

# Generalization Guarantees for Neural Networks via Harnessing the Low-rank Structure of the Jacobian

Samet Oymak\*    Zalan Fabian<sup>†α</sup>    Mingchen Li<sup>‡α</sup>    Mahdi Soltanolkotabi<sup>†</sup>

June 13, 2019

## Abstract

Modern neural network architectures often generalize well despite containing many more parameters than the size of the training dataset. This paper explores the generalization capabilities of neural networks trained via gradient descent. We develop a data-dependent optimization and generalization theory which leverages the low-rank structure of the Jacobian matrix associated with the network. Our results help demystify why training and generalization is easier on clean and structured datasets and harder on noisy and unstructured datasets as well as how the network size affects the evolution of the train and test errors during training. Specifically, we use a control knob to split the Jacobian spectrum into “information” and “nuisance” spaces associated with the large and small singular values. We show that over the information space learning is fast and one can quickly train a model with zero training loss that can also generalize well. Over the nuisance space training is slower and early stopping can help with generalization at the expense of some bias. We also show that the overall generalization capability of the network is controlled by how well the label vector is aligned with the information space. A key feature of our results is that even constant width neural nets can provably generalize for sufficiently nice datasets. We conduct various numerical experiments on deep networks that corroborate our theoretical findings and demonstrate that: (i) the Jacobian of typical neural networks exhibit low-rank structure with a few large singular values and many small ones leading to a low-dimensional information space, (ii) over the information space learning is fast and most of the label vector falls on this space, and (iii) label noise falls on the nuisance space and impedes optimization/generalization.

## 1 Introduction

### 1.1 Motivation and contributions

Deep neural networks (DNN) are ubiquitous in a growing number of domains ranging from computer vision to healthcare. State-of-the-art DNN models are typically overparameterized and contain more parameters than the size of the training dataset. It is well understood that in this overparameterized regime, DNNs are highly expressive and have the capacity to (over)fit arbitrary training datasets including pure noise [59]. Mysteriously however neural network models trained via simple algorithms such as (stochastic) gradient descent continue to predict well or *generalize* on yet unseen test data. In this paper we wish to take a step towards demystifying this phenomenon and help explain why neural nets can overfit to noise yet have the ability to generalize when real data sets are used for training. In particular we explore the generalization dynamics of neural nets trained via gradient descent. Using the Jacobian mapping associated to the neural network we characterize directions where learning is fast and generalizable versus directions where learning is slow and leads to overfitting. The main contributions of this work are as follows.

---

\*Department of Electrical and Computer Engineering, University of California, Riverside, CA

<sup>†</sup>Ming Hsieh Department of Electrical Engineering, University of Southern California, Los Angeles, CA

<sup>‡</sup>Department of Computer Science and Engineering, University of California, Riverside, CA

<sup>α</sup>Equal contribution.

- **Leveraging dataset structure:** We develop new optimization and generalization results that can harness the low-rank representation of semantically meaningful datasets via the Jacobian mapping of the neural net. This sheds light as to why training and generalization is easier using datasets where the features and labels are semantically linked versus others where there is no meaningful relationship between the features and labels (even when the same network is used for training).
- **Bias–variance tradeoffs:** We develop a bias–variance theory based on the Jacobian which decouples the learning process into *information* and *nuisance* spaces. We show that gradient descent almost perfectly interpolates the data over the information space (incurring only a small bias). In contrast, optimization over the nuisance space is slow and results in overfitting due to higher variance.
- **Network size vs prediction bias:** We obtain data-dependent tradeoffs between the network size and prediction bias. Specifically, we show that larger networks result in smaller prediction bias, but small networks can still generalize well, especially when the dataset is sufficiently structured, but typically incur a larger bias. This is in stark contrast to recent literature on optimization and generalization of neural networks [4, 21, 2, 15, 40, 1, 14] where guarantees only hold for very wide networks with the width of the network growing inversely proportional to the distance between the input samples or class margins or related notions. See Section 3.4 for further detail.
- **Pretrained models:** In our framework we do not require the initialization to be random and our results continue to apply even with arbitrary initialization. Therefore, our results may shed light on the generalization capabilities of networks initialized with pre-trained models such as those commonly used in meta/transfer learning.

## 1.2 Model and training

In our theoretical analysis we focus on neural networks consisting of one hidden layer with  $d$  input features,  $k$  hidden neurons and  $K$  outputs as depicted in Figure 4. We use  $\mathbf{W} \in \mathbb{R}^{k \times d}$  and  $\mathbf{V} \in \mathbb{R}^{K \times k}$  to denote the input-to-hidden and hidden-to-output weights. The overall input-output relationship of the neural network in this case is a function  $f(\cdot; \mathbf{W}) : \mathbb{R}^d \rightarrow \mathbb{R}^K$  that maps an input vector  $\mathbf{x} \in \mathbb{R}^d$  into an output of size  $K$  via

$$\mathbf{x} \mapsto f(\mathbf{x}; \mathbf{W}) := \mathbf{V}\phi(\mathbf{W}\mathbf{x}). \quad (1.1)$$

Given a dataset consisting of  $n$  feature/label pairs  $(\mathbf{x}_i, \mathbf{y}_i)$  with  $\mathbf{x}_i \in \mathbb{R}^d$  representing the features and  $\mathbf{y}_i \in \mathbb{R}^K$  the associated labels representing one of  $K$  classes with one-hot encoding (i.e.  $\mathbf{y}_i \in \{\mathbf{e}_1, \mathbf{e}_2, \dots, \mathbf{e}_K\}$  where  $\mathbf{e}_\ell \in \mathbb{R}^K$  has all zero entries except for the  $\ell$ th entry which is equal to one). To learn this dataset, we fix the output layer and train over  $\mathbf{W}$  via<sup>1</sup>

$$\min_{\mathbf{W} \in \mathbb{R}^{k \times d}} \mathcal{L}(\mathbf{W}) := \frac{1}{2} \sum_{i=1}^n \|\mathbf{V}\phi(\mathbf{W}\mathbf{x}_i) - \mathbf{y}_i\|_{\ell_2}^2. \quad (1.2)$$

It will be convenient to concatenate the labels and prediction vectors as follows

$$\mathbf{y} = \begin{bmatrix} \mathbf{y}_1 \\ \vdots \\ \mathbf{y}_n \end{bmatrix} \in \mathbb{R}^{nK} \quad \text{and} \quad f(\mathbf{W}) = \begin{bmatrix} \mathbf{V}f(\mathbf{x}_1; \mathbf{W}) \\ \vdots \\ \mathbf{V}f(\mathbf{x}_n; \mathbf{W}) \end{bmatrix} \in \mathbb{R}^{nK}. \quad (1.3)$$

Using this shorthand we can rewrite the loss (1.2) as

$$\min_{\mathbf{W} \in \mathbb{R}^{k \times d}} \mathcal{L}(\mathbf{W}) := \frac{1}{2} \|\mathbf{V}f(\mathbf{W}) - \mathbf{y}\|_{\ell_2}^2. \quad (1.4)$$

---

<sup>1</sup>For clarity of exposition, we focus only on optimizing over the input layer. However, as shown in the supplementary material, the technical approach is quite general and applies to arbitrary multiclass nonlinear least-squares problems. In particular, the proofs are stated so as to apply (or easily extend) to one-hidden layer networks where both layers are used for training. These results when combined can be used to prove variations of Theorems 3.2 and 3.3 when both layers are trained.

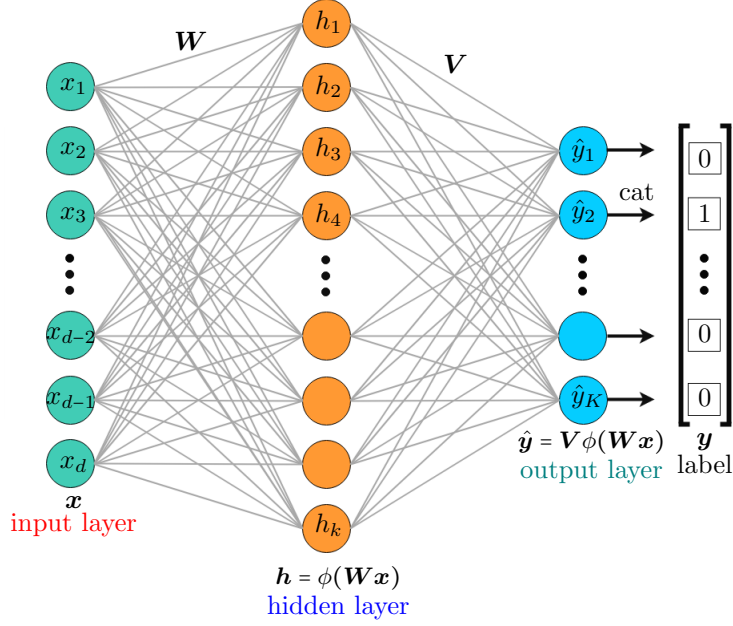


Figure 1: Illustration of a one-hidden layer neural net with  $d$  inputs,  $k$  hidden units and  $K$  outputs along with a one-hot encoded label.

To optimize this loss starting from an initialization  $\mathbf{W}_0$  we run gradient descent iterations of the form

$$\mathbf{W}_{\tau+1} = \mathbf{W}_{\tau} - \eta \nabla \mathcal{L}(\mathbf{W}_{\tau}), \quad (1.5)$$

with a step size  $\eta$ . In this paper we wish to explore the theoretical properties of the model found by such iterative updates with an emphasis on the generalization ability.

## 2 Components of a Jacobian-based theory of generalization

### 2.1 Prelude: fitting a linear model

To gain better insights into what governs the generalization capability of gradient based iterations let us consider the simple problem of fitting a linear model via gradient descent. This model maps an input/feature vector  $\mathbf{x} \in \mathbb{R}^d$  into a one-dimensional output/label via  $\mathbf{x} \mapsto f(\mathbf{x}, \mathbf{w}) := \mathbf{w}^T \mathbf{x}$ . We wish to fit a model of this form to  $n$  training data consisting of input/label pairs  $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}$ . Aggregating this training data as rows of a feature matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$  and label vector  $\mathbf{y} \in \mathbb{R}^n$ , the training problem takes the form

$$\mathcal{L}(\mathbf{w}) = \frac{1}{2} \|\mathbf{X}\mathbf{w} - \mathbf{y}\|_{\ell_2}^2. \quad (2.1)$$

We focus on an overparameterized model where there are fewer training data than the number of parameters i.e.  $n \leq d$ . We assume the feature matrix can be decomposed into the form  $\mathbf{X} = \overline{\mathbf{X}} + \mathbf{Z}$  where  $\overline{\mathbf{X}}$  is low-rank (i.e.  $\text{rank}(\overline{\mathbf{X}}) = r \ll n$ ) with singular value decomposition  $\overline{\mathbf{X}} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T$  with  $\mathbf{U} \in \mathbb{R}^{n \times r}$ ,  $\mathbf{\Sigma} \in \mathbb{R}^{r \times r}$ ,  $\mathbf{V} \in \mathbb{R}^{d \times r}$ , and  $\mathbf{Z} \in \mathbb{R}^{n \times d}$  is a matrix with i.i.d.  $\mathcal{N}(0, \sigma_x^2/n)$  entries. We shall also assume the labels are equal to  $\mathbf{y} = \overline{\mathbf{y}} + \mathbf{z}$  with  $\overline{\mathbf{y}} = \overline{\mathbf{X}}\mathbf{w}^*$  for some  $\mathbf{w}^* \in \text{Range}(\mathbf{V})$  and  $\mathbf{z} \in \mathbb{R}^n$  a Gaussian random vector with i.i.d.  $\mathcal{N}(0, \sigma_y^2/n)$  entries. One can think of this as a linear regression model where the features and labels are corrupted with Gaussian noise. The goal of course is to learn a model which fits to the clean uncorrupted data and not the corruption. In this case the population loss (i.e. test error) takes the form

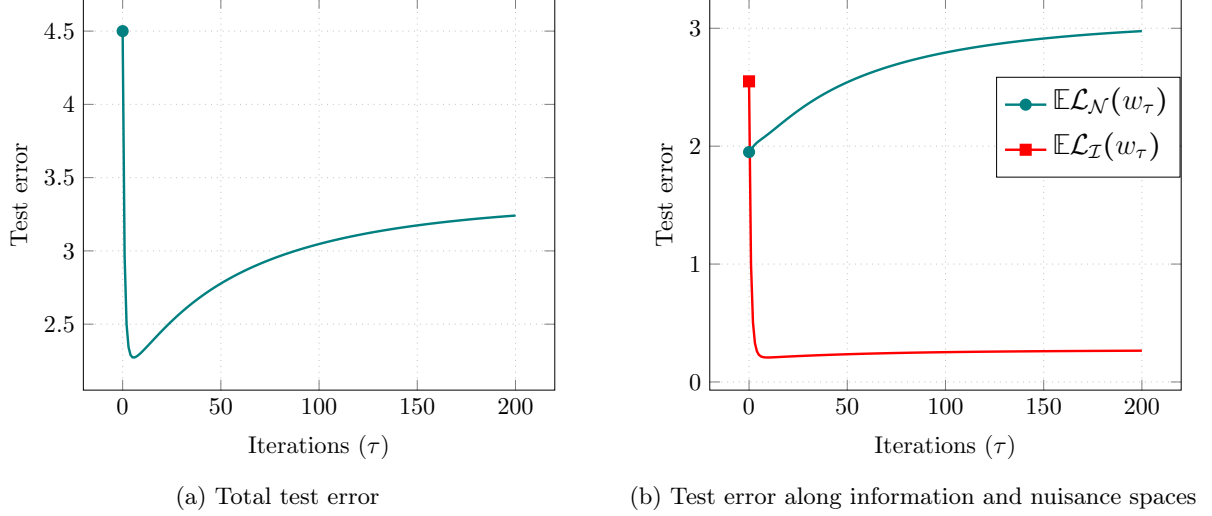


Figure 2: Plots of the (a) total test error and (b) the test error components for the model in Section 2.1. The test error decreases rapidly over the information subspace but slowly increases over the nuisance subspace.

$$\mathbb{E}[\mathcal{L}(\mathbf{w})] = \frac{1}{2} \|\bar{\mathbf{X}}\mathbf{w} - \bar{\mathbf{y}}\|_{\ell_2}^2 + \frac{1}{2}\sigma_x^2 \|\mathbf{w}\|_{\ell_2}^2 + \frac{1}{2}\sigma_y^2,$$

Now let us consider gradient descent iterations with a step size  $\eta$  which take the form

$$\mathbf{w}_{\tau+1} = \mathbf{w}_\tau - \eta \nabla \mathcal{L}(\mathbf{w}_\tau) = (\mathbf{I} - \eta \mathbf{X}^T \mathbf{X}) \mathbf{w}_\tau + \eta \mathbf{X}^T \mathbf{y}. \quad (2.2)$$

To gain further insight into the generalization capabilities of the gradient descent iterations we shall consider an instance of this problem where the subspaces  $\mathbf{U}$  and  $\mathbf{V}$  are chosen uniformly at random,  $\Sigma = \mathbf{I}_r$  with  $n = 200$ ,  $d = 500$ ,  $r = 5$ , and  $\sigma_x = 0.2$ ,  $\sigma_y = 2$ . In Figure 2a we plot the population loss evaluated at different iterations. We observe an interesting phenomenon, in the first few iterations the test error goes down quickly but it then slowly increases. To better understand this behavior we decompose the population loss into two parts by tracking the projection of the misfit  $\mathbf{X}\mathbf{w} - \mathbf{y}$  on the column space of the uncorrupted portion of the input data ( $\mathbf{U}$ ) and its complement. That is,

$$\mathbb{E} \mathcal{L}(\mathbf{w}) = \mathbb{E} \mathcal{L}_{\mathcal{I}}(\mathbf{w}) + \mathbb{E} \mathcal{L}_{\mathcal{N}}(\mathbf{w}).$$

where

$$\begin{aligned} \mathbb{E} \mathcal{L}_{\mathcal{I}}(\mathbf{w}) &:= \mathbb{E} \left[ \|\Pi_{\mathcal{I}}(\mathbf{X}\mathbf{w} - \mathbf{y})\|_{\ell_2}^2 \right] = \|\bar{\mathbf{X}}\mathbf{w} - \bar{\mathbf{y}}\|_{\ell_2}^2 + \frac{r}{2n} \sigma_x^2 \|\mathbf{w}\|_{\ell_2}^2 + \frac{r}{2n} \sigma_y^2, \\ \mathbb{E} \mathcal{L}_{\mathcal{N}}(\mathbf{w}) &:= \mathbb{E} \left[ \|\Pi_{\mathcal{N}}(\mathbf{X}\mathbf{w} - \mathbf{y})\|_{\ell_2}^2 \right] = \frac{1}{2} \left( 1 - \frac{r}{n} \right) (\sigma_x^2 \|\mathbf{w}\|_{\ell_2}^2 + \sigma_y^2), \end{aligned}$$

with  $\Pi_{\mathcal{I}} = \mathbf{U}\mathbf{U}^T$  and  $\Pi_{\mathcal{N}} = \mathbf{I} - \mathbf{U}\mathbf{U}^T$ . In Figure 2b we plot these two components. This plot clearly shows that  $\mathbb{E} \mathcal{L}_{\mathcal{I}}(\mathbf{w})$  goes down quickly while  $\mathbb{E} \mathcal{L}_{\mathcal{N}}(\mathbf{w})$  slowly increases with their sum creating the dip in the test error. Since  $\mathbf{U}$  is a basis for the range of the uncorrupted portion of the features ( $\bar{\mathbf{X}}$ ) one can think of  $\text{span}(\mathbf{U})$  as the “information” subspace and  $\mathbb{E} \mathcal{L}_{\mathcal{I}}(\mathbf{w})$  as the test error on this information subspace. Similarly, one can think of the complement of this subspace as the “nuisance” subspace and  $\mathbb{E} \mathcal{L}_{\mathcal{N}}(\mathbf{w})$  as the test error on this nuisance subspace. Therefore, one can interpret Figure 2a as the test error decreasing rapidly in the first few iterations over the information subspace but slowly increasing due to the contributions of the nuisance subspace.

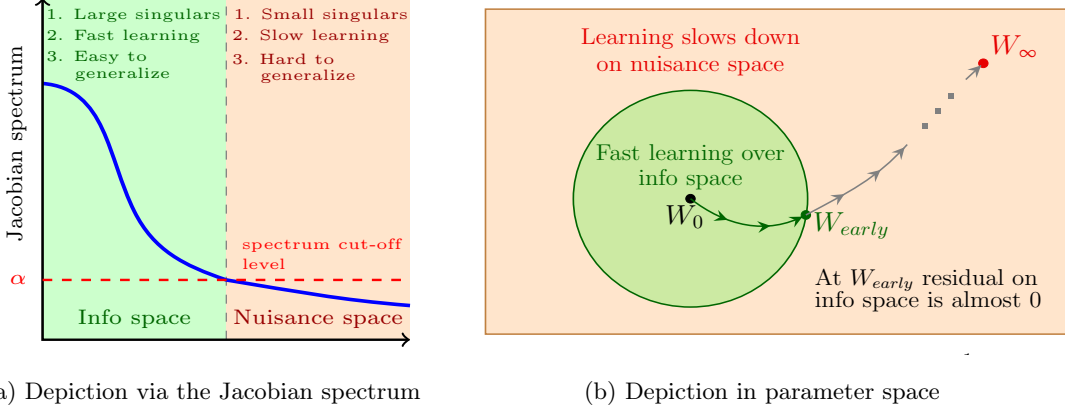


Figure 3: Depiction of the training and generalization dynamics of gradient methods based on the information and nuisance spaces associated with the neural net Jacobian.

To help demystify this behavior note that using the gradient descent updates from (2.2) the update in terms of the misfit/residual  $\mathbf{r}_\tau = \mathbf{X}\mathbf{w}_\tau - \mathbf{y}$  takes the form

$$\mathbf{r}_{\tau+1} = (\mathbf{I} - \eta \mathbf{X} \mathbf{X}^T) \mathbf{r}_\tau = (\mathbf{I} - \eta \bar{\mathbf{X}} \bar{\mathbf{X}}^T) (\bar{\mathbf{X}} \mathbf{w}_\tau - \bar{\mathbf{y}}) + \text{noise}$$

Based on the form of this update when the information subspace is closely aligned with the prominent singular vectors of  $\mathbf{X}$  the test error on the information subspace ( $\mathbb{E} \mathcal{L}_{\mathcal{I}}(\mathbf{w}) \approx \|\bar{\mathbf{X}} \mathbf{w}_\tau - \bar{\mathbf{y}}\|_{\ell_2}^2$ ) quickly decreases in the first few iterations. However, the further we iterate the parts of the residual aligned with the less prominent eigen-directions of  $\mathbf{X}$  (which correspond to the nuisance subspace) slowly pick up more energy contributing to a larger total test error.

