sim \sum_{\ell=1}^k v_\ell (\phi(g_\ell^T x) - \phi(g_\ell^T \bar{x}))$$

where $g_1, g_2, \ldots, g_k$ are i.i.d. vectors with $\mathcal{N}(0, I_d)$ distribution. Also for $v$ obeying $1^T v = 0$ this random variable has mean zero. Hence, using the fact that weighted sum of subGaussian random variables are subgaussian combined with (G.2) we conclude that $f(W_0, x) - f(W_0, \bar{x})$

is also subGaussian obeying $\|f(W_0, x) - f(W_0, \bar{x})\|_2 \leq c \Gamma \|v\|_{L_2} \|x - \bar{x}\|_{L_2}$. Thus

$$|f(W_0, x) - f(W_0, \bar{x})| \leq c \Gamma \|v\|_{L_2} \|x - \bar{x}\|_{L_2} \leq c \Gamma \|x - \bar{x}\|_{L_2}, \hfill (G.3)$$

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

$$|f(W_0, x_i) - f(W_0, \bar{x}_i)| \leq 2c \Gamma \sqrt{\log n} \|x_i - \bar{x}_i\|_{L_2},$$

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

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