## 2.2 Information and nuisance spaces of the Jacobian

In this section we build upon the intuition gained from the linear case to develop a better understanding of generalization dynamics for nonlinear data fitting problems. As in the linear case, in order to understand the generalization capabilities of models trained via gradient descent we need to develop better insights into the form of the gradient updates and how it affects the training dynamics. To this aim let us aggregate the weights at each iteration into one large vector  $\mathbf{w}_\tau := \text{vect}(\mathbf{W}_\tau) \in \mathbb{R}^{kd}$ , define the misfit/residual vector  $\mathbf{r}(\mathbf{w}) := f(\mathbf{w}) - \mathbf{y}$  and note that the gradient updates take the form

$$\mathbf{w}_{\tau+1} = \mathbf{w}_\tau - \eta \nabla \mathcal{L}(\mathbf{w}_\tau) \quad \text{where} \quad \nabla \mathcal{L}(\mathbf{w}) = \nabla \mathcal{L}(\mathbf{w}) = \mathcal{J}^T(\mathbf{w}) \mathbf{r}(\mathbf{w}).$$

Here,  $\mathcal{J}(\mathbf{w}) \in \mathbb{R}^{nK \times kd}$  denotes the Jacobian mapping associated with  $f$  defined as  $\mathcal{J}(\mathbf{w}) = \frac{\partial f(\mathbf{w})}{\partial \mathbf{w}}$ . Due to the form of the gradient updates the dynamics of training is dictated by the spectrum of the Jacobian matrix as well as the interaction between the residual vector and the Jacobian. If the residual vector is very well aligned with the singular vectors associated with the top singular values of  $\mathcal{J}(\mathbf{w})$ , the gradient update significantly reduces the misfit allowing substantial reduction in the train error. In a similar fashion we will also show that if the labels  $\mathbf{y}$  are well-aligned with the prominent directions of the Jacobian the test error of the trained network will be low. Thus to provide a more precise understanding of the training dynamics and generalization capabilities of neural networks it is crucial to develop a better understanding of the interaction between the Jacobian and the misfit and label vectors. To capture these interactions we require a few definitions.

**Definition 2.1 (Information & Nuisance Spaces)** Consider a matrix  $\mathbf{J} \in \mathbb{R}^{nK \times p}$  with singular value decomposition given by

$$\mathbf{J} = \sum_{s=1}^{nK} \lambda_s \mathbf{u}_s \mathbf{v}_s^T = \mathbf{U} \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_{nK}) \mathbf{V}^T,$$

with  $\lambda_1 \geq \lambda_2 \geq \dots \geq \lambda_{nK}$  denoting the singular values of  $\mathbf{J}$  in decreasing order and  $\{\mathbf{u}_s\}_{s=1}^{nK} \in \mathbb{R}^{nK}$  and  $\{\mathbf{v}_s\}_{s=1}^{nK} \in \mathbb{R}^p$  the corresponding left and right singular vectors forming the orthonormal basis matrices  $\mathbf{U} \in \mathbb{R}^{nK \times nK}$  and  $\mathbf{V} \in \mathbb{R}^{p \times nK}$ . For a spectrum cutoff  $\alpha$  obeying  $0 \leq \alpha \leq \lambda_1$  let  $r := r(\alpha)$  denote the index of the smallest singular value above the threshold  $\alpha$ .

We define the information and nuisance spaces associated with  $\mathbf{J}$  as  $\mathcal{I} := \text{span}(\{\mathbf{u}_s\}_{s=1}^r)$  and  $\mathcal{N} := \text{span}(\{\mathbf{u}_s\}_{s=r+1}^{nK})$ . We also define the truncated Jacobian

$$\mathbf{J}_{\mathcal{I}} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_r] \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r) [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_r]^T$$

which is the part of the reference Jacobian that acts on the information space  $\mathcal{I}$ .

In this paper we shall use either the expected value of the Jacobian at the random initialization or the Jacobian at one of the iterates to define the matrix  $\mathbf{J}$  and the corresponding information/nuisance spaces. More, specifically we will set  $\mathbf{J}$  to either  $\mathbf{J} = (\mathbb{E}[\mathcal{J}(\mathbf{W}_0)\mathcal{J}^T(\mathbf{W}_0)])^{1/2}$  or  $\mathbf{J} = \mathcal{J}(\mathbf{W}_\tau)$ . Therefore, one can effectively think of the information space as the span of the prominent singular vectors of the Jacobian and the nuisance space as its complement. In particular, as we demonstrate in Section 4 the Jacobian mapping associated with neural networks exhibit low-rank structure with a few large singular values and many small ones leading to natural choices for the cut-off value  $\alpha$  as well as the information and nuisance spaces. Furthermore, we demonstrate both (empirically and theoretically) that learning is fast over the information space leading to a significant reduction in both train/test accuracy in the early stages of training. However, after a certain number of iterations learning shifts to the nuisance space and reduction in the training error significantly slows down. Furthermore, subsequent iterations in this stage lead to a slight increase in test error. We provide a cartoon depiction of this behavior in Figure 3.

### 3 Main results

Our main results establish multi-class generalization bounds for neural networks trained via gradient descent. First, we will focus on networks where both layers are randomly initialized. Next we will provide guarantees for arbitrary initialization with the goal of characterizing the generalization ability of subsequent iterative updates for a given (possibly pre-trained) network in terms of its Jacobian mapping. In this paper we focus on activations  $\phi$  which are smooth and have bounded first and second order derivatives. This would for instance apply to the softplus activation  $\phi(z) = \log(1 + e^z)$ . We note that utilizing a proof technique developed in [48] for going from smooth to ReLU activations it is possible to extend our results to ReLU activations with proper modifications. We avoid doing this in the current paper for clarity of exposition. Before we begin discussing our main results we discuss some notation used throughout the paper. For a matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$  we use  $s_{\min}(\mathbf{X})$  and  $s_{\max}(\mathbf{X}) = \|\mathbf{X}\|$  to denote the minimum and maximum singular value of  $\mathbf{X}$ . For two matrices  $\mathbf{A}$  and  $\mathbf{B}$  we use  $\mathbf{A} \odot \mathbf{B}$  and  $\mathbf{A} \otimes \mathbf{B}$  to denote their Hadamard and Kronecker products, respectively. For a PSD matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  with eigenvalue decomposition  $\mathbf{A} = \sum_{i=1}^n \lambda_i \mathbf{u}_i \mathbf{u}_i^T$ , the square root matrix is defined as  $\mathbf{A}^{1/2} := \sum_{i=1}^n \sqrt{\lambda_i} \mathbf{u}_i \mathbf{u}_i^T$ . We also use  $\mathbf{A}^\dagger$  to denote the pseudo-inverse of  $\mathbf{A}$ . In this paper we mostly focus on label vectors  $\mathbf{y}$  which are one-hot encoded i.e. all entries are zero except one of them. For a subspace  $\mathcal{S} \subset \mathbb{R}^n$  and point  $\mathbf{x} \in \mathbb{R}^n$ ,  $\Pi_{\mathcal{S}}(\mathbf{x})$  denotes the projection of  $\mathbf{x}$  onto  $\mathcal{S}$ . Finally, before stating our results we need to provide a quantifiable measure of performance for a trained model. Given a sample  $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^K$  from a distribution  $\mathcal{D}$ , the classification error of the network  $\mathbf{W}$  with respect to  $\mathcal{D}$  is defined as

$$\text{Err}_{\mathcal{D}}(\mathbf{W}) = \mathbb{P} \left\{ \arg \max_{1 \leq \ell \leq K} \mathbf{y}_\ell \neq \arg \max_{1 \leq \ell \leq K} f_\ell(\mathbf{x}; \mathbf{W}) \right\}. \quad (3.1)$$

### 3.1 Results for random initialization

To explore the generalization of randomly initialized networks, we utilize the neural tangent kernel.

**Definition 3.1 (Multiclass Neural Tangent Kernel (M-NTK) [31])** Let  $\mathbf{w} \in \mathbb{R}^d$  be a vector with  $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$  distribution. Consider a set of  $n$  input data points  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$  aggregated into the rows of a data matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$ . Associated to the activation  $\phi$  and the input data matrix  $\mathbf{X}$  we define the multiclass kernel matrix as

$$\Sigma(\mathbf{X}) := \mathbf{I}_K \otimes \mathbb{E} \left[ \left( \phi'(\mathbf{X}\mathbf{w}) \phi'(\mathbf{X}\mathbf{w})^T \right) \odot (\mathbf{X}\mathbf{X}^T) \right],$$

where  $\mathbf{I}_K$  is the identity matrix of size  $K$ . Here, the  $\ell$  th diagonal block of  $\Sigma(\mathbf{X})$  corresponds to the kernel matrix associated with the  $\ell$  th network output for  $1 \leq \ell \leq K$ . This kernel is intimately related to the multiclass Jacobian mapping. In particular, suppose the initial input weights  $\mathbf{W}_0$  are distributed i.i.d.  $\mathcal{N}(0, 1)$  and the output layer  $\mathbf{V}$  has i.i.d. zero-mean entries with  $\nu^2/K$  variance. Then  $\mathbb{E}[\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T] = \nu^2 \Sigma(\mathbf{X})$ . We use the square root of this multiclass kernel matrix (i.e.  $\Sigma(\mathbf{X})^{1/2}$ ) to define the information and nuisance spaces for our random initialization result.

**Theorem 3.2** Let  $\zeta, \Gamma, \bar{\alpha}$  be scalars obeying  $\zeta \leq 1/2$ ,  $\Gamma \geq 1$ , and  $\bar{\alpha} \geq 0$  which determine the overall precision, cut-off and learning duration, respectively.<sup>2</sup> Consider a training data set  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^K$  generated i.i.d. according to a distribution  $\mathcal{D}$  where the input samples have unit Euclidean norm and the concatenated label vector obeys  $\|\mathbf{y}\|_{\ell_2} = \sqrt{n}$  (e.g. one-hot encoding). Consider a neural net with  $k$  hidden layers as described in (1.1) parameterized by  $\mathbf{W}$  where the activation function  $\phi$  obeys  $|\phi'(z)|, |\phi''(z)| \leq B$ . Let  $\mathbf{W}_0$  be the initial weight matrix with i.i.d.  $\mathcal{N}(0, 1)$  entries. Fix a precision level  $\zeta$  and set  $\nu = \zeta/(50B\sqrt{\log(2K)})$ . Also assume the output layer  $\mathbf{V}$  has i.i.d. Rademacher entries scaled by  $\frac{\nu}{\sqrt{kK}}$ . Furthermore, set  $\mathbf{J} := (\Sigma(\mathbf{X}))^{1/2}$  and define the information  $\mathcal{I}$  and nuisance  $\mathcal{N}$  spaces and the truncated Jacobian  $\mathbf{J}_{\mathcal{I}}$  associated with the Jacobian  $\mathbf{J}$  based on a cut-off spectrum value of  $\alpha_0 = \bar{\alpha} \sqrt[4]{n} \sqrt{K} \|\mathbf{X}\| B$  per Definition 2.1. Assume

$$k \gtrsim \frac{\Gamma^4 \log n}{\zeta^4 \bar{\alpha}^8} \quad (3.2)$$

with  $\Gamma \geq 1$ . We run gradient descent iterations of the form (1.5) with a learning rate  $\eta \leq \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ . Then, after  $T = \frac{\Gamma K}{\eta \nu^2 \alpha_0^2}$  iterations, classification error  $\text{Err}_{\mathcal{D}}(\mathbf{W}_T)$  is upper bounded by

$$\underbrace{\frac{2\|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2}}{\sqrt{n}}}_{\text{bias term}} + \underbrace{\frac{12B\sqrt{K}}{\sqrt{n}} \left( \|\mathbf{J}_{\mathcal{I}}^T \mathbf{y}\|_{\ell_2} + \frac{\Gamma}{\alpha_0} \|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2} \right) + 12 \left( 1 + \frac{\Gamma}{\bar{\alpha} \sqrt[4]{n} \|\mathbf{X}\|^2} \right) \zeta + 5 \sqrt{\frac{\log(2/\delta)}{n}} + 2e^{-\Gamma}}_{\text{variance term}},$$

holds with probability at least  $1 - (2K)^{-100} - \delta$ .

This theorem shows that even networks of moderate width can achieve a small generalization error if (1) the data has low-dimensional representation i.e. the kernel is approximately low-rank and (2) the inputs and labels are semantically-linked i.e. the label vector  $\mathbf{y}$  mostly lies on the information space.

• **Bias–Variance decomposition:** The generalization error has two core components: bias and variance. The bias component  $\|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2}/\sqrt{n}$  arises from the portion of the labels that falls over the nuisance space leading to a nonzero training error. The variance component is proportional to the distance  $\|\mathbf{W}_T - \mathbf{W}_0\|_F$  and arises from the growing model complexity as gradient descent strays further away from the initialization while fitting the label vector over the information space. If the label vector is aligned with the information space, bias term  $\Pi_{\mathcal{N}}(\mathbf{y})$  will be small. Additionally, if the kernel matrix is approximately low-rank, we can

<sup>2</sup>Note that this theorem and its conclusions hold for any choice of these parameters in the specified range.



set  $\bar{\alpha}$  to ensure small variance even when the width grows at most logarithmically with the size of the training data as required by (6.86). In particular, using  $\|\mathbf{J}_{\mathcal{I}}^\dagger \mathbf{y}\|_{\ell_2} \leq \|\mathbf{y}\|_{\ell_2}/\alpha_0 \leq \sqrt{n}/\alpha_0$ , the bound simplifies to

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \leq \frac{2}{\sqrt{n}} \|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2} + \frac{36\Gamma}{\bar{\alpha}^4 \sqrt{n} \|\mathbf{X}\|^2} + 12\zeta + 5\sqrt{\frac{\log(2/\delta)}{n}} + 2e^{-\Gamma}, \quad (3.3)$$

which is small as soon as the label vector is well-aligned with the information subspace. We note however that our results continue to apply even when the kernel is not approximately low-rank. In particular, consider the extreme case where we select  $\alpha_0 = \sqrt{\lambda} := \sqrt{\lambda_{\min}(\Sigma(\mathbf{X}))}$ . Then, the information space  $\mathcal{I}$  spans  $\mathbb{R}^{K^n}$  and the bias term disappears ( $\|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2} = 0$ ) and

$$\|\mathbf{J}_{\mathcal{I}}^\dagger \mathbf{y}\|_{\ell_2} = \|\mathbf{J}^\dagger \mathbf{y}\|_{\ell_2} = \sqrt{\mathbf{y}^T \Sigma(\mathbf{X})^{-1} \mathbf{y}}.$$

In this case our results guarantee that

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \lesssim \frac{\sqrt{K}}{\sqrt{n}} \sqrt{\mathbf{y}^T \Sigma^{-1}(\mathbf{X}) \mathbf{y}} + \sqrt{\frac{\log(2/\delta)}{n}}, \quad (3.4)$$

holds as long as  $\Sigma(\mathbf{X})$  is invertible and the width of the network obeys

$$k \gtrsim \frac{n^2 K^4 \|\mathbf{X}\|^4 \log n}{\lambda^4} \quad (3.5)$$

We note that in this special case our results improve upon the required width in recent literature [4]<sup>3</sup> that focuses on  $K = 1$  and a conclusion of the form (3.4). However, as we demonstrate in our numerical experiments in practice  $\lambda$  can be rather small or even zero (e.g. see the toy model in Section 3.3) so that requirements of the form (3.5) may require unrealistically (or even infinitely) wide networks. In contrast, as discussed above by harnessing the low-rank structure of the Jacobian our results show that neural networks generalize well as soon as the width grows at most logarithmically in the size of the training data (even when  $\lambda = 0$ ).

• **Small width is sufficient for generalization:** Based on our simulations the M-NTK (or more specifically Jacobian at random initialization) indeed has low-rank structure with a few large eigenvalues and many smaller ones. As a result a typical scaling of the cut-off  $\alpha_0$  is so that  $\bar{\alpha}$  scales like a constant. In that case our result states that as soon as the number of hidden nodes are moderately large (e.g. logarithmic in  $n$ ) then good generalization can be achieved. Specifically we can achieve good generalization by using width on the order of  $\log n$  and picking small values for  $\zeta$  and  $\bar{\alpha}$  and large values for  $\Gamma$ .

• **Network size–Bias tradeoff:** Based on the requirement (6.86) if the network is large (in terms of # of hidden units  $k$ ), we can choose a small cut-off  $\alpha_0$ . This in turn allows us to enlarge the information space and reduce the training bias. In summary, as network capacity grows, we can gradually interpolate finer detail and reduce bias. On the other hand, choosing a properly large  $\alpha_0$ , we can obtain good bounds for even small network sizes  $k$  as long as the portion of the labels that fall on the nuisance space is small. This is in stark contrast to related works [4, 21, 2, 15] where network size grows inversely proportional to the distance between the input samples or other notions of margin.

• **Fast convergence:** We note that by setting learning rate to  $\eta = \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ , the number of gradient iterations is upper bounded by  $\frac{\Gamma}{\bar{\alpha}^2}$ . Hence, the training speed is dictated by and is inversely proportional to the the smallest singular value over the information space. Specifically, when the Jacobian is sufficiently low-rank so that we can pick  $\bar{\alpha}$  to be a constant, convergence on the information space is rather fast requiring only a *constant number of iterations* to converge to any fixed constant accuracy. See the proofs for further detail on the optimization dynamics of the training problem (e.g. results/proofs for linear convergence of the empirical loss).

---

<sup>3</sup>Based on our understanding [4] requires the number of hidden units to be at least on the order of  $k \gtrsim \frac{n^8}{\lambda^6}$ . Note that using the fact that  $\|\mathbf{X}\| \leq \sqrt{n}$  our result reduces the dependence on width by a factor of at least  $\frac{n^4}{\lambda^2}$ . We note that  $\|\mathbf{X}\|$  often scales with  $\sqrt{\frac{n}{d}}$  so that the improvement in width is even more pronounced in typical instances.



### 3.2 Generalization guarantees with arbitrary initialization

Our next result provides generalization guarantees from an arbitrary initialization which applies to pre-trained networks (e.g. those that arise in transfer learning applications) as well as intermediate gradient iterates as the weights evolve. This result has a similar flavor to Theorem 3.2 with the key difference that the information and nuisance spaces are defined with respect to any arbitrary initial Jacobian. This shows that if a pre-trained model<sup>4</sup> provides a better low-rank representation of the data in terms of its Jacobian, it is more likely to generalize well. Furthermore, given its deterministic nature the theorem can be applied at any iteration, implying that if the Jacobians of any of the iterates provides a better low-rank representation of the data then one can provide sharper generalization guarantees.

**Theorem 3.3** *Let  $\zeta, \Gamma, \bar{\alpha}$  be scalars obeying  $\zeta \leq 1/2$ ,  $\Gamma \geq 1$ , and  $\bar{\alpha} \geq 0$  which determine the overall precision, cut-off and learning duration, respectively.<sup>5</sup> Consider a training data set  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^K$  generated i.i.d. according to a distribution  $\mathcal{D}$  where the input samples have unit Euclidean norm. Also consider a neural net with  $k$  hidden layers as described in (1.1) parameterized by  $\mathbf{W}$  where the activation function  $\phi$  obeys  $|\phi'(z)|, |\phi''(z)| \leq B$ . Let  $\mathbf{W}_0$  be the initial weight matrix with i.i.d.  $\mathcal{N}(0, 1)$  entries. Also assume the output matrix has bounded entries obeying  $\|\mathbf{V}\|_{\ell_\infty} \leq \frac{\nu}{\sqrt{kK}}$ . Furthermore, set  $\mathbf{J} := \mathcal{J}(\mathbf{W}_0)$  and define the information  $\mathcal{I}$  and nuisance  $\mathcal{N}$  subspaces and the truncated Jacobian  $\mathbf{J}_{\mathcal{I}}$  associated with the reference/initial Jacobian  $\mathbf{J}$  based on a cut-off spectrum value  $\alpha = \nu B \bar{\alpha} \sqrt[4]{n} \sqrt{\|\mathbf{X}\|}$ . Also define the initial residual  $\mathbf{r}_0 = f(\mathbf{W}_0) - \mathbf{y} \in \mathbb{R}^{nK}$  and pick  $C_r > 0$  so that  $\frac{\|\mathbf{r}_0\|_{\ell_2}}{\sqrt{n}} \leq C_r$ . Suppose number of hidden nodes  $k$  obeys*

$$k \gtrsim \frac{C_r^2 \Gamma^4}{\bar{\alpha}^8 \nu^2 \zeta^2}, \quad (3.6)$$

with  $\Gamma \geq 1$  and tolerance level  $\zeta$ . Run gradient descent updates (1.5) with learning rate  $\eta \leq \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ . Then, after  $T = \frac{\Gamma}{\eta \alpha^2}$  iterations, with probability at least  $1 - \delta$ , the generalization error obeys

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \leq \underbrace{\frac{2\|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}}{\sqrt{n}}}_{\text{bias term}} + \underbrace{\frac{12\nu B}{\sqrt{n}} \left( \|\mathbf{J}_{\mathcal{I}}^\dagger \mathbf{r}_0\|_{\ell_2} + \frac{\Gamma}{\alpha} \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} \right)}_{\text{variance term}} + 5\sqrt{\frac{\log(2/\delta)}{n}} + 2C_r(e^{-\Gamma} + \zeta).$$

As with the random initialization result, this theorem shows that as long as the initial residual is sufficiently correlated with the information space, then high accuracy can be achieved for neural networks with moderate width. As with its randomized counterpart this result also allows us to study various tradeoffs between bias-variance and network size-bias. Crucially however this result does not rely on random initialization. The reason this is particularly important is two fold. First, in many scenarios neural networks are not initialized at random. For instance, in transfer learning the network is pre-trained via data from a different domain. Second, as we demonstrate in Section 4 as the iterates progress the Jacobian mapping seems to develop more favorable properties with the labels/initial residuals becoming more correlated with the information space of the Jacobian. As mentioned earlier, due its deterministic nature the theorem above applies in both of these scenarios. In particular, if a pre-trained model provides a better low-rank representation of the data in terms of its Jacobian, it is more likely to generalize well. Furthermore, given its deterministic nature the theorem can be applied at any iteration by setting  $\boldsymbol{\theta}_0 = \boldsymbol{\theta}_\tau$ , implying that if the Jacobians of any of the iterates provides a better low-rank representation of the data then one can provide sharper generalization guarantees. Our numerical experiments demonstrate that the Jacobian of the neural network seems to adapt to the dataset over time with a more substantial amount of the labels lying on the information space. While we have not formally proven such an adaptation behavior in this paper, we hope to develop rigorous theory demonstrating this adaptation in our future work. Such a result when combined with our arbitrary initialization guarantee above can potentially provide significantly tighter generalization bounds. This is particularly important in

<sup>4</sup>e.g. obtained by training with data in a related problem as is common in transfer learning.

<sup>5</sup>Note that this theorem and its conclusions hold for any choice of these parameters in the specified range.

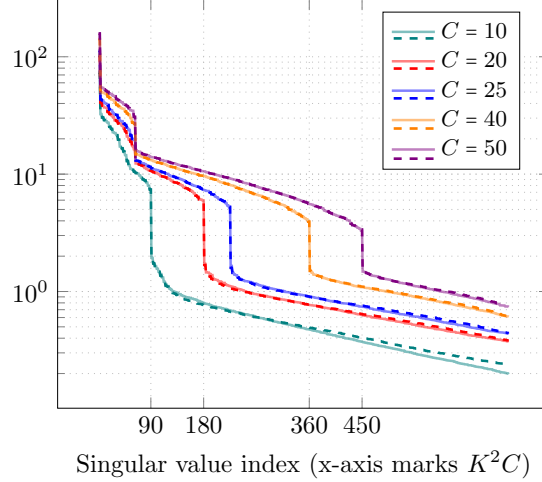


Figure 4: The singular values of the normalized Jacobian spectrum  $\sqrt{\frac{KC}{n}} \mathcal{J}(\mathbf{W}_0)$  of a one-hidden layer neural network with  $K = 3$  outputs. Here, the data set is generated according to the Gaussian mixture model in Definition 3.4 with  $K = 3$  classes and  $\sigma = 0.1$ . We pick the cluster center so that the distance between any two is at least 0.5. We consider two cases:  $n = 30C$  (solid line) and  $n = 60C$  (dashed line). These plots demonstrate that the top  $KC$  singular values grow with the square root of the size of the data set ( $\sqrt{n}$ ).

light of a few recent literature [17, 23, 56] suggesting a significant gap between generalization capabilities of kernel methods/linearized neural nets when compared with neural nets operating beyond a linear or NTK learning regime (e.g. mean field regime). As a result we view our deterministic result as a first step towards moving beyond the NTK regime.

### 3.3 Case Study: Gaussian mixture model

To illustrate a concrete example, we consider a distribution based on a Gaussian mixture model consisting of  $K$  classes where each class consists of  $C$  clusters.

**Definition 3.4 (Gaussian mixture model)** Consider a data set of size  $n$  consisting of input/label pairs  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^K$ . We assume this data set consists of  $K$  classes each comprising of  $C$  clusters with a total of  $KC$  clusters. We use the class/cluster pair to index the clusters with  $(\ell, \tilde{\ell})$  denoting the  $\tilde{\ell}$ th cluster from the  $\ell$ th class. We assume the data set in cluster  $(\ell, \tilde{\ell})$  is centered around a cluster center  $\boldsymbol{\mu}_{\ell, \tilde{\ell}} \in \mathbb{R}^d$  with unit Euclidian norm. We assume the data set is generated i.i.d. with the cluster membership assigned uniformlych of the clusters with probability  $\frac{1}{KC}$ <sup>6</sup> and the input data points associated with the cluster indexed by  $(\ell, \tilde{\ell})$  are generated i.i.d. according to  $\mathcal{N}(\boldsymbol{\mu}_{\ell, \tilde{\ell}}, \frac{\sigma^2}{d} \mathbf{I}_d)$  with the corresponding label set to the one-hot encoded vector associated with class  $\ell$  i.e.  $\mathbf{e}_\ell$ . We note that in this model the cluster indexed by  $(\ell, \tilde{\ell})$  contains  $\tilde{n}_{\ell, \tilde{\ell}}$  data points satisfying  $\mathbb{E}[\tilde{n}_{\ell, \tilde{\ell}}] = \tilde{n} = \frac{n}{KC}$ .

This distribution is an ideal candidate to demonstrate why the Jacobian of the network exhibits low-rank or bimodal structure. Let us consider the extreme case  $\sigma = 0$  where we have a discrete input distribution over the cluster centers. In this scenario, we can show that the multi-class Jacobian matrix is at most rank

$$K^2C = \# \text{ of output nodes} \times \# \text{ of distinct inputs.}$$

<sup>6</sup>This assumption is for simplicity of exposition. Our results (with proper modification) apply to any discrete probability distribution over the clusters.

as there are (i) only  $KC$  distinct input vectors and (ii)  $K$  output nodes. We can thus set the information space to be the top  $K^2C$  eigenvectors of the multiclass kernel matrix  $\Sigma(\mathbf{X})$ . As formalized in the appendix, it can be shown that

- The singular values of the information space grow proportionally with  $n/KC$ .
- The concatenated label vector  $\mathbf{y}$  perfectly lies on the information space.

In Figure 4 we numerically verify that the approximate rank and singular values of the Jacobian indeed scale as above even when  $\sigma > 0$ . The following informal theorem leverages these observations to establish a generalization bound for this mixture model. This informal statement is for exposition purposes. See Theorem A.3 in Appendix A for a more detailed result capturing the exact dependencies (e.g.  $\zeta, B, \log n$ ). In this theorem we use  $\gtrsim$  to denote inequality up to constant/logarithmic factors.

**Theorem 3.5 (Generalization for Gaussian Mixture Models-simplified)** *Consider a data set of size  $n$  consisting of input/label pairs  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^K$  generated according to a Gaussian mixture model per Definition 3.4 with the standard deviation obeying  $\sigma \lesssim \frac{K}{n}$ . Let  $\mathbf{M} = [\boldsymbol{\mu}_{1,1} \dots \boldsymbol{\mu}_{K,C}]^T$  be the matrix obtained by aggregating all the cluster centers as rows and let  $\mathbf{g} \in \mathbb{R}^d$  be a Gaussian random vector distributed as  $\mathcal{N}(0, \mathbf{I}_d)$ . Also let  $\Sigma(\mathbf{M}) \in \mathbb{R}^{KC \times KC}$  be the M-NTK associated with the cluster centers  $\mathbf{M}$  per Definition 3.1. Furthermore, set  $\lambda_{\mathbf{M}} = \lambda_{\min}(\Sigma(\mathbf{M}))$ , and assume  $\lambda_{\mathbf{M}} > 0$ . Also, assume the number of hidden nodes obeys*

$$k \gtrsim \frac{\Gamma^4 K^8 C^4}{\lambda_{\mathbf{M}}^4}.$$

*Then, after running gradient descent for  $T = \frac{2\Gamma K^2 C}{\lambda_{\mathbf{M}}}$  iterations, the model obeys*

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \lesssim \Gamma \sqrt{\frac{K^2 C}{n \lambda_{\mathbf{M}}}},$$

*with high probability.*

We note that  $\lambda_{\mathbf{M}}$  captures how diverse the cluster centers are. In this sense  $\lambda_{\mathbf{M}} > 0$  intuitively means that neural network, specifically the neural tangent kernel, is sufficiently expressive to interpolate the cluster centers. In fact when the cluster centers are in generic position  $\lambda_{\mathbf{M}}$  scales like a constant [48]. This theorem focuses on the regime where the noise level  $\sigma$  is small. In this case we show that one can achieve good generalization as soon as the number of data points scale with the square of the number classes times the total number of cluster (i.e.  $n \gtrsim K^2 C$ ) which is the effective rank of the M-NTK matrix. We note that this result follows from our main result with random initialization by setting the cutoff level at  $\alpha_0^2 \sim \frac{\lambda_{\mathbf{M}} n}{KC}$ . This demonstrates that in this model  $\bar{\alpha}$  does indeed scale as a constant. Finally, the required network width is independent of  $n$  and only depends on  $K$  and  $C$  specifically we require  $k \gtrsim K^8 C^4$ . This is in stark contrast with [4] in the binary case. To the best of understanding [4] requires  $k \gtrsim \frac{n^8}{\lambda_{\mathbf{X}}^8}$  which depends on  $n$  (in lieu of  $K$  and  $C$ ) and the minimum eigenvalue  $\lambda_{\mathbf{X}}$  of the NTK matrix  $\Sigma(\mathbf{X})$  (rather than  $\lambda_{\mathbf{M}}$ ). Furthermore, in this case as  $\sigma \rightarrow 0$ ,  $\Sigma(\mathbf{X})$  becomes rank deficient and  $\lambda_{\mathbf{X}} \rightarrow 0$  so that the required width of [4] grows to infinity.

### 3.4 Prior Art

Neural networks have impressive generalization abilities even when they are trained with more parameters than the size of the dataset [59]. Thus, optimization and generalization properties of neural networks have been the topic of many recent works [59]. Below we discuss related work on classical learning theory as well as optimization and implicit bias.

**Statistical learning theory:** Statistical properties of neural networks have been studied since 1990's [3, 8, 7]. With the success of deep networks, there is a renewed interest in understanding capacity of the neural networks under different norm constraints or network architectures [22, 5, 44, 26]. [6, 45] established

tight sample complexity results for deep networks based on the product of appropriately normalized spectral norms. See also [43] for improvements via leveraging various properties of the inter-layer Jacobian and [39] for results with convolutional networks. Related, [5] leverages compression techniques for constructing tighter bounds. [57] jointly studies statistical learning and adversarial robustness. These interesting results, provide generalization guarantees for the optimal solution to the empirical risk minimizer. In contrast, we focus on analyzing the generalization dynamics of gradient descent iterations.

**Properties of gradient descent:** There is a growing understanding that solutions found by first-order methods such as gradient descent have often favorable properties. Generalization properties of stochastic gradient descent is extensively studied empirically [34, 28, 50, 16, 29, 24, 25]. For linearly separable datasets, [55, 27, 14, 32, 33] show that first-order methods find solutions that generalize well without an explicit regularization for logistic regression. An interesting line of work establish connection between kernel methods and neural networks and study the generalization abilities of kernel methods when the model interpolates the training data [19, 10, 11, 12, 38, 13]. [18, 54, 42, 53, 49] relate the distribution of the network weights to Wasserstein gradient flows using mean field analysis. This literature is focused on asymptotic characterizations rather than finite-size networks.

**Global convergence and generalization of neural nets:** Closer to this work, recent literature [15, 4, 40, 1] provides generalization bounds for overparameterized networks trained via gradient descent. Also see [36, 30] for interesting visualization of the optimization and generalization landscape. Similar to Theorem 3.2, [4] uses the NTK to provide generalization guarantees. [37] leverages low-rank Jacobian structure to establish robustness to label noise. These works build on global convergence results of randomly initialized neural networks [21, 20, 2, 17, 60, 46, 47, 61] which study the gradient descent trajectory via comparisons to a linearized Neural Tangent Kernel (NTK) learning problem. These results however typically require unrealistically wide networks for optimization where the width grows poly-inversely proportional to the distance between the input samples. Example distance measures are class margin for logistic loss and minimum eigenvalue of the kernel matrix for least-squares. Our work circumvents this by allowing a capacity-dependent interpolation. We prove that even rather small networks (e.g. of constant width) can interpolate the data over a low-dimensional information space without making restrictive assumptions on the input. This approach also leads to faster convergence rates. In terms of generalization, our work has three distinguishing features: (a) bias-variance tradeoffs by identifying information/nuisance spaces, (b) no margin/distance/minimum eigenvalue assumptions on data, (c) the bounds apply to multiclass classification as well as pre-trained networks (Theorem 3.3).

## 4 Numerical experiments

**Experimental setup.** We present experiments supporting our theoretical findings on the CIFAR-10 dataset, which consists of  $50k$  training images and  $10k$  test images in 10 classes. For our experiments, we reduced the number of classes to 3 (automobile, airplane, bird) and subsampled the training data such that each class is represented by 3333 images (9999 in total). This is due to the fact that calculating the full spectrum of the Jacobian matrix over the entire data set is computationally intensive<sup>7</sup>. For testing, we used all examples of the 3 classes (3000 in total). In all of our experiments we set the information space to be the span of the top 50 singular vectors (out of total dimension of  $Kn \approx 30000$ ).

We demonstrate our results on ResNet20, a state-of-the-art architecture with a fairly low test error on this dataset (8.75% test error reported on 10 classes) and relatively few parameters ( $0.27M$ ). In order to be consistent with our theoretical formulation we made the following modifications to the default architecture: (1) we turned off batch normalization and (2) we did not pass the network output through a soft-max function. We trained the network using a least-squares loss with SGD with batch size 128 and standard data augmentation (e.g. random crop and flip). We set the initial learning rate to 0.01 and adjusted the learning rate schedule and number of epochs depending on the particular experiment so as to achieve a good fit to the training data quickly. The figures in this section depict the minimum error over a window consisting of

<sup>7</sup>We plan to perform more comprehensive set of experiments by calculating the Jacobian spectrum in a distributed manner.

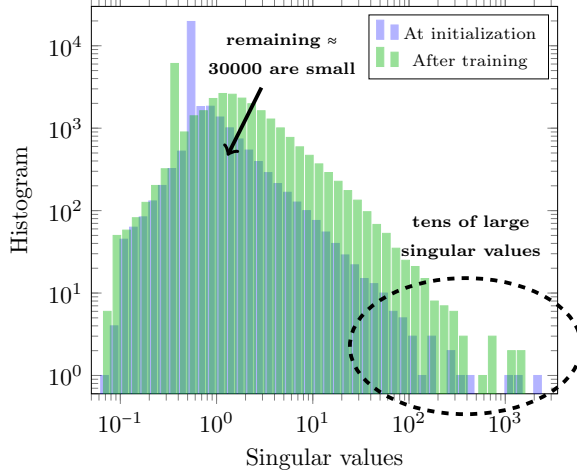


Figure 5: Histogram of the singular values of the initial and final Jacobian of the neural network during training.

	$\frac{\ \Pi_{\mathcal{I}}(\mathbf{y})\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{N}}(\mathbf{y})\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \mathbf{J}_{\mathcal{I}}^{\dagger}\mathbf{y}\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{I}}(\mathbf{r}_0)\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{N}}(\mathbf{r}_0)\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$	$\frac{\ \mathbf{J}_{\mathcal{I}}^{\dagger}\mathbf{r}_0\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$
$\mathbf{J}_{init}$	0.724	0.690	$5.44 \cdot 10^{-3}$	0.886	0.465	$4.10 \cdot 10^{-3}$
$\mathbf{J}_{final}$	0.987	0.158	$3.16 \cdot 10^{-3}$	0.976	0.217	$3.43 \cdot 10^{-3}$

Table 1: Depiction of the alignment of the initial label/residual with the information/nuisance space using uncorrupted data and a Multi-class ResNet20 model trained with SGD.

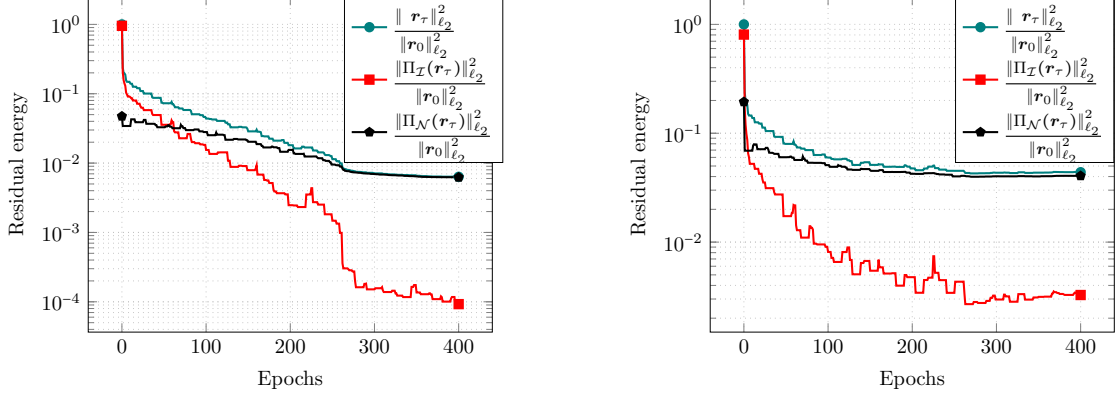
the last 10 epochs for visual clarity. We also conducted two sets of experiments to illustrate the results on uncorrupted and corrupted data.

**Experiments without label corruption.** First, we present experiments on the original training data described above with no label corruption. We train the network to fit to the training data by using 400 epochs and decreasing the learning rate at 260 and 360 epochs by a factor of 10.

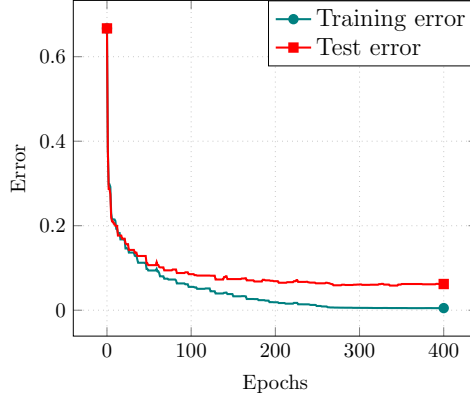
In Figure 5 we plot the histogram of the eigenvalues of the Jacobian calculated on the training data at initialization and after training. This figure clearly demonstrates that the Jacobian has low-rank structure as there are tens of large singular values with the remaining majority of the spectrum consisting of small singular values. This observation serves as a natural basis for decomposition of the label space into the information space  $\mathcal{I}$  (large singular values, low-dimensional) and nuisance space  $\mathcal{N}$  (small singular values, high-dimensional).

Our theory predicts that the sum of  $\|\mathbf{J}_{\mathcal{I}}^{\dagger}\mathbf{y}\|_{\ell_2}$  and  $\|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2}$  determines the classification error (Theorem 3.2). Table 1 collects these values for the initial and final Jacobian. These values demonstrate that the label vector is indeed correlated with the top eigenvectors of both the initial and final Jacobians. An interesting aspect of these results is that this correlation increases from the initial to the final Jacobian so that more of the label energy lies on the information space of the final Jacobian in comparison with the initial Jacobian. Stated differently, we observe a significant adaptation of the Jacobian to the labels after training compared to the initial Jacobian so that our predictions become more and more accurate as the iterates progress. In particular, Table 1 shows that more of the energy of both labels and initial residual  $\mathbf{r}_0$  lies on the information space of the Jacobian after training. Consequentially, less energy falls on the nuisance space, while  $\|\mathbf{J}_{\mathcal{I}}^{\dagger}\mathbf{y}\|_{\ell_2}$  remains relatively small resulting in better generalization.

We also track the projection of the residual  $\mathbf{r}_{\tau}$  on the information and nuisance subspaces throughout training on both training and test data and depict the results in Figures 6a and 6b. In agreement with our



(a) Residual along the information and nuisance spaces of the final Jacobian  $\mathbf{J}_{\text{final}}$  using training data. (b) Residual along the information/nuisance spaces of the final Jacobian  $\mathbf{J}_{\text{final}}$  using test data.



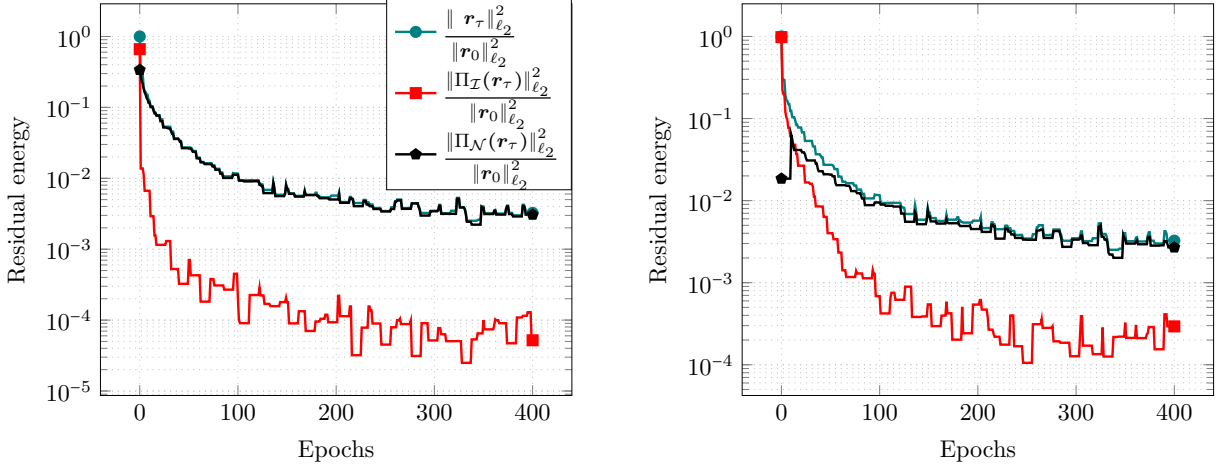
(c) Training and test miss-classification error.

Figure 6: Evolution of the residual ( $\mathbf{r}_\tau = f(\mathbf{W}_\tau) - \mathbf{y}$ ) and misclassification error on training and test data without label corruption using SGD.

theory, these plots show that learning on  $\mathcal{I}$  is fast and the residual energy decreases rapidly on this space. On the other hand, residual energy on  $\mathcal{N}$  goes down rather slowly and the decrease in total residual energy is overwhelmingly governed by  $\mathcal{I}$ , suggesting that most information relevant to learning lies in this space. We also plot the training and test error in Figure 6c. We observe that as learning progresses, the residual on both spaces decrease in tandem with training and test error.

In our final experiment with uncorrupted data we focus on training the model with an Adam optimizer with a learning rate of 0.001. We depict the results in Figure 7. We observe that due to the built-in learning rate adaptation of Adam, perfect fitting to training data is achieved in fewer iterations compared to SGD. Interestingly, the residual energy on the information space drops significantly faster than in the previous experiment with simple SGD (without Adam). In particular, after 100 epochs the fraction of the residual on the information space falls below  $4 \cdot 10^{-4}$  with Adam ( $\|\Pi_{\mathcal{I}}(\mathbf{r}_\tau)\|_{\ell_2}^2 / \|\mathbf{r}_0\|_{\ell_2}^2 \leq 4 \cdot 10^{-4}$ ) versus  $10^{-2}$  for the SGD on the final Jacobian. This suggests Adam obtains semantically relevant features significantly faster. Moreover, Table 2 shows that the Jacobian adapts to both the labels and initial residual even faster than SGD on this dataset.

**Experiments with 50% label corruption.** In our next series of experiments we study the effect of corruption. Specifically, we corrupt 50% of the labels by randomly picking a label from a (strictly) different



(a) Residual along the information and nuisance spaces of the initial Jacobian  $\mathbf{J}_{init}$  using training data. (b) Residual along the information and nuisance spaces of the final Jacobian  $\mathbf{J}_{final}$  using training data.

Figure 7: Evolution of the residual ( $\mathbf{r}_\tau = f(\mathbf{W}_\tau) - \mathbf{y}$ ) on training data without label corruption using ADAM.

	$\frac{\ \Pi_{\mathcal{I}}(\mathbf{y})\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{N}}(\mathbf{y})\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \mathbf{J}_{\mathcal{I}}^\dagger \mathbf{y}\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{I}}(\mathbf{r}_0)\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{N}}(\mathbf{r}_0)\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$	$\frac{\ \mathbf{J}_{\mathcal{I}}^\dagger \mathbf{r}_0\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$
$\mathbf{J}_{init}$	0.702	0.712	$5.36 \cdot 10^{-3}$	0.814	0.582	$4.43 \cdot 10^{-3}$
$\mathbf{J}_{final}$	0.997	0.078	$3.10 \cdot 10^{-3}$	0.991	0.136	$3.06 \cdot 10^{-3}$

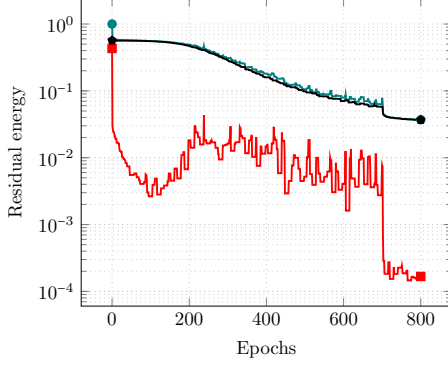
Table 2: Depiction of the alignment of the initial label/residual with the information/nuisance space using uncorrupted data and a Multi-class ResNet20 model trained with Adam.

class. We train the network for 800 epochs and divide the learning rate by 10 at 700 epochs to fit to the training data.

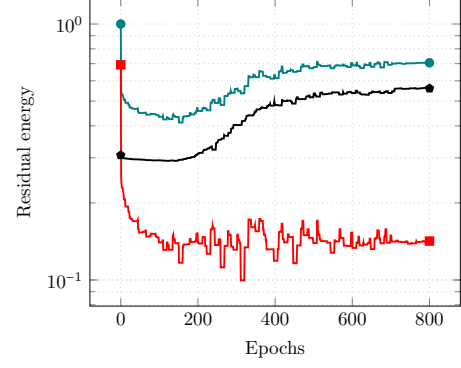
Similar to the uncorrupted case, we track the projection of the residual  $\mathbf{r}_\tau$  on the information and nuisance spaces throughout training on both training and test data and depict the results in Figures 8a and 8b. We also track the train and test misclassification error in Figure 8c. From Figure 8c it is evident that while the training error steadily decreases, test error exhibits a very different behavior from the uncorrupted experiment. In the first phase, test error drops rapidly as the network learns from information contained in the uncorrupted data, accompanied by a corresponding decrease in residual energy on the information subspace on the training data (Figure 8a). The lowest test error is observed at 100 epochs after which a steady increase follows. In the second phase, the network overfits to the corrupted data resulting in larger test error on the uncorrupted test data (Figure 8b). More importantly, the increase of the test error is due to the nuisance space as the error over information space is stable while it increases over the nuisance space. In particular the residual on  $\mathcal{N}$  slowly increases while residual on  $\mathcal{I}$  drops sharply creating a dip in both test error and total residual energy at approximately 100 epochs. This phenomenon closely resembles the population loss decomposition of the linear model discussed in Section 2.1 (see Figure 2), where we observe a dip in total test error caused by an increasing component along the nuisance space and a simultaneously decreasing component along information space.

In Table 3 we again depict the fraction of the energy of the labels and the initial residual that lies on the information/nuisance spaces. The Jacobian continues to adapt to the labels/initial residual even in the presence of label corruption, albeit to a smaller degree. We note that due to corruption, labels are less correlated with the information space of the Jacobian and the fraction of the energy on the nuisance space is

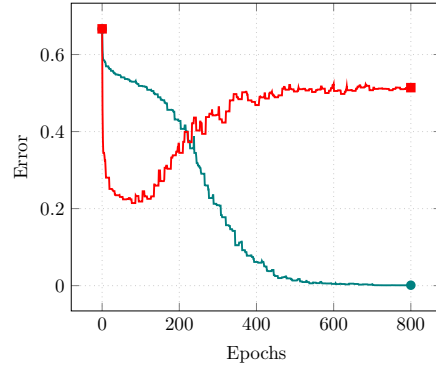




(a) Residual along the info./nuisance spaces of the Jacobian evaluated at 100 epoch ( $\mathbf{J}(\mathbf{w}_\tau)$ ) using training data.



(b) Residual along the information/nuisance spaces of the Jacobian evaluated at 100 epoch ( $\mathbf{J}(\mathbf{w}_\tau)$ ) using test data.



(c) Training and test error

Figure 8: Evolution of the residual ( $\mathbf{r}_\tau = f(\mathbf{W}_\tau) - \mathbf{y}$ ) and misclassification error on training and test data with 50% label corruption using SGD.

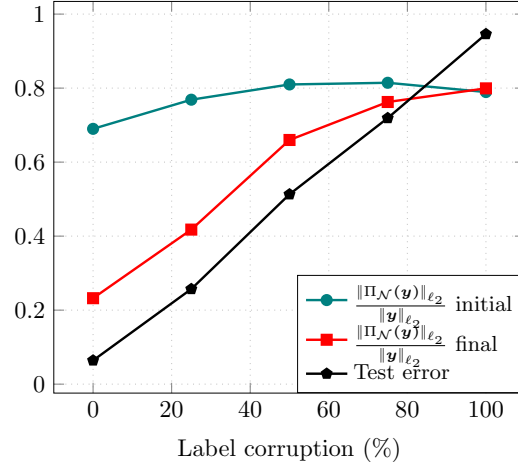


Figure 9: Fraction of the energy of the label vector that lies on the nuisance space of the initial Jacobian (cyan with circles) and final Jacobian (red with squares) as well as the test error (black with pentagons) as a function of the amount of label corruption.

	$\frac{\ \Pi_{\mathcal{I}}(\mathbf{y})\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{N}}(\mathbf{y})\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \mathbf{J}_{\mathcal{I}}^\top \mathbf{y}\ _{\ell_2}}{\ \mathbf{y}\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{I}}(\mathbf{r}_0)\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$	$\frac{\ \Pi_{\mathcal{N}}(\mathbf{r}_0)\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$	$\frac{\ \mathbf{J}_{\mathcal{I}}^\top \mathbf{r}_0\ _{\ell_2}}{\ \mathbf{r}_0\ _{\ell_2}}$
$\mathbf{J}_{init}$	0.587	0.810	$1.72 \cdot 10^{-3}$	0.643	0.766	$1.98 \cdot 10^{-3}$
$\mathbf{J}_{final}$	0.751	0.660	$1.87 \cdot 10^{-3}$	0.763	0.646	$1.20 \cdot 10^{-3}$

Table 3: Depiction of the alignment of the initial label/residual with the information/nuisance space using 50% label corrupted data and a Multi-class ResNet20 model trained with SGD.

higher which results in worse generalization (as also predicted by our theory).

In order to demonstrate the connection between generalization error and information/nuisance spaces of the Jacobian, we repeat the experiment with 25%, 75% and 100% label corruption and depict the results after 800 epochs in Figure 9. As expected, the test error increases with the corruption level. Furthermore, the corrupted labels become less correlated with the information space, with more of the label energy falling onto the nuisance space. This is consistent with our theory which predicts worse generalization in this case.

## 5 Technical approach and General Theory

In this section, we outline our approach to proving robustness of over-parameterized neural networks. Towards this goal, we consider a general formulation where we aim to fit a general nonlinear model of the form  $\mathbf{x} \mapsto f(\mathbf{x}; \boldsymbol{\theta})$  with  $\mathbf{x} \in \mathbb{R}^d$  denoting the input features,  $\boldsymbol{\theta} \in \mathbb{R}^p$  denoting the parameters, and  $f(\mathbf{x}; \boldsymbol{\theta}) \in \mathbb{R}^K$  the  $K$  outputs of the model denoted by  $f_1(\mathbf{x}; \boldsymbol{\theta}), f_2(\mathbf{x}; \boldsymbol{\theta}), \dots, f_K(\mathbf{x}; \boldsymbol{\theta})$ . For instance in the case of neural networks  $\boldsymbol{\theta}$  represents its weights. Given a data set of  $n$  input/label pairs  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}^K$ , we fit to this data by minimizing a nonlinear least-squares loss of the form

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \sum_{i=1}^n \|f(\mathbf{x}_i; \boldsymbol{\theta}) - \mathbf{y}_i\|_{\ell_2}^2. \quad (5.1)$$

To continue let us first aggregate the predictions and labels into larger vectors based on class. In particular define

$$f_\ell(\boldsymbol{\theta}) = \begin{bmatrix} f_\ell(\mathbf{x}_1; \boldsymbol{\theta}) \\ \vdots \\ f_\ell(\mathbf{x}_n; \boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^n \quad \text{and} \quad \mathbf{y}^{(\ell)} = \begin{bmatrix} (\mathbf{y}_1)_\ell \\ \vdots \\ (\mathbf{y}_n)_\ell \end{bmatrix} \in \mathbb{R}^n \quad \text{for } \ell = 1, 2, \dots, K.$$

Concatenating these vectors we arrive at

$$f(\boldsymbol{\theta}) = \begin{bmatrix} f_1(\boldsymbol{\theta}) \\ \vdots \\ f_K(\boldsymbol{\theta}) \end{bmatrix} \in \mathbb{R}^{Kn} \quad \text{and} \quad \mathbf{y} = \begin{bmatrix} \mathbf{y}^{(1)} \\ \vdots \\ \mathbf{y}^{(K)} \end{bmatrix} \in \mathbb{R}^{Kn}. \quad (5.2)$$

Using the latter we can rewrite the optimization problem (5.1) into the more compact form

$$\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \|f(\boldsymbol{\theta}) - \mathbf{y}\|_{\ell_2}^2. \quad (5.3)$$

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

$$\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_\tau) \quad \text{with} \quad \nabla \mathcal{L}(\boldsymbol{\theta}) = \mathcal{J}^T(\boldsymbol{\theta}) (f(\boldsymbol{\theta}) - \mathbf{y}). \quad (5.4)$$

As mentioned earlier due to the form of the gradient the convergence/generalization of gradient descent naturally depends on the spectral properties of the Jacobian. To capture these spectral properties we will use a reference Jacobian  $\mathbf{J}$  (formally defined below) that is close to the Jacobian at initialization  $\mathcal{J}(\boldsymbol{\theta}_0)$ .

**Definition 5.1 (Reference Jacobian and its SVD)** Consider an initial point  $\theta_0 \in \mathbb{R}^p$  and the Jacobian mapping  $\mathcal{J}(\theta_0) \in \mathbb{R}^{Kn \times p}$ . For  $\varepsilon_0, \beta > 0$ , we call  $\mathbf{J} \in \mathbb{R}^{Kn \times \max(Kn, p)}$  an  $(\varepsilon_0, \beta)$  reference Jacobian matrix if it obeys the following conditions,

$$\|\mathbf{J}\| \leq \beta, \quad \|\mathcal{J}(\theta_0)\mathcal{J}^T(\theta_0) - \mathbf{J}\mathbf{J}^T\| \leq \varepsilon_0^2, \quad \text{and} \quad \|\overline{\mathcal{J}}(\theta_0) - \mathbf{J}\| \leq \varepsilon_0.$$

where  $\overline{\mathcal{J}}(\theta_0) \in \mathbb{R}^{Kn \times \max(Kn, p)}$  is a matrix obtained by augmenting  $\mathcal{J}(\theta_0)$  with  $\max(0, Kn - p)$  zero columns. Furthermore, consider the singular value decomposition of  $\mathbf{J}$  given by

$$\mathbf{J} = \mathbf{U} \text{diag}(\boldsymbol{\lambda}) \mathbf{V}^T = \sum_{s=1}^{Kn} \lambda_s \mathbf{u}_s \mathbf{v}_s^T. \quad (5.5)$$

where  $\boldsymbol{\lambda} \in \mathbb{R}^{Kn}$  are the vector of singular values and  $\mathbf{u}_s \in \mathbb{R}^{Kn}$  and  $\mathbf{v}_s \in \mathbb{R}^p$  are the left/right singular vectors.

One natural choice for this reference Jacobian is  $\mathbf{J} = \overline{\mathcal{J}}(\theta_0)$ . However, we shall also use other reference Jacobians in our results. We will compare the gradient iterations (5.4) to the iterations associated with fitting a linearized model around  $\theta_0$  defined as  $f_{\text{lin}}(\tilde{\theta}) = f(\theta_0) + \mathbf{J}(\tilde{\theta} - \theta_0)$ , where  $\tilde{\theta}_0 \in \mathbb{R}^{\max(Kn, p)}$  is obtained from  $\theta_0$  by adding  $\max(Kn - p, 0)$  zero entries at the end of  $\theta_0$ . The optimization problem for fitting the linearized problem has the form

$$\mathcal{L}_{\text{lin}}(\theta) = \frac{1}{2} \|f_{\text{lin}}(\theta) - \mathbf{y}\|_{\ell_2}^2. \quad (5.6)$$

Thus starting from  $\tilde{\theta}_0 = \bar{\theta}_0$  the iterates  $\tilde{\theta}_\tau$  on the linearized problem take the form

$$\begin{aligned} \tilde{\theta}_{\tau+1} &= \tilde{\theta}_\tau - \eta \nabla \mathcal{L}_{\text{lin}}(\tilde{\theta}_\tau), \\ &= \tilde{\theta}_\tau - \eta \mathbf{J}^T (f(\theta_0) + \mathbf{J}(\tilde{\theta}_\tau - \theta_0) - \mathbf{y}), \\ &= \tilde{\theta}_\tau - \eta \mathbf{J}^T \mathbf{J}(\tilde{\theta}_\tau - \bar{\theta}_0) - \eta \mathbf{J}^T (f(\theta_0) - \mathbf{y}). \end{aligned} \quad (5.7)$$

The iterates based on the linearized problem will provide a useful reference to keep track of the evolution of the original iterates (5.4). Specifically we study the evolution of misfit/residuals associated with the two problems

$$\text{Original residual: } \mathbf{r}_\tau = f(\theta_\tau) - \mathbf{y}. \quad (5.8)$$

$$\text{Linearized residual: } \tilde{\mathbf{r}}_\tau = f_{\text{lin}}(\tilde{\theta}_\tau) - \mathbf{y} = (\mathbf{I} - \eta \mathbf{J} \mathbf{J}^T)^\tau \mathbf{r}_0. \quad (5.9)$$

To better understand the dynamics of convergence of the linearized iterates next we define two subspaces associated with the reference Jacobian and its spectrum.

**Definition 5.2 (Information/Nuisance Subspaces)** Let  $\mathbf{J}$  denote the reference Jacobian per Definition 5.1 with eigenvalue decomposition  $\mathbf{J} = \mathbf{U} \text{diag}(\boldsymbol{\lambda}) \mathbf{V}^T$  per (5.5). For a spectrum cutoff  $\alpha$  obeying  $0 \leq \alpha \leq \lambda_1$  let  $r(\alpha)$  denote the index of the smallest singular value above the threshold  $\alpha$ , that is,

$$r(\alpha) = \min(\{s \in \{1, 2, \dots, nK\} \text{ such that } \lambda_s \geq \alpha\}).$$

We define the information and nuisance subspaces associated with  $\mathbf{J}$  as  $\mathcal{I} := \text{span}(\{\mathbf{u}_s\}_{s=1}^r)$  and  $\mathcal{N} := \text{span}(\{\mathbf{u}_s\}_{s=r+1}^{Kn})$ . We also define the truncated reference Jacobian

$$\mathbf{J}_{\mathcal{I}} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_r] \text{diag}(\lambda_1, \lambda_2, \dots, \lambda_r) [\mathbf{v}_1 \quad \mathbf{v}_2 \quad \dots \quad \mathbf{v}_r]^T$$

which is the part of the reference Jacobian that acts on the information subspace  $\mathcal{I}$ .

We will show rigorously that the information and nuisance subspaces associated with the reference Jacobian dictate the directions where learning is fast and generalizable versus the directions where learning is slow and overfitting occurs. Before we make this precise we list two assumptions that will be utilized in our result.

**Assumption 1 (Bounded spectrum)** For any  $\boldsymbol{\theta} \in \mathbb{R}^p$  the Jacobian mapping associated with the nonlinearity  $f : \mathbb{R}^p \mapsto \mathbb{R}^n$  has bounded spectrum, i.e.  $\|\mathcal{J}(\boldsymbol{\theta})\| \leq \beta$ .

**Assumption 2 (Bounded perturbation)** Consider a point  $\boldsymbol{\theta}_0 \in \mathbb{R}^p$  and positive scalars  $\varepsilon, R > 0$ . Assume that for any  $\boldsymbol{\theta}$  obeying  $\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \leq R$ , we have

$$\|\mathcal{J}(\boldsymbol{\theta}) - \mathcal{J}(\boldsymbol{\theta}_0)\| \leq \frac{\varepsilon}{2}.$$

With these assumptions in place we are now ready to discuss our meta theorem that demonstrates that the misfit/residuals associated to the original and linearized iterates do in fact track each other rather closely.

**Theorem 5.3 (Meta Theorem)** Consider a nonlinear least squares problem of the form  $\mathcal{L}(\boldsymbol{\theta}) = \frac{1}{2} \|f(\boldsymbol{\theta}) - \mathbf{y}\|_{\ell_2}^2$  with  $f : \mathbb{R}^p \mapsto \mathbb{R}^{nK}$  the multi-class nonlinear mapping,  $\boldsymbol{\theta} \in \mathbb{R}^p$  the parameters of the model, and  $\mathbf{y} \in \mathbb{R}^{nK}$  the concatenated labels as in (5.2). Let  $\tilde{\boldsymbol{\theta}}$  be zero-padding of  $\boldsymbol{\theta}$  till size  $\max(Kn, p)$ . Also, consider a point  $\boldsymbol{\theta}_0 \in \mathbb{R}^p$  with  $\mathbf{J}$  an  $(\varepsilon_0, \beta)$  reference Jacobian associated with  $\mathcal{J}(\boldsymbol{\theta}_0)$  per Definition 5.1 and fitting the linearized problem  $f_{lin}(\tilde{\boldsymbol{\theta}}) = f(\boldsymbol{\theta}_0) + \mathbf{J}(\tilde{\boldsymbol{\theta}} - \tilde{\boldsymbol{\theta}}_0)$  via the loss  $\mathcal{L}_{lin}(\boldsymbol{\theta}) = \frac{1}{2} \|f_{lin}(\boldsymbol{\theta}) - \mathbf{y}\|_{\ell_2}^2$ . Furthermore, define the information  $\mathcal{I}$  and nuisance  $\mathcal{N}$  subspaces and the truncated Jacobian  $\mathbf{J}_{\mathcal{I}}$  associated with the reference Jacobian  $\mathbf{J}$  based on a cut-off spectrum value of  $\alpha$  per Definition 5.2. Furthermore, assume the Jacobian mapping  $\mathcal{J}(\boldsymbol{\theta}) \in \mathbb{R}^{nK \times p}$  associated with  $f$  obeys Assumptions 1 and 2 for all  $\boldsymbol{\theta} \in \mathbb{R}^p$  obeying

$$\|\boldsymbol{\theta} - \boldsymbol{\theta}_0\|_{\ell_2} \leq R := 2 \left( \|\mathbf{J}_{\mathcal{I}}^\dagger \mathbf{r}_0\|_{\ell_2} + \frac{\Gamma}{\alpha} \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \delta \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \right), \quad (5.10)$$

around a point  $\boldsymbol{\theta}_0 \in \mathbb{R}^p$  for a tolerance level  $\delta$  obeying  $0 < \delta \leq 1$  and stopping time  $\Gamma$  obeying  $\Gamma \geq 1$ . Finally, assume the following inequalities hold

$$\varepsilon_0 \leq \frac{\min(\delta\alpha, \sqrt{\delta\alpha^3/\Gamma\beta})}{5} \quad \text{and} \quad \varepsilon \leq \frac{\delta\alpha^3}{5\Gamma\beta^2}. \quad (5.11)$$

We run gradient descent iterations of the form  $\boldsymbol{\theta}_{\tau+1} = \boldsymbol{\theta}_\tau - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_\tau)$  and  $\tilde{\boldsymbol{\theta}}_{\tau+1} = \tilde{\boldsymbol{\theta}}_\tau - \eta \nabla \mathcal{L}_{lin}(\tilde{\boldsymbol{\theta}}_\tau)$  on the original and linearized problems starting from  $\boldsymbol{\theta}_0$  with step size  $\eta$  obeying  $\eta \leq 1/\beta^2$ . Then for all iterates  $\tau$  obeying  $0 \leq \tau \leq T := \frac{\Gamma}{\eta\alpha^2}$  the iterates of the original ( $\boldsymbol{\theta}_\tau$ ) and linearized ( $\tilde{\boldsymbol{\theta}}_\tau$ ) problems and the corresponding residuals  $\mathbf{r}_\tau := f(\boldsymbol{\theta}_\tau) - \mathbf{y}$  and  $\tilde{\mathbf{r}}_\tau := f_{lin}(\tilde{\boldsymbol{\theta}}_\tau) - \mathbf{y}$  closely track each other. That is,

$$\|\mathbf{r}_\tau - \tilde{\mathbf{r}}_\tau\|_{\ell_2} \leq \frac{3}{5} \frac{\delta\alpha}{\beta} \|\mathbf{r}_0\|_{\ell_2} \quad \text{and} \quad \|\tilde{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_\tau\|_{\ell_2} \leq \delta \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \quad (5.12)$$

Furthermore, for all iterates  $\tau$  obeying  $0 \leq \tau \leq T := \frac{\Gamma}{\eta\alpha^2}$

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} \leq \frac{R}{2} = \|\mathbf{J}_{\mathcal{I}}^\dagger \mathbf{r}_0\|_{\ell_2} + \frac{\Gamma}{\alpha} \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \delta \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2}. \quad (5.13)$$

and after  $\tau = T$  iteration we have

$$\|\mathbf{r}_T\|_{\ell_2} \leq e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \frac{\delta\alpha}{\beta} \|\mathbf{r}_0\|_{\ell_2}. \quad (5.14)$$

## 6 Proofs

Before we proceed with the proof let us briefly discuss some notation used throughout. For a matrix  $\mathbf{W} \in \mathbb{R}^{k \times d}$  we use  $\text{vect}(\mathbf{W}) \in \mathbb{R}^{kd}$  to denote a vector obtained by concatenating the rows  $\mathbf{w}_1, \mathbf{w}_2, \dots, \mathbf{w}_k \in \mathbb{R}^d$  of  $\mathbf{W}$ . That is,  $\text{vect}(\mathbf{W}) = [\mathbf{w}_1^T \ \mathbf{w}_2^T \ \dots \ \mathbf{w}_k^T]^T$ . Similarly, we use  $\text{mat}(\mathbf{w}) \in \mathbb{R}^{k \times d}$  to denote a  $k \times d$  matrix obtained by reshaping the vector  $\mathbf{w} \in \mathbb{R}^{kd}$  across its rows. Throughout, for a differentiable function  $\phi : \mathbb{R} \mapsto \mathbb{R}$  we use  $\phi'$  and  $\phi''$  to denote the first and second derivative.

## 6.1 Proofs for General Theory (Proof of Theorem 5.3)

In this section we prove our result for general nonlinearities. We begin with a few notations and definitions and preliminary lemmas in Section 6.1.1. Next in Section 6.1.2 we prove some key lemmas regarding the evolution of the linearized residuals  $\tilde{\mathbf{r}}_\tau$ . In Section 6.3 we establish some key Rademacher complexity results used in our generalization bounds. Finally, in Section 6.1.3 we use these results to complete the proof of Theorem 5.3.

### 6.1.1 Preliminary definitions and lemmas

Throughout we use

$$\mathbf{U}_{\mathcal{I}} = [\mathbf{u}_1 \quad \mathbf{u}_2 \quad \dots \quad \mathbf{u}_r] \in \mathbb{R}^{nK \times r} \quad \text{and} \quad \mathbf{U}_{\mathcal{N}} = [\mathbf{u}_{r+1} \quad \mathbf{u}_{r+2} \quad \dots \quad \mathbf{u}_{nK}] \in \mathbb{R}^{nK \times (nK-r)}.$$

to denote the basis matrices for the information and nuisance subspaces from Definition 5.2. Similarly, we define the information and nuisance spectrum as

$$\boldsymbol{\lambda}_{\mathcal{I}} = [\lambda_1 \quad \lambda_2 \quad \dots \quad \lambda_r]^T \quad \text{and} \quad \boldsymbol{\lambda}_{\mathcal{N}} = [\lambda_{r+1} \quad \lambda_{r+2} \quad \dots \quad \lambda_{nK}]^T.$$

We also define the diagonal matrices

$$\boldsymbol{\Lambda} = \text{diag}(\boldsymbol{\lambda}), \quad \boldsymbol{\Lambda}_{\mathcal{I}} = \text{diag}(\boldsymbol{\lambda}_{\mathcal{I}}), \quad \text{and} \quad \boldsymbol{\Lambda}_{\mathcal{N}} = \text{diag}(\boldsymbol{\lambda}_{\mathcal{N}}).$$

**Definition 6.1 (early stopping value and distance)** Consider Definition 5.2 and let  $\Gamma > 0$  be a positive scalar. Associated with the initial residual  $\mathbf{r}_0 = f(\boldsymbol{\theta}_0) - \mathbf{y}$  and the information/nuisance subspaces of the reference Jacobian  $\mathbf{J}$  (with a cut-off level  $\alpha$ ) we define the  $(\alpha, \Gamma)$  early stopping value as

$$\mathcal{B}_{\alpha, \Gamma} = \left( \sum_{s=1}^r \frac{\alpha^2}{\lambda_s^2} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 + \Gamma^2 \sum_{s=r+1}^{nK} \frac{\lambda_s^2}{\alpha^2} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \right)^{1/2}. \quad (6.1)$$

We also define the early stopping distance as

$$\mathcal{D}_{\alpha, \Gamma} = \frac{\mathcal{B}_{\alpha, \Gamma}}{\alpha}.$$

The goal of early stopping value/distance is understanding the behavior of the algorithm at a particular stopping time that depends on  $\Gamma$  and the spectrum cutoff  $\alpha$ . In particular, as we will see later on the early stopping distance characterizes the distance from initialization at an appropriate early stopping time. We continue by stating and proving a few simple lemmas. The first Lemma provides upper/lower bounds on the early stopping value.

**Lemma 6.2 (Bounds on Early-Stopping Value)** The early stopping value  $\mathcal{B}_{\alpha, \Gamma}$  from Definition 6.1 obeys

$$\mathcal{B}_{\alpha, \Gamma} \leq \left( \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2}^2 + \Gamma^2 \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}^2 \right)^{1/2} \leq \Gamma \|\mathbf{r}_0\|_{\ell_2} \quad (6.2)$$

$$\mathcal{B}_{\alpha, \Gamma} \geq \frac{\alpha}{\lambda_1} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2}. \quad (6.3)$$

**Proof** To prove the upper bound we use the fact that  $\alpha \leq \lambda_s$  for  $s \leq r$  and  $\alpha \geq \lambda_s$  for  $s \geq r$  to conclude that

$$\begin{aligned} \mathcal{B}_{\alpha, \Gamma} &\leq \left( \sum_{s=1}^r (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 + \Gamma^2 \sum_{s=r+1}^{nK} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \right)^{1/2} \\ &= \left( \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2}^2 + \Gamma^2 \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}^2 \right)^{1/2} \\ &\leq \Gamma \|\mathbf{r}_0\|_{\ell_2}. \end{aligned}$$

To prove the lower bound, we use the facts that  $\alpha^2/\lambda_s^2 \geq \alpha^2/\lambda_1^2$  to conclude that

$$\begin{aligned}\mathcal{B}_{\alpha,\Gamma} &= \left( \sum_{s=1}^r \frac{\alpha^2}{\lambda_s^2} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 + \Gamma^2 \sum_{s=r+1}^{nK} \frac{\lambda_s^2}{\alpha^2} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \right)^{1/2}, \\ &\geq \left( \sum_{s=1}^r \frac{\alpha^2}{\lambda_s^2} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \right)^{1/2}, \\ &\geq \frac{\alpha}{\lambda_1} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2}.\end{aligned}$$

■

It is of course well known that the mapping  $(\mathbf{I} - \eta \mathbf{A} \mathbf{A}^T)$  is a contraction for sufficiently small values of  $\eta$ . The next lemma shows that if we replace one of the  $\mathbf{A}$  matrices with a matrix  $\mathbf{B}$  which is close to  $\mathbf{A}$  the resulting matrix  $(\mathbf{I} - \eta \mathbf{A} \mathbf{B}^T)$ , while may not be contractive, is not too expansive.

**Lemma 6.3 (Asymmetric PSD increase)** *Let  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times p}$  be matrices obeying*

$$\|\mathbf{A}\| \leq \beta, \quad \|\mathbf{B}\| \leq \beta, \quad \text{and} \quad \|\mathbf{B} - \mathbf{A}\| \leq \varepsilon.$$

*Then, for all  $\mathbf{r} \in \mathbb{R}^n$  and  $\eta \leq 1/\beta^2$  we have*

$$\|(\mathbf{I} - \eta \mathbf{A} \mathbf{B}^T) \mathbf{r}\|_{\ell_2} \leq (1 + \eta \varepsilon^2) \|\mathbf{r}\|_{\ell_2}.$$

**Proof** Note that using  $\eta \leq 1/\beta^2$  and  $\|\mathbf{B} - \mathbf{A}\| \leq \varepsilon$  we conclude that

$$\begin{aligned}\|(\mathbf{I} - \eta \mathbf{A} \mathbf{B}^T) \mathbf{r}\|_{\ell_2}^2 &= \|(\mathbf{I} - \eta \mathbf{B} \mathbf{B}^T - \eta(\mathbf{A} - \mathbf{B}) \mathbf{B}^T) \mathbf{r}\|_{\ell_2}^2 \\ &= \|\mathbf{r} - \eta(\mathbf{A} - \mathbf{B} + \mathbf{B}) \mathbf{B}^T \mathbf{r}\|_{\ell_2}^2 \\ &= \|\mathbf{r}\|_{\ell_2}^2 - 2\eta \mathbf{r}^T (\mathbf{A} - \mathbf{B} + \mathbf{B}) \mathbf{B}^T \mathbf{r} + \eta^2 \|\mathbf{A} \mathbf{B}^T \mathbf{r}\|_{\ell_2}^2 \\ &\leq \|\mathbf{r}\|_{\ell_2}^2 - 2\eta \|\mathbf{B}^T \mathbf{r}\|_{\ell_2}^2 + 2\eta \|(\mathbf{A} - \mathbf{B})^T \mathbf{r}\|_{\ell_2} \|\mathbf{B}^T \mathbf{r}\|_{\ell_2} + \eta^2 \|\mathbf{A}\|^2 \|\mathbf{B}^T \mathbf{r}\|_{\ell_2}^2 \\ &= \|\mathbf{r}\|_{\ell_2}^2 - \eta \|\mathbf{B}^T \mathbf{r}\|_{\ell_2}^2 + 2\eta \|(\mathbf{A} - \mathbf{B})^T \mathbf{r}\|_{\ell_2} \|\mathbf{B}^T \mathbf{r}\|_{\ell_2} + (\eta^2 \|\mathbf{A}\|^2 \|\mathbf{B}^T \mathbf{r}\|_{\ell_2}^2 - \eta \|\mathbf{B}^T \mathbf{r}\|_{\ell_2}^2) \\ &\stackrel{\eta \leq 1/\beta^2}{\leq} \|\mathbf{r}\|_{\ell_2}^2 - \eta \|\mathbf{B}^T \mathbf{r}\|_{\ell_2}^2 + 2\eta \|(\mathbf{A} - \mathbf{B})^T \mathbf{r}\|_{\ell_2} \|\mathbf{B}^T \mathbf{r}\|_{\ell_2} \\ &\stackrel{\|\mathbf{A} - \mathbf{B}\| \leq \varepsilon}{\leq} \|\mathbf{r}\|_{\ell_2}^2 - \eta \|\mathbf{B}^T \mathbf{r}\|_{\ell_2}^2 + 2\eta \varepsilon \|\mathbf{B}^T \mathbf{r}\|_{\ell_2} \|\mathbf{r}\|_{\ell_2} \\ &= (1 + \eta \varepsilon^2) \|\mathbf{r}\|_{\ell_2}^2 - \eta \left( \varepsilon \|\mathbf{r}\|_{\ell_2} - \|\mathbf{B}^T \mathbf{r}\|_{\ell_2} \right)^2 \\ &\leq (1 + \eta \varepsilon^2) \|\mathbf{r}\|_{\ell_2}^2,\end{aligned}$$

completing the proof. ■

The next lemma shows that if two PSD matrices are close to each other then an appropriate square root of these matrices will also be close.

**Lemma 6.4** *Let  $\mathbf{A}$  and  $\mathbf{B}$  be  $n \times n$  positive semi-definite matrices satisfying  $\|\mathbf{A} - \mathbf{B}\| \leq \alpha^2$  for a scalar  $\alpha \geq 0$ . Then for any  $\mathbf{X} \in \mathbb{R}^{n \times p}$  with  $p \geq n$  obeying  $\mathbf{A} = \mathbf{X} \mathbf{X}^T$ , there exists a matrix  $\mathbf{Y} \in \mathbb{R}^{n \times p}$  obeying  $\mathbf{B} = \mathbf{Y} \mathbf{Y}^T$  such that*

$$\|\mathbf{Y} - \mathbf{X}\| \leq 2\alpha$$

**Proof** First we note that for any two PSD matrices  $\mathbf{A}_+, \mathbf{B}_+ \in \mathbb{R}^{n \times n}$  obeying  $\mathbf{A}_+, \mathbf{B}_+ \geq \frac{\alpha^2}{4} \mathbf{I}_n$ , Lemma 2.2 of [51] guarantees that

$$\|\mathbf{A}_+^{1/2} - \mathbf{B}_+^{1/2}\| \leq \frac{\|\mathbf{A}_+ - \mathbf{B}_+\|}{\alpha}.$$

In the above for a PSD matrix  $\mathbf{A} \in \mathbb{R}^{n \times n}$  with an eigenvalue decomposition  $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T$  we use  $\mathbf{A}^{1/2}$  to denote the square root of the matrix given by  $\mathbf{A} = \mathbf{U}\mathbf{\Lambda}^{1/2}\mathbf{U}^T$ . We shall use this result with  $\mathbf{A}_+ = \mathbf{A} + \frac{\alpha^2}{4}\mathbf{I}_n$  and  $\mathbf{B}_+ = \mathbf{B} + \frac{\alpha^2}{4}\mathbf{I}_n$  to conclude that

$$\|\mathbf{A}_+^{1/2} - \mathbf{B}_+^{1/2}\| \leq \frac{\|\mathbf{A}_+ - \mathbf{B}_+\|}{\alpha} = \frac{\|\mathbf{A} - \mathbf{B}\|}{\alpha}.$$

Furthermore, using the fact that the eigenvalues of  $\mathbf{A}_+$  and  $\mathbf{B}_+$  are just shifted versions of the eigenvalues of  $\mathbf{A}$  and  $\mathbf{B}$  by  $\alpha^2/4$  we can conclude that

$$\|\mathbf{A}_+^{1/2} - \mathbf{A}^{1/2}\| \leq \frac{\alpha}{2} \quad \text{and} \quad \|\mathbf{B}_+^{1/2} - \mathbf{B}^{1/2}\| \leq \frac{\alpha}{2}.$$

Combining the latter two inequalities with the assumption that  $\|\mathbf{A} - \mathbf{B}\| \leq \alpha^2$  we conclude that

$$\begin{aligned} \|\mathbf{A}^{1/2} - \mathbf{B}^{1/2}\| &\leq \|\mathbf{A}_+^{1/2} - \mathbf{B}_+^{1/2}\| + \|\mathbf{A}_+^{1/2} - \mathbf{A}^{1/2}\| + \|\mathbf{B}_+^{1/2} - \mathbf{B}^{1/2}\| \\ &\leq \frac{\|\mathbf{A} - \mathbf{B}\|}{\alpha} + \frac{\alpha}{2} + \frac{\alpha}{2} \\ &\leq 2\alpha. \end{aligned} \tag{6.4}$$

Suppose  $p \geq n$  and assume the matrices  $\mathbf{A}$  and  $\mathbf{B}$  have eigenvalue decompositions given by  $\mathbf{A} = \mathbf{U}_\mathbf{A}\mathbf{\Lambda}_\mathbf{A}\mathbf{U}_\mathbf{A}^T$  and  $\mathbf{B} = \mathbf{U}_\mathbf{B}\mathbf{\Lambda}_\mathbf{B}\mathbf{U}_\mathbf{B}^T$ . Then, any  $\mathbf{X} \in \mathbb{R}^{n \times p}$  with  $p \geq n$  has the form  $\mathbf{X} = \mathbf{U}_\mathbf{A}\mathbf{\Lambda}_\mathbf{A}^{1/2}\mathbf{V}_\mathbf{A}^T$  with  $\mathbf{V}_\mathbf{A} \in \mathbb{R}^{p \times n}$  an orthonormal matrix. Now pick

$$\mathbf{Y} = \mathbf{U}_\mathbf{B}\mathbf{\Lambda}_\mathbf{B}^{1/2}\mathbf{U}_\mathbf{B}^T\mathbf{U}_\mathbf{A}\mathbf{V}_\mathbf{A}^T.$$

Then clearly  $\mathbf{Y}\mathbf{Y}^T = \mathbf{B}$ . Furthermore, we have

$$\begin{aligned} \|\mathbf{X} - \mathbf{Y}\| &= \|\mathbf{U}_\mathbf{A}\mathbf{\Lambda}_\mathbf{A}^{1/2}\mathbf{V}_\mathbf{A}^T - \mathbf{U}_\mathbf{B}\mathbf{\Lambda}_\mathbf{B}^{1/2}\mathbf{U}_\mathbf{B}^T\mathbf{U}_\mathbf{A}\mathbf{V}_\mathbf{A}^T\| \\ &= \|\mathbf{U}_\mathbf{A}\mathbf{\Lambda}_\mathbf{A}^{1/2}\mathbf{U}_\mathbf{A}^T\mathbf{U}_\mathbf{A}\mathbf{V}_\mathbf{A}^T - \mathbf{U}_\mathbf{B}\mathbf{\Lambda}_\mathbf{B}^{1/2}\mathbf{U}_\mathbf{B}^T\mathbf{U}_\mathbf{A}\mathbf{V}_\mathbf{A}^T\| \\ &= \|\left(\mathbf{U}_\mathbf{A}\mathbf{\Lambda}_\mathbf{A}^{1/2}\mathbf{U}_\mathbf{A}^T - \mathbf{U}_\mathbf{B}\mathbf{\Lambda}_\mathbf{B}^{1/2}\mathbf{U}_\mathbf{B}^T\right)\mathbf{U}_\mathbf{A}\mathbf{V}_\mathbf{A}^T\| \\ &= \|\left(\mathbf{A}^{1/2} - \mathbf{B}^{1/2}\right)\mathbf{U}_\mathbf{A}\mathbf{V}_\mathbf{A}^T\| \\ &= \|\mathbf{A}^{1/2} - \mathbf{B}^{1/2}\|. \end{aligned}$$

Combining the latter with (6.4) completes the proof. ■

### 6.1.2 Key lemmas for general nonlinearities

Throughout this section we assume  $\mathbf{J}$  is the reference Jacobian per Definition 5.1 with eigenvalue decomposition  $\mathbf{J} = \mathbf{U}\mathbf{\Lambda}\mathbf{V}^T = \sum_{s=1}^{Kn} \lambda_s \mathbf{u}_s \mathbf{v}_s^T$  with  $\mathbf{\Lambda} = \text{diag}(\boldsymbol{\lambda})$ . We also define  $\mathbf{a} = \mathbf{U}^T \mathbf{r}_0 = \mathbf{U}^T \tilde{\mathbf{r}}_0 \in \mathbb{R}^{nK}$  be the coefficients of the initial residual in the span of the column space of this reference Jacobian.

We shall first characterize the evolution of the linearized parameter  $\tilde{\boldsymbol{\theta}}_\tau$  and residual  $\tilde{\mathbf{r}}_\tau$  vectors from (5.9) in the following lemma.

**Lemma 6.5** *The linearized residual vector  $\tilde{\mathbf{r}}_\tau$  can be written in the form*

$$\tilde{\mathbf{r}}_\tau = \mathbf{U} \left( \mathbf{I} - \eta \mathbf{\Lambda}^2 \right)^\tau \mathbf{a} = \sum_{s=1}^{nK} (1 - \eta \lambda_s^2)^\tau a_s \mathbf{u}_s. \tag{6.5}$$

Furthermore, assuming  $\eta \leq 1/\lambda_1^2$  the linear updates  $\tilde{\boldsymbol{\theta}}_\tau$  obey

$$\|\tilde{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_0\|_{\ell_2}^2 \leq \sum_{s=1}^r \frac{a_s^2}{\lambda_s^2} + \tau^2 \eta^2 \sum_{s=r+1}^{nK} \lambda_s^2 a_s^2. \tag{6.6}$$



**Proof** Using the fact that  $\mathbf{J}\mathbf{J}^T = \mathbf{U}\mathbf{\Lambda}^2\mathbf{U}^T$  we have

$$(\mathbf{I} - \eta\mathbf{J}\mathbf{J}^T)^\tau = \mathbf{U}(\mathbf{I} - \eta\mathbf{\Lambda}^2)^\tau \mathbf{U}^T$$

Using the latter combined with (5.9) we thus have

$$\begin{aligned}\tilde{\mathbf{r}}_\tau &= (\mathbf{I} - \eta\mathbf{J}\mathbf{J}^T)^\tau \mathbf{r}_0, \\ &= \mathbf{U}(\mathbf{I} - \eta\mathbf{\Lambda}^2)^\tau \mathbf{U}^T \mathbf{r}_0, \\ &= \mathbf{U}(\mathbf{I} - \eta\mathbf{\Lambda}^2)^\tau \mathbf{a}, \\ &= \sum_{s=1}^{nK} (1 - \eta\lambda_s^2)^\tau a_s \mathbf{u}_s,\end{aligned}$$

completing the proof of (6.5).

We now turn our attention to proving (6.6) by tracking the representation of  $\tilde{\boldsymbol{\theta}}_\tau$  in terms of the right singular vectors of  $\mathbf{J}$ . To do this note that using (6.5) we have

$$\mathbf{J}^T \tilde{\mathbf{r}}_t = \mathbf{V}\mathbf{\Lambda}\mathbf{U}^T \tilde{\mathbf{r}}_t = \mathbf{V}\mathbf{\Lambda}(\mathbf{I} - \eta\mathbf{\Lambda}^2)^t \mathbf{a}.$$

Using the latter together with the gradient update on the linearized problem we have

$$\tilde{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_0 = -\eta \left( \sum_{t=0}^{\tau-1} \nabla \mathcal{L}_{lin}(\tilde{\boldsymbol{\theta}}_t) \right) = -\eta \left( \sum_{t=0}^{\tau-1} \mathbf{J}^T \tilde{\mathbf{r}}_t \right) = -\eta \mathbf{V} \left( \sum_{t=0}^{\tau-1} \mathbf{\Lambda}(\mathbf{I} - \eta\mathbf{\Lambda}^2)^t \right) \mathbf{a}.$$

Thus for any  $s \in \{1, 2, \dots, nK\}$

$$\mathbf{v}_s^T (\tilde{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_0) = -\eta \lambda_s \mathbf{a}_s \left( \sum_{t=0}^{\tau-1} (1 - \eta\lambda_s^2)^t \right) = -\eta \lambda_s \mathbf{a}_s \frac{1 - (1 - \eta\lambda_s^2)^\tau}{\eta\lambda_s^2} = -\mathbf{a}_s \frac{1 - (1 - \eta\lambda_s^2)^\tau}{\lambda_s}.$$

Noting that for  $\eta \leq 1/\lambda_1^2 \leq 1/\lambda_s^2$  we have  $1 - \eta\lambda_s^2 \geq 0$ , the latter identity implies that

$$|\mathbf{v}_s^T (\tilde{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_0)| \leq \frac{|\mathbf{a}_s|}{\lambda_s}. \quad (6.7)$$

Furthermore, using the fact that  $1 - \eta\lambda_s^2 \leq 1$  we have

$$|\mathbf{v}_s^T (\tilde{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_0)| = \eta \lambda_s |\mathbf{a}_s| \left( \sum_{t=0}^{\tau-1} (1 - \eta\lambda_s^2)^t \right) \leq \eta \lambda_s |\mathbf{a}_s| \tau \quad (6.8)$$

Combining (6.7) for  $1 \leq s \leq r$  and (6.8) for  $s > r$  we have

$$\|\tilde{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_0\|_{\ell_2}^2 = \sum_{s=1}^{nK} |\mathbf{v}_s^T (\tilde{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_0)|^2 \leq \sum_{s=1}^r \frac{a_s^2}{\lambda_s^2} + \tau^2 \eta^2 \sum_{s=r+1}^{nK} \lambda_s^2 a_s^2,$$

completing the proof of (6.6). ■

For future use we also state a simple corollary of the above Lemma below.

**Corollary 6.6** *Consider the setting and assumptions of Lemma 6.5. Then, after  $\tau$  iterations we have*

$$\|\tilde{\mathbf{r}}_\tau\|_{\ell_2} \leq (1 - \eta\alpha^2)^\tau \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}. \quad (6.9)$$

Furthermore, after  $T = \frac{\Gamma}{\eta\alpha^2}$  iterations we have

$$\|\tilde{\mathbf{r}}_T\|_{\ell_2} \leq e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}. \quad (6.10)$$

and

$$\|\tilde{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_0\|_{\ell_2}^2 \leq \sum_{s=1}^r \frac{a_s^2}{\lambda_s^2} + \Gamma^2 \sum_{s=r+1}^{nK} \frac{\lambda_s^2 a_s^2}{\alpha^4} = \frac{\mathcal{B}_{\alpha, \Gamma}^2}{\alpha^2}.$$

with  $\mathcal{B}_{\alpha, \Gamma}$  given by (6.1) per Definition 6.2.

**Proof** To prove the first bound on the residual ((6.9))note that using (6.5) we have

$$\mathbf{U}_{\mathcal{I}}^T \tilde{\mathbf{r}}_\tau = (\mathbf{I} - \eta \Lambda_{\mathcal{I}}^2)^\tau \mathbf{U}_{\mathcal{I}}^T \tilde{\mathbf{r}}_0 \quad \text{and} \quad \mathbf{U}_{\mathcal{N}}^T \tilde{\mathbf{r}}_\tau = (\mathbf{I} - \eta \Lambda_{\mathcal{N}}^2)^\tau \mathbf{U}_{\mathcal{N}}^T \tilde{\mathbf{r}}_0$$

Thus, using the fact that for  $s \leq r$  we have  $\lambda_s \geq \alpha$  we have  $(1 - \eta \lambda_s^2)^\tau \leq (1 - \eta \alpha^2)^\tau$  and for  $s > r$  we have  $(1 - \eta \lambda_s^2)^\tau \leq 1$ , we can conclude that

$$\|\mathbf{U}_{\mathcal{I}}^T \tilde{\mathbf{r}}_\tau\|_{\ell_2} \leq (1 - \eta \alpha^2)^\tau \|\mathbf{U}_{\mathcal{I}}^T \mathbf{r}_0\|_{\ell_2} \quad \text{and} \quad \|\mathbf{U}_{\mathcal{N}}^T \tilde{\mathbf{r}}_\tau\|_{\ell_2} \leq \|\mathbf{U}_{\mathcal{N}}^T \mathbf{r}_0\|_{\ell_2}.$$

Combining these with the triangular inequality we have

$$\|\tilde{\mathbf{r}}_\tau\|_{\ell_2} = \left\| \begin{bmatrix} \mathbf{U}_{\mathcal{I}}^T \tilde{\mathbf{r}}_\tau \\ \mathbf{U}_{\mathcal{N}}^T \tilde{\mathbf{r}}_\tau \end{bmatrix} \right\|_{\ell_2} \leq \|\mathbf{U}_{\mathcal{I}}^T \tilde{\mathbf{r}}_\tau\|_{\ell_2} + \|\mathbf{U}_{\mathcal{N}}^T \tilde{\mathbf{r}}_\tau\|_{\ell_2} \leq (1 - \eta \alpha^2)^\tau \|\mathbf{U}_{\mathcal{I}}^T \mathbf{r}_0\|_{\ell_2} + \|\mathbf{U}_{\mathcal{N}}^T \mathbf{r}_0\|_{\ell_2},$$

concluding the proof of (6.9). The second bound on the residual simply follows from the fact that  $(1 - \eta \alpha^2)^T \leq e^{-\Gamma}$ . The bound on  $\|\tilde{\boldsymbol{\theta}}_T - \tilde{\boldsymbol{\theta}}_0\|_{\ell_2}^2$  is trivially obtained by using  $T^2 = \frac{\Gamma^2}{\eta^2 \alpha^4}$  in (6.6). ■

The lemma above shows that with enough iterations, gradient descent on the linearized problem fits the residual over the information space and the residual is (in the worst case) unchanged over the nuisance subspace  $\mathcal{N}$ . Our hypothesis is that, when the model is generalizable the residual mostly lies on the information space  $\mathcal{I}$  which contains the directions aligned with the top singular vectors. Hence, the smaller term  $\|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}$  over the nuisance space will not affect generalization significantly. To make this intuition precise however we need to connect the residual of the original problem to that of the linearized problem. The following lemma sheds light on the evolution of the original problem (5.4) by characterizing the evolution of the difference between the residuals of the original and linearized problems from one iteration to the next.

**Lemma 6.7 (Keeping track of perturbation - one step)** Assume Assumptions 1 and 2 hold and  $\boldsymbol{\theta}_\tau$  and  $\boldsymbol{\theta}_{\tau+1}$  are within an  $R$  neighborhood of  $\boldsymbol{\theta}_0$ , that is,

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} \leq R \quad \text{and} \quad \|\boldsymbol{\theta}_{\tau+1} - \boldsymbol{\theta}_0\|_{\ell_2} \leq R.$$

Then with a learning rate obeying  $\eta \leq 1/\beta^2$ , the deviation in the residuals of the original and linearized problems  $\mathbf{e}_{\tau+1} = \mathbf{r}_{\tau+1} - \tilde{\mathbf{r}}_{\tau+1}$  obey

$$\|\mathbf{e}_{\tau+1}\|_{\ell_2} \leq \eta(\varepsilon_0^2 + \varepsilon\beta) \|\tilde{\mathbf{r}}_\tau\|_{\ell_2} + (1 + \eta\varepsilon^2) \|\mathbf{e}_\tau\|_{\ell_2}. \quad (6.11)$$

**Proof** For simplicity, denote  $\mathbf{B}_1 = \mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau)$ ,  $\mathbf{B}_2 = \mathcal{J}(\boldsymbol{\theta}_\tau)$ ,  $\mathbf{A} = \mathcal{J}(\boldsymbol{\theta}_0)$  where

$$\mathcal{J}(\mathbf{b}, \mathbf{a}) = \int_0^1 \mathcal{J}(t\mathbf{b} + (1-t)\mathbf{a}) dt.$$

We can write the predictions due to  $\boldsymbol{\theta}_{\tau+1}$  as

$$\begin{aligned} f(\boldsymbol{\theta}_{\tau+1}) &= f(\boldsymbol{\theta}_\tau - \eta \nabla \mathcal{L}(\boldsymbol{\theta}_\tau)) = f(\boldsymbol{\theta}_\tau) + \eta \mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau) \nabla \mathcal{L}(\boldsymbol{\theta}_\tau) \\ &= f(\boldsymbol{\theta}_\tau) + \eta \mathcal{J}(\boldsymbol{\theta}_{\tau+1}, \boldsymbol{\theta}_\tau) \mathcal{J}^T(\boldsymbol{\theta}_\tau) (f(\boldsymbol{\theta}_\tau) - \mathbf{y}). \end{aligned}$$

This implies that

$$\mathbf{r}_{\tau+1} = f(\boldsymbol{\theta}_{\tau+1}) - \mathbf{y} = (\mathbf{I} - \eta \mathbf{B}_1 \mathbf{B}_2^T) \mathbf{r}_\tau.$$

Similarly, for linearized problem we have  $\tilde{\mathbf{r}}_{\tau+1} = (\mathbf{I} - \eta \mathbf{J} \mathbf{J}^T) \tilde{\mathbf{r}}_\tau$ . Thus,

$$\begin{aligned} \|\mathbf{e}_{\tau+1}\|_{\ell_2} &= \|(\mathbf{I} - \eta \mathbf{B}_1 \mathbf{B}_2^T) \mathbf{r}_\tau - (\mathbf{I} - \eta \mathbf{J} \mathbf{J}^T) \tilde{\mathbf{r}}_\tau\|_{\ell_2} \\ &= \|(\mathbf{I} - \eta \mathbf{B}_1 \mathbf{B}_2^T) \mathbf{e}_\tau - \eta (\mathbf{B}_1 \mathbf{B}_2^T - \mathbf{J} \mathbf{J}^T) \tilde{\mathbf{r}}_\tau\|_{\ell_2} \\ &\leq \|(\mathbf{I} - \eta \mathbf{B}_1 \mathbf{B}_2^T) \mathbf{e}_\tau\|_{\ell_2} + \eta \|(\mathbf{B}_1 \mathbf{B}_2^T - \mathbf{J} \mathbf{J}^T) \tilde{\mathbf{r}}_\tau\|_{\ell_2} \\ &\leq \|(\mathbf{I} - \eta \mathbf{B}_1 \mathbf{B}_2^T) \mathbf{e}_\tau\|_{\ell_2} + \eta \|(\mathbf{B}_1 \mathbf{B}_2^T - \mathbf{J} \mathbf{J}^T)\| \|\tilde{\mathbf{r}}_\tau\|_{\ell_2}. \end{aligned} \quad (6.12)$$

We proceed by bounding each of these two terms. For the first term, we apply Lemma 6.3 with  $\mathbf{A} = \mathbf{B}_1$  and  $\mathbf{B} = \mathbf{B}_2$  and use  $\|\mathbf{B}_1 - \mathbf{B}_2\| \leq \varepsilon$  to conclude that

$$\|(I - \eta \mathbf{B}_1 \mathbf{B}_2^T) \mathbf{e}_\tau\|_{\ell_2} \leq (1 + \eta \varepsilon^2) \|\mathbf{e}_\tau\|_{\ell_2}. \quad (6.13)$$

Next we turn our attention to bounding the second term. To this aim note that

$$\begin{aligned} \|\mathbf{B}_1 \mathbf{B}_2^T - \mathbf{J} \mathbf{J}^T\| &= \|\mathbf{B}_1 \mathbf{B}_2^T - \mathbf{A} \mathbf{A}^T + \mathbf{A} \mathbf{A}^T - \mathbf{J} \mathbf{J}^T\| \\ &\leq \|\mathbf{B}_1 \mathbf{B}_2^T - \mathbf{A} \mathbf{A}^T\| + \|\mathbf{A} \mathbf{A}^T - \mathbf{J} \mathbf{J}^T\| \\ &\leq \|(\mathbf{B}_1 - \mathbf{A}) \mathbf{B}_2^T\| + \|\mathbf{A} (\mathbf{B}_2 - \mathbf{A})^T\| + \|\mathbf{A} \mathbf{A}^T - \mathbf{J} \mathbf{J}^T\| \\ &\leq \|\mathbf{B}_1 - \mathbf{A}\| \|\mathbf{B}_2\| + \|\mathbf{B}_2 - \mathbf{A}\| \|\mathbf{A}\| + \|\mathbf{A} \mathbf{A}^T - \mathbf{J} \mathbf{J}^T\| \\ &\leq \beta \frac{\varepsilon}{2} + \beta \frac{\varepsilon}{2} + \varepsilon_0^2 \\ &= \varepsilon_0^2 + \varepsilon \beta. \end{aligned} \quad (6.14)$$

In the last inequality we use the fact that per Assumption 2 we have  $\|\mathbf{B}_1 - \mathbf{A}\| \leq \varepsilon/2$  and  $\|\mathbf{B}_2 - \mathbf{A}\| \leq \varepsilon/2$  as well as the fact that per Definition 5.1  $\|\mathbf{A} \mathbf{A}^T - \mathbf{J} \mathbf{J}^T\| \leq \varepsilon_0^2$ . Plugging (6.13) and (6.14) in (6.12) completes the proof.  $\blacksquare$

Next we prove a result about the growth of sequences obeying certain assumptions. As we will see later on in the proofs this lemma allows us to control the growth of the perturbation between the original and linearized residuals ( $e_\tau = \|\mathbf{e}_\tau\|_{\ell_2}$ ).

**Lemma 6.8 (Bounding residual perturbation growth for general nonlinearities)** *Consider positive scalars  $\Gamma, \alpha, \varepsilon, \eta > 0$ . Also assume  $\eta \leq 1/\alpha^2$  and  $\alpha \geq \sqrt{2\Gamma\varepsilon}$  and set  $T = \frac{\Gamma}{\eta\alpha^2}$ . Assume the scalar sequences  $e_\tau$  (with  $e_0 = 0$ ) and  $\tilde{r}_\tau$  obey the following identities*

$$\begin{aligned} \tilde{r}_\tau &\leq (1 - \eta\alpha^2)^\tau \rho_+ + \rho_-, \\ e_\tau &\leq (1 + \eta\varepsilon^2) e_{\tau-1} + \eta\Theta \tilde{r}_{\tau-1}, \end{aligned} \quad (6.15)$$

for all  $0 \leq \tau \leq T$  and non-negative values  $\rho_-, \rho_+ \geq 0$ . Then, for all  $0 \leq \tau \leq T$ ,

$$e_\tau \leq \Theta \Lambda \quad \text{holds with} \quad \Lambda = \frac{2(\Gamma\rho_- + \rho_+)}{\alpha^2}. \quad (6.16)$$

**Proof** We shall prove the result inductively. Suppose (6.16) holds for all  $t \leq \tau - 1$ . Consequently, we have

$$\begin{aligned} e_{t+1} &\leq (1 + \eta\varepsilon^2) e_t + \eta\Theta \tilde{r}_t \\ &\leq e_t + \eta\varepsilon^2 e_t + \eta\Theta ((1 - \eta\alpha^2)^t \rho_+ + \rho_-) \\ &\leq e_t + \eta\Theta (\varepsilon^2 \Lambda + (1 - \eta\alpha^2)^t \rho_+ + \rho_-). \end{aligned}$$

Thus

$$\frac{e_{t+1} - e_t}{\Theta} \leq \eta (\varepsilon^2 \Lambda + (1 - \eta\alpha^2)^t \rho_+ + \rho_-). \quad (6.17)$$

Summing up both sides of (6.17) for  $0 \leq t \leq \tau - 1$  we conclude that

$$\begin{aligned}
\frac{e_\tau}{\Theta} &= \sum_{t=0}^{\tau-1} \frac{e_{t+1} - e_t}{\Theta} \\
&\leq \eta\tau(\varepsilon^2\Lambda + \rho_-) + \eta\rho_+ \sum_{t=0}^{\tau-1} (1 - \eta\alpha^2)^t \\
&= \eta\tau(\varepsilon^2\Lambda + \rho_-) + \eta\rho_+ \frac{1 - (1 - \eta\alpha^2)^\tau}{\eta\alpha^2} \\
&\leq \eta\left(\tau\varepsilon^2\Lambda + \frac{\rho_+}{\eta\alpha^2} + \tau\rho_-\right) \\
&= \eta\tau(\varepsilon^2\Lambda + \rho_-) + \frac{\rho_+}{\alpha^2} \\
&\leq \eta T(\varepsilon^2\Lambda + \rho_-) + \frac{\rho_+}{\alpha^2} \\
&= \frac{\Gamma\varepsilon^2\Lambda + \Gamma\rho_- + \rho_+}{\alpha^2} \\
&= \frac{\Gamma\varepsilon^2\Lambda}{\alpha^2} + \frac{\Lambda}{2} \\
&\leq \Lambda,
\end{aligned}$$

where in the last inequality we used the fact that  $\alpha^2 \geq 2\Gamma\varepsilon^2$ . This completes the proof of the induction step and the proof of the lemma.  $\blacksquare$

### 6.1.3 Completing the proof of Theorem 5.3

With the key lemmas in place in this section we wish to complete the proof of Theorem 5.3. We will use induction to prove the result. Suppose the statement is true for some  $\tau - 1 \leq T - 1$ . In particular, we assume the identities (5.12) and (5.13) hold for all  $0 \leq t \leq \tau - 1$ . We aim to prove these identities continue to hold for iteration  $\tau$ . We will prove this result in multiple steps.

**Step I: Next iterate obeys  $\|\theta_\tau - \theta_0\|_{\ell_2} \leq R$ .**

We first argue that  $\theta_\tau$  lies in the domain of interest as dictated by (5.10), i.e.  $\|\theta_\tau - \theta_0\|_{\ell_2} \leq R$ . To do this note that per the induction assumption (5.13) holds for iteration  $\tau - 1$  and thus  $\|\theta_{\tau-1} - \theta_0\|_{\ell_2} \leq R/2$ . As a result using the triangular inequality to show  $\|\theta_\tau - \theta_0\|_{\ell_2} \leq R$  holds it suffices to show that  $\|\theta_\tau - \theta_{\tau-1}\|_{\ell_2} \leq R/2$

holds. To do this note that

$$\begin{aligned}
\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_{\tau-1}\|_{\ell_2} &= \eta \|\nabla \mathcal{L}(\boldsymbol{\theta}_{\tau-1})\|_{\ell_2} \\
&= \eta \|\mathcal{J}^T(\boldsymbol{\theta}_{\tau-1}) \mathbf{r}_{\tau-1}\|_{\ell_2} \\
&= \eta \|\overline{\mathcal{J}}^T(\boldsymbol{\theta}_{\tau-1}) \mathbf{r}_{\tau-1}\|_{\ell_2} \\
&\stackrel{(a)}{\leq} \eta \|\overline{\mathcal{J}}^T(\boldsymbol{\theta}_{\tau-1}) \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + \eta \|\overline{\mathcal{J}}^T(\boldsymbol{\theta}_{\tau-1})(\mathbf{r}_{\tau-1} - \tilde{\mathbf{r}}_{\tau-1})\|_{\ell_2} \\
&\stackrel{(b)}{\leq} \eta \|\mathbf{J}^T \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + \eta \|\overline{\mathcal{J}}(\boldsymbol{\theta}_{\tau-1}) - \mathbf{J}\| \|\tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + \eta \|\overline{\mathcal{J}}(\boldsymbol{\theta}_{\tau-1})\| \|\mathbf{r}_{\tau-1} - \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} \\
&\stackrel{(c)}{\leq} \eta \|\mathbf{J}^T \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + \frac{\varepsilon_0 + \varepsilon}{\beta^2} \|\tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + \frac{1}{\beta} \|\mathbf{r}_{\tau-1} - \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} \\
&\stackrel{(d)}{\leq} \eta \|\mathbf{J}^T \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + \frac{2\delta\alpha}{5\beta^2} \|\mathbf{r}_0\|_{\ell_2} + \frac{1}{\beta} \|\mathbf{r}_{\tau-1} - \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} \\
&\stackrel{(e)}{\leq} \eta \|\mathbf{J}^T \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + \frac{2\delta\alpha}{5\beta^2} \|\mathbf{r}_0\|_{\ell_2} + \frac{3\delta\alpha}{5\beta^2} \|\mathbf{r}_0\|_{\ell_2} \\
&= \eta \|\mathbf{J}^T \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + \frac{\delta\alpha}{\beta^2} \|\mathbf{r}_0\|_{\ell_2} \\
&\stackrel{(f)}{\leq} \eta \beta^2 \frac{\mathcal{B}_{\alpha,\Gamma}}{\alpha} + \frac{\delta\alpha}{\beta^2} \|\mathbf{r}_0\|_{\ell_2} \\
&\stackrel{(g)}{\leq} \frac{\mathcal{B}_{\alpha,\Gamma}}{\alpha} + \frac{\delta\alpha}{\beta^2} \|\mathbf{r}_0\|_{\ell_2} \\
&\stackrel{(h)}{\leq} \frac{\mathcal{B}_{\alpha,\Gamma}}{\alpha} + \frac{\delta\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \\
&= \frac{R}{2}.
\end{aligned}$$

Here, (a) and (b) follow from a simple application of the triangular inequality, (c) from the fact that  $\|\mathcal{J}(\boldsymbol{\theta}_{\tau-1}) - \mathbf{J}\| \leq \|\mathcal{J}(\boldsymbol{\theta}_{\tau-1}) - \mathcal{J}(\boldsymbol{\theta}_0)\| + \|\mathcal{J}(\boldsymbol{\theta}_0) - \mathbf{J}\| \leq \varepsilon + \varepsilon_0$ , (d) from combining the bounds in (5.11), (e) from the induction hypothesis that postulates (5.12) holds for iteration  $\tau - 1$ , (f) from considering the SVD  $\mathbf{J} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{V}^T$  which implies that

$$\begin{aligned}
\|\mathbf{J}^T \tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2}^2 &= \|\mathbf{J}^T (\mathbf{I} - \eta \mathbf{J} \mathbf{J}^T)^{\tau-1} \mathbf{r}_0\|_{\ell_2}^2 = \|\mathbf{V} \boldsymbol{\Lambda} (\mathbf{I} - \eta \boldsymbol{\Lambda}^2)^{\tau-1} \mathbf{U}^T \mathbf{r}_0\|_{\ell_2}^2 \\
&= \|\boldsymbol{\Lambda} (\mathbf{I} - \eta \boldsymbol{\Lambda}^2)^{\tau-1} \mathbf{U}^T \mathbf{r}_0\|_{\ell_2}^2 \\
&= \sum_{s=1}^{nK} \lambda_s^2 (1 - \eta \lambda_s^2)^{2(\tau-1)} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \\
&\leq \sum_{s=1}^{nK} \lambda_s^2 (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \\
&= \sum_{s=1}^r \lambda_s^2 (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 + \sum_{s=r+1}^{nK} \lambda_s^2 (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \\
&\leq \beta^4 \sum_{s=1}^r \frac{1}{\lambda_s^2} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 + \sum_{s=r+1}^{nK} \lambda_s^2 (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \\
&\leq \beta^4 \left( \sum_{s=1}^r \frac{1}{\lambda_s^2} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 + \Gamma^2 \sum_{s=r+1}^{nK} \frac{\lambda_s^2}{\alpha^4} (\langle \mathbf{u}_s, \mathbf{r}_0 \rangle)^2 \right) \\
&= \beta^4 \left( \frac{\mathcal{B}_{\alpha,\Gamma}}{\alpha} \right)^2
\end{aligned}$$

(g) from the fact that  $\eta \leq \frac{1}{\beta^2}$ , and (h) from the fact that  $\alpha \leq \beta$  and  $\Gamma \geq 1$ .

**Step II: Original and linearized residuals are close (first part of (5.12)).**

In this step we wish to show that the first part of (5.12) holds for iteration  $\tau$ . Since we established in the previous step that  $\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} \leq R$  the assumption of Lemma 6.7 holds for iterations  $\tau - 1$  and  $\tau$ . Hence, using Lemma 6.7 equation (6.11) we conclude that

$$\|\mathbf{e}_\tau\|_{\ell_2} \leq \eta(\varepsilon_0^2 + \varepsilon\beta)\|\tilde{\mathbf{r}}_{\tau-1}\|_{\ell_2} + (1 + \eta\varepsilon^2)\|\mathbf{e}_{\tau-1}\|_{\ell_2}.$$

This combined with the induction assumption implies that

$$\|\mathbf{e}_t\|_{\ell_2} \leq \eta(\varepsilon_0^2 + \varepsilon\beta)\|\tilde{\mathbf{r}}_{t-1}\|_{\ell_2} + (1 + \eta\varepsilon^2)\|\mathbf{e}_{t-1}\|_{\ell_2}, \quad (6.18)$$

holds for all  $t \leq \tau \leq T$ . Furthermore, using Lemma 6.5 equation (6.9) for all  $t \leq \tau \leq T$  we have

$$\|\tilde{\mathbf{r}}_t\|_{\ell_2} \leq (1 - \eta\alpha^2)^t \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}, \quad (6.19)$$

To proceed, we shall apply Lemma 6.8 with the following variable substitutions

$$\Theta := \varepsilon_0^2 + \varepsilon\beta, \quad \rho_+ = \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2}, \quad \rho_- = \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}, \quad e_\tau := \|\mathbf{e}_\tau\|_{\ell_2}, \quad \tilde{r}_\tau := \|\tilde{\mathbf{r}}_\tau\|_{\ell_2}. \quad (6.20)$$

We note that Lemma 6.8 is applicable since (i)  $\eta \leq 1/\beta^2 \leq 1/\alpha^2$ , (ii) based on (5.11) we have  $\frac{\alpha}{\varepsilon} \geq \frac{5\Gamma}{\delta} \frac{\beta^2}{\alpha^2} \geq \sqrt{2\Gamma}$ , (iii)  $\tau$  obeys  $\tau \leq T = \frac{\Gamma}{\eta\alpha^2}$ , and (iv) (6.15) holds based on (6.18) and (6.19). Thus using Lemma 6.8 we can conclude that

$$\begin{aligned} \|\mathbf{e}_\tau\|_{\ell_2} &\leq 2(\varepsilon_0^2 + \varepsilon\beta) \frac{(\|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \Gamma\|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2})}{\alpha^2} \\ &\leq \frac{2\Gamma(\varepsilon_0^2 + \varepsilon\beta)\|\mathbf{r}_0\|_{\ell_2}}{\alpha^2} \end{aligned} \quad (6.21)$$

$$\leq \left(\frac{2}{25} + \frac{2}{5}\right) \frac{\delta\alpha}{\beta} \|\mathbf{r}_0\|_{\ell_2} \leq \frac{3}{5} \frac{\delta\alpha}{\beta} \|\mathbf{r}_0\|_{\ell_2}, \quad (6.22)$$

where in the last inequality we used (5.11). This completes the first part of (5.12) via induction.

**Step III: Original and linearized parameters are close (second part of (5.12)).**

In this step we wish to show that the second part of (5.12) holds for iteration  $\tau$ . To do this we begin by noting that by the fact that  $\mathbf{J}$  is a reference Jacobian we have  $\|\bar{\mathcal{J}}(\boldsymbol{\theta}_0) - \mathbf{J}\| \leq \varepsilon_0$  where  $\bar{\mathcal{J}}$  augments  $\mathcal{J}(\boldsymbol{\theta}_0)$  by padding zero columns to match size of  $\mathbf{J}$ . Also by Assumption 2 we have  $\|\mathcal{J}(\boldsymbol{\theta}) - \bar{\mathcal{J}}(\boldsymbol{\theta}_0)\| \leq \frac{\varepsilon}{2}$ . Combining the latter two via the triangular inequality we conclude that

$$\|\bar{\mathcal{J}}(\boldsymbol{\theta}_\tau) - \mathbf{J}\| \leq \varepsilon_0 + \varepsilon. \quad (6.23)$$

Let  $\bar{\boldsymbol{\theta}}$  and  $\nabla \bar{\mathcal{L}}(\boldsymbol{\theta})$  be vectors augmented by zero padding  $\boldsymbol{\theta}, \nabla \mathcal{L}(\boldsymbol{\theta})$  so that they have dimension  $\max(Kn, p)$ . Now, we track the difference between  $\bar{\boldsymbol{\theta}}$  and linearized  $\boldsymbol{\theta}$  as follows

$$\begin{aligned} \frac{\|\bar{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_\tau\|_{\ell_2}}{\eta} &= \left\| \sum_{t=0}^{\tau-1} \nabla \bar{\mathcal{L}}(\boldsymbol{\theta}_t) - \nabla \mathcal{L}_{lin}(\tilde{\boldsymbol{\theta}}_t) \right\|_{\ell_2} \\ &= \left\| \sum_{t=0}^{\tau-1} \bar{\mathcal{J}}(\boldsymbol{\theta}_t)^T \mathbf{r}_t - \mathbf{J}^T \tilde{\mathbf{r}}_t \right\|_{\ell_2} \\ &\leq \sum_{t=0}^{\tau-1} \|\bar{\mathcal{J}}(\boldsymbol{\theta}_t)^T \mathbf{r}_t - \mathbf{J}^T \tilde{\mathbf{r}}_t\|_{\ell_2} \\ &\leq \sum_{t=0}^{\tau-1} \|(\bar{\mathcal{J}}(\boldsymbol{\theta}_t) - \mathbf{J})^T \tilde{\mathbf{r}}_t\|_{\ell_2} + \|\bar{\mathcal{J}}(\boldsymbol{\theta}_t)^T (\mathbf{r}_t - \tilde{\mathbf{r}}_t)\|_{\ell_2} \\ &= \sum_{t=0}^{\tau-1} \|(\bar{\mathcal{J}}(\boldsymbol{\theta}_t) - \mathbf{J})^T \tilde{\mathbf{r}}_t\|_{\ell_2} + \|\bar{\mathcal{J}}(\boldsymbol{\theta}_t)^T \mathbf{e}_t\|_{\ell_2} \\ &\leq \sum_{t=0}^{\tau-1} (\varepsilon + \varepsilon_0) \|\tilde{\mathbf{r}}_t\|_{\ell_2} + \beta \|\mathbf{e}_t\|_{\ell_2}. \end{aligned} \quad (6.24)$$

In the last inequality we used the fact that  $\|\bar{\mathcal{J}}(\boldsymbol{\theta}_t) - \mathbf{J}\| \leq \varepsilon + \varepsilon_0$  and  $\|\mathbf{J}\| \leq \beta$ . We proceed by bounding each of the two terms in (6.24) above. For the first term we use the fact that  $\|\tilde{\mathbf{r}}_\tau\|_{\ell_2} \leq \|\mathbf{r}_0\|_{\ell_2}$  to conclude

$$\sum_{t=0}^{\tau-1} \|\tilde{\mathbf{r}}_t\|_{\ell_2} \leq \tau \|\mathbf{r}_0\|_{\ell_2} \leq T \|\mathbf{r}_0\|_{\ell_2} = \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\eta \alpha^2}. \quad (6.25)$$

To bound the second term in (6.24) we use (6.21) together with  $\tau \leq T \leq \frac{\Gamma}{\eta \alpha^2}$  to conclude that

$$\sum_{t=0}^{\tau-1} \|\mathbf{e}_t\|_{\ell_2} \leq \tau \frac{2(\varepsilon\beta + \varepsilon_0^2)}{\alpha^2} \Gamma \|\mathbf{r}_0\|_{\ell_2} \leq \frac{2\Gamma^2(\varepsilon\beta + \varepsilon_0^2)}{\eta \alpha^4} \|\mathbf{r}_0\|_{\ell_2}. \quad (6.26)$$

Combining (6.25) and (6.26) in (6.24), we conclude that

$$\begin{aligned} \|\bar{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_\tau\|_{\ell_2} &\leq \left( \frac{2\Gamma(\varepsilon\beta^2 + \varepsilon_0^2\beta)}{\alpha^3} + \frac{\varepsilon + \varepsilon_0}{\alpha} \right) \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \\ &= \left( \varepsilon \frac{2\Gamma\beta^2}{\alpha^3} + \varepsilon_0^2 \frac{2\Gamma\beta}{\alpha^3} + \frac{\varepsilon + \varepsilon_0}{\alpha} \right) \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \\ &\stackrel{(a)}{\leq} \left( \frac{2}{5}\delta + \varepsilon_0^2 \frac{2\Gamma\beta}{\alpha^3} + \frac{\varepsilon + \varepsilon_0}{\alpha} \right) \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \\ &\stackrel{(b)}{\leq} \left( \frac{2}{5}\delta + \frac{2}{25}\delta + \frac{\varepsilon + \varepsilon_0}{\alpha} \right) \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \\ &\stackrel{(c)}{\leq} \left( \frac{2}{5}\delta + \frac{2}{25}\delta + \frac{1}{5}\delta + \frac{\varepsilon_0}{\alpha} \right) \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \\ &\stackrel{(d)}{\leq} \left( \frac{2}{5}\delta + \frac{2}{25}\delta + \frac{1}{5}\delta + \frac{1}{5}\delta \right) \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \\ &= \frac{22}{25} \frac{\delta}{\alpha} \Gamma \|\mathbf{r}_0\|_{\ell_2}. \end{aligned}$$

Here, (a) follows from  $\varepsilon \leq \frac{\delta \alpha^3}{5\Gamma\beta^2}$  per Assumption (5.11), (b) from  $\varepsilon_0 \leq \frac{1}{5}\sqrt{\frac{\delta \alpha^3}{\Gamma\beta}}$  per Assumption (5.11), (c) from  $\varepsilon \leq \frac{\delta \alpha^3}{5\Gamma\beta^2} \leq \frac{\delta \alpha}{5\Gamma} \leq \frac{\delta \alpha}{5}$  per Assumption (5.11), and (d) from  $\varepsilon_0 \leq \frac{\delta \alpha}{5}$  per Assumption (5.11). Thus,

$$\|\bar{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_\tau\|_{\ell_2} \leq \frac{\delta}{\alpha} \Gamma \|\mathbf{r}_0\|_{\ell_2}.$$

Combining the latter with the fact that  $\|\tilde{\boldsymbol{\theta}}_\tau - \bar{\boldsymbol{\theta}}_0\|_{\ell_2} \leq \frac{\mathcal{B}_{\alpha,\Gamma}}{\alpha}$  (which follows from Lemma 6.5 equation (6.6)) we conclude that

$$\|\boldsymbol{\theta}_\tau - \boldsymbol{\theta}_0\|_{\ell_2} = \|\bar{\boldsymbol{\theta}}_\tau - \bar{\boldsymbol{\theta}}_0\|_{\ell_2} \leq \|\tilde{\boldsymbol{\theta}}_\tau - \bar{\boldsymbol{\theta}}_0\|_{\ell_2} + \|\bar{\boldsymbol{\theta}}_\tau - \tilde{\boldsymbol{\theta}}_\tau\|_{\ell_2} \leq \frac{\mathcal{B}_{\alpha,\Gamma}}{\alpha} + \frac{\delta}{\alpha} \Gamma \|\mathbf{r}_0\|_{\ell_2} \leq \|\mathbf{J}_{\mathcal{I}}^\dagger \mathbf{r}_0\|_{\ell_2} + \frac{\Gamma}{\alpha} \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \frac{\delta}{\alpha} \Gamma \|\mathbf{r}_0\|_{\ell_2}$$

The completes the proof of the bound (5.13).

**Step V: Bound on residual with early stopping.**

In this step we wish to prove (5.14). To this aim note that

$$\begin{aligned} \|\mathbf{r}_T\|_{\ell_2} &\stackrel{(a)}{\leq} \|\tilde{\mathbf{r}}_T\|_{\ell_2} + \|\tilde{\mathbf{r}}_T - \mathbf{r}_T\|_{\ell_2} \\ &\stackrel{(b)}{\leq} \|\tilde{\mathbf{r}}_T\|_{\ell_2} + \frac{\delta \alpha}{\beta} \|\mathbf{r}_0\|_{\ell_2} \\ &\stackrel{(c)}{\leq} e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \frac{\delta \alpha}{\beta} \|\mathbf{r}_0\|_{\ell_2} \end{aligned}$$

where (a) follows from the triangular inequality, (b) from the conclusion of Step II (first part of (5.12)), and (c) from Corollary 6.6 equation (6.10). This completes the proof of (5.14).



## 6.2 Key lemmas and identities for neural networks

In this section we prove some key lemmas and identities regarding the Jacobian of one-hidden layer networks as well as the size of the initial residual that when combined with Theorem 5.3 allows us to prove theorems involving neural networks. We begin with some preliminary identities and calculations in Section 6.2.1. Next, in Section 6.2.2 we prove a few key properties of the Jacobian mapping of a one-hidden layer neural network. Section 6.2.3 focuses on a few further properties of the Jacobian at a random initialization. Finally, in Section 6.2.4 we provide bounds on the initial misfit.

For two matrices

$$\mathbf{A} = \begin{bmatrix} \mathbf{A}_1 \\ \mathbf{A}_2 \\ \vdots \\ \mathbf{A}_p \end{bmatrix} \in \mathbb{R}^{p \times m} \quad \text{and} \quad \mathbf{B} = \begin{bmatrix} \mathbf{B}_1 \\ \mathbf{B}_2 \\ \vdots \\ \mathbf{B}_p \end{bmatrix} \in \mathbb{R}^{p \times n},$$

we define their Khatri-Rao product as  $\mathbf{A} * \mathbf{B} = [\mathbf{A}_1 \otimes \mathbf{B}_1, \dots, \mathbf{A}_p \otimes \mathbf{B}_p] \in \mathbb{R}^{p \times mn}$ , where  $\otimes$  denotes the Kronecker product.

### 6.2.1 Preliminary identities and calculations

We begin by discussing a few notations. Throughout we use  $\mathbf{w}_\ell$  and  $\mathbf{v}_\ell$  to denote the  $\ell$ th row of input and output weight matrices  $\mathbf{W}$  and  $\mathbf{V}$ . Given a matrix  $\mathbf{M}$  we use  $\|\mathbf{M}\|_{2,\infty}$  to denote the largest Euclidean norm of the rows of  $\mathbf{M}$ . We begin by noting that for a one-hidden layer neural network of the form  $\mathbf{x} \mapsto \mathbf{V}\phi(\mathbf{W}\mathbf{x})$ , the Jacobian matrix with respect to  $\text{vect}(\mathbf{W}) \in \mathbb{R}^{kd}$  takes the form

$$\mathcal{J}(\mathbf{W}) = \begin{bmatrix} \mathcal{J}_1(\mathbf{W}) \\ \vdots \\ \mathcal{J}_K(\mathbf{W}) \end{bmatrix} \in \mathbb{R}^{Kn \times kd} \quad (6.27)$$

where  $\mathcal{J}_\ell(\mathbf{W})$  is the Jacobian matrix associated with the  $\ell$ th class. In particular,  $\mathcal{J}_\ell(\mathbf{W})$  is given by

$$\mathcal{J}_\ell(\mathbf{W}) = [\mathcal{J}_\ell(\mathbf{w}_1) \quad \dots \quad \mathcal{J}_\ell(\mathbf{w}_k)] \in \mathbb{R}^{n \times kd} \quad \text{with} \quad \mathcal{J}_\ell(\mathbf{w}_s) := \mathbf{V}_{\ell,s} \text{diag}(\phi'(\mathbf{X}\mathbf{w}_s)) \mathbf{X}.$$

Alternatively using Khatri-Rao products this can be rewritten in the more compact form

$$\mathcal{J}_\ell(\mathbf{W}) = (\phi'(\mathbf{X}\mathbf{W}^T) \text{diag}(\mathbf{v}_\ell)) * \mathbf{X}. \quad (6.28)$$

An alternative characterization of the Jacobian is via its matrix representation. Given a vector  $\mathbf{u} \in \mathbb{R}^{Kn}$  let us partition it into  $K$  size  $n$  subvectors so that  $\mathbf{u} = [\mathbf{u}_1^T \dots \mathbf{u}_K^T]^T$ . We have

$$\text{mat}(\mathcal{J}^T(\mathbf{W})\mathbf{u}) = \sum_{\ell=1}^K \text{diag}(\mathbf{v}_\ell) \phi'(\mathbf{W}\mathbf{X}^T) \text{diag}(\mathbf{u}_\ell) \mathbf{X}. \quad (6.29)$$

### 6.2.2 Fundamental properties of the Jacobian of the neural network

In this section we prove a few key properties of the Jacobian mapping of a one-hidden layer neural network.

**Lemma 6.9 (Properties of Single Output Neural Net Jacobian)** *Let  $K = 1$  so that  $\mathbf{V}^T = \mathbf{v} \in \mathbb{R}^n$ . Suppose  $\phi$  is an activation obeying  $|\phi'(z)| \leq B$  for all  $z$ . Then, for any  $\mathbf{W} \in \mathbb{R}^{k \times d}$  and any unit length vector  $\mathbf{u}$ , we have*

$$\|\mathcal{J}(\mathbf{W})\| \leq B\sqrt{k} \|\mathbf{v}\|_{\ell_\infty} \|\mathbf{X}\|$$

and

$$\|\text{mat}(\mathcal{J}^T(\mathbf{W})\mathbf{u})\|_{2,\infty} \leq B \|\mathbf{v}\|_{\ell_\infty} \|\mathbf{X}\| \quad (6.30)$$

Furthermore, suppose  $\phi$  is twice differentiable and  $|\phi''(z)| \leq B$  for all  $z$ . Also assume all data points have unit Euclidean norm ( $\|\mathbf{x}_i\|_{\ell_2} = 1$ ). Then the Jacobian mapping is Lipschitz with respect to spectral norm i.e. for all  $\widetilde{\mathbf{W}}, \mathbf{W} \in \mathbb{R}^{k \times d}$  we have

$$\|\mathcal{J}(\widetilde{\mathbf{W}}) - \mathcal{J}(\mathbf{W})\| \leq B \|\mathbf{v}\|_{\ell_\infty} \|\mathbf{X}\| \|\widetilde{\mathbf{W}} - \mathbf{W}\|_F.$$

**Proof** The result on spectral norm and Lipschitzness of  $\mathcal{J}(\mathbf{W})$  have been proven in [48]. To show the row-wise bound (6.30), we use (6.29) to conclude that

$$\begin{aligned} \|\text{mat}(\mathcal{J}^T(\mathbf{W})\mathbf{u})\|_{2,\infty} &= \|\text{diag}(\mathbf{v})\phi'(\mathbf{W}\mathbf{X}^T)\text{diag}(\mathbf{u})\mathbf{X}\|_{2,\infty} \\ &\leq \|\mathbf{v}\|_{\ell_\infty} \max_{1 \leq \ell \leq k} \|\phi'(\mathbf{w}_\ell^T \mathbf{X}^T)\text{diag}(\mathbf{u})\mathbf{X}\|_{\ell_2} \\ &\leq \|\mathbf{v}\|_{\ell_\infty} \|\mathbf{X}\| \max_{1 \leq \ell \leq k} \|\phi'(\mathbf{w}_\ell^T \mathbf{X}^T)\text{diag}(\mathbf{u})\|_{\ell_2} \\ &\leq B \|\mathbf{v}\|_{\ell_\infty} \|\mathbf{X}\| \|\mathbf{u}\|_{\ell_2} \\ &= B \|\mathbf{v}\|_{\ell_\infty} \|\mathbf{X}\|. \end{aligned}$$

■

Next we extend the lemma above to the multi-class setting.

**Lemma 6.10 (Properties of Multiclass Neural Net Jacobian)** Suppose  $\phi$  is an activation obeying  $|\phi'(z)| \leq B$  for all  $z$ . Then, for any  $\mathbf{W} \in \mathbb{R}^{k \times d}$  and any unit length vector  $\mathbf{u}$ , we have

$$\|\mathcal{J}(\mathbf{W})\| \leq B\sqrt{K} \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\|$$

and

$$\|\text{mat}(\mathcal{J}^T(\mathbf{W})\mathbf{u})\|_{2,\infty} \leq B\sqrt{K} \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\|. \quad (6.31)$$

Furthermore, suppose  $\phi$  is twice differentiable and  $|\phi''(z)| \leq B$  for all  $z$ . Also assume all data points have unit Euclidean norm ( $\|\mathbf{x}_i\|_{\ell_2} = 1$ ). Then the Jacobian mapping is Lipschitz with respect to spectral norm i.e. for all  $\widetilde{\mathbf{W}}, \mathbf{W} \in \mathbb{R}^{k \times d}$  we have

$$\|\mathcal{J}(\widetilde{\mathbf{W}}) - \mathcal{J}(\mathbf{W})\| \leq B\sqrt{K} \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\| \|\widetilde{\mathbf{W}} - \mathbf{W}\|_F.$$

**Proof** The proof will follow from Lemma 6.9. First, given  $\mathbf{A} = [\mathbf{A}_1^T \dots \mathbf{A}_K^T]^T$  and  $\mathbf{B} = [\mathbf{B}_1^T \dots \mathbf{B}_K^T]^T$ , observe that

$$\|\mathbf{A}\| \leq \sqrt{K} \sup_{1 \leq \ell \leq K} \|\mathbf{A}_\ell\| \quad \text{and} \quad \|\mathbf{A} - \mathbf{B}\| \leq \sqrt{K} \sup_{1 \leq \ell \leq K} \|\mathbf{A}_\ell - \mathbf{B}_\ell\|.$$

These two identities applied to the components  $\mathcal{J}_\ell(\mathbf{W})$  and  $\mathcal{J}_\ell(\widetilde{\mathbf{W}}) - \mathcal{J}_\ell(\mathbf{W})$  completes the proof of the bound on the spectral norm and the perturbation. To prove the bound in (6.31) we use the identity (6.29) to conclude that

$$\begin{aligned} \|\text{mat}(\mathcal{J}^T(\mathbf{W})\mathbf{u})\|_{2,\infty} &= \left\| \sum_{\ell=1}^K \text{diag}(\mathbf{v}_\ell) \phi'(\mathbf{W}\mathbf{X}^T) \text{diag}(\mathbf{u}_\ell) \mathbf{X} \right\|_{2,\infty} \\ &\leq \sum_{\ell=1}^K \|\text{diag}(\mathbf{v}_\ell) \phi'(\mathbf{W}\mathbf{X}^T) \text{diag}(\mathbf{u}_\ell) \mathbf{X}\|_{2,\infty} \\ &\leq \sum_{\ell=1}^K B \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\| \|\mathbf{u}_\ell\|_{\ell_2} \\ &= B \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\| \left( \sum_{\ell=1}^K \|\mathbf{u}_\ell\|_{\ell_2} \right) \\ &\leq B \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\| \sqrt{K} \left( \sum_{\ell=1}^K \|\mathbf{u}_\ell\|_{\ell_2}^2 \right)^{1/2} \\ &= B \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\| \sqrt{K}, \end{aligned}$$

where the penultimate inequality follows from Cauchy Schwarz, completing the proof.  $\blacksquare$

### 6.2.3 Properties of the Jacobian at random initialization

In this section we prove a few lemmas characterizing the properties of the Jacobian at the random initialization.

**Lemma 6.11 (Multiclass covariance)** *Given input and output layer weights  $\mathbf{V}$  and  $\mathbf{W}$ , consider the Jacobian described in (6.27). Given an  $Kn \times Kn$  matrix  $\mathbf{M}$ , for  $1 \leq \ell, \tilde{\ell} \leq K$ , let  $\mathbf{M}[\ell, \tilde{\ell}]$  denote the  $(\ell, \tilde{\ell})$ th submatrix. For  $\mathbf{C}(\mathbf{W}) = \mathcal{J}(\mathbf{W})\mathcal{J}(\mathbf{W})^T$  we have*

$$\mathbf{C}(\mathbf{W})[\ell, \tilde{\ell}] = \sum_{s=1}^k (\mathbf{X}\mathbf{X}^T) \odot (\mathbf{V}_{\ell,s}\mathbf{V}_{\tilde{\ell},s}\phi'(\mathbf{X}\mathbf{w}_s)\phi'(\mathbf{X}\mathbf{w}_s)^T).$$

Suppose  $\mathbf{W} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$  and  $\mathbf{V}$  has i.i.d. zero-mean entries with  $\nu^2$  variance. Then  $\mathbb{E}[\mathbf{C}(\mathbf{W})]$  is a block diagonal matrix given by the Kronecker product

$$\mathbb{E}[\mathbf{C}(\mathbf{W})] = k\nu^2\mathbf{\Sigma}(\mathbf{X}).$$

where  $\mathbf{\Sigma}(\mathbf{X})$  is equal to  $\mathbf{I}_K \otimes [(\mathbf{X}\mathbf{X}^T) \odot \mathbb{E}[\phi'(\mathbf{X}\mathbf{w}_s)\phi'(\mathbf{X}\mathbf{w}_s)^T]]$ .

**Proof** The  $(\ell, \tilde{\ell})$ th submatrix of  $\mathbf{C}(\mathbf{W})$  is given by

$$\begin{aligned} \mathbf{C}(\mathbf{W})[\ell, \tilde{\ell}] &= ((\text{diag}(\mathbf{v}_\ell)\phi'(\mathbf{W}\mathbf{X}^T)) * \mathbf{X}^T)((\text{diag}(\mathbf{v}_{\tilde{\ell}})\phi'(\mathbf{W}\mathbf{X}^T)) * \mathbf{X}^T)^T \\ &= \sum_{s=1}^k \mathcal{J}_\ell(\mathbf{w}_s)\mathcal{J}_{\tilde{\ell}}(\mathbf{w}_s)^T \\ &= \sum_{s=1}^k \mathbf{V}_{\ell,s}\mathbf{V}_{\tilde{\ell},s}(\text{diag}(\phi'(\mathbf{X}\mathbf{w}_s))\mathbf{X})(\text{diag}(\phi'(\mathbf{X}\mathbf{w}_s))\mathbf{X})^T \\ &= \sum_{s=1}^k \mathbf{V}_{\ell,s}\mathbf{V}_{\tilde{\ell},s}(\mathbf{X}\mathbf{X}^T) \odot (\phi'(\mathbf{X}\mathbf{w}_s)\phi'(\mathbf{X}\mathbf{w}_s)^T) \\ &= \sum_{s=1}^k (\mathbf{X}\mathbf{X}^T) \odot (\mathbf{V}_{\ell,s}\mathbf{V}_{\tilde{\ell},s}\phi'(\mathbf{X}\mathbf{w}_s)\phi'(\mathbf{X}\mathbf{w}_s)^T). \end{aligned} \tag{6.32}$$

Setting  $\mathbf{W} \stackrel{i.i.d.}{\sim} \mathcal{N}(0, 1)$  and  $\mathbf{V}$  with i.i.d. zero-mean and  $\nu^2$ -variance entries, we conclude that

$$\begin{aligned} \mathbb{E}[\mathbf{C}(\mathbf{W})[\ell, \tilde{\ell}]] &= \sum_{s=1}^k (\mathbf{X}\mathbf{X}^T) \odot (\mathbb{E}[\mathbf{V}_{\ell,s}\mathbf{V}_{\tilde{\ell},s}]\mathbb{E}[\phi'(\mathbf{X}\mathbf{w}_s)\phi'(\mathbf{X}\mathbf{w}_s)^T]) \\ &= \sum_{s=1}^k \nu^2\delta(\ell - \tilde{\ell})[(\mathbf{X}\mathbf{X}^T) \odot \mathbb{E}[\phi'(\mathbf{X}\mathbf{w}_s)\phi'(\mathbf{X}\mathbf{w}_s)^T]] \\ &= k\delta(\ell - \tilde{\ell})\nu^2\tilde{\mathbf{\Sigma}}(\mathbf{X}), \end{aligned}$$

where  $\delta(x)$  is the discrete  $\delta$  function which is 0 for  $x \neq 0$  and 1 for  $x = 0$  and  $\tilde{\mathbf{\Sigma}}(\mathbf{X})$  is single output kernel matrix which concludes the proof.  $\blacksquare$

Next we state a useful lemma from [52] which allows us to bound the eigenvalues of the Hadamard product of the two PSD matrices.

**Lemma 6.12 ([52])** *Let  $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$  be two Positive Semi-Definite (PSD) matrices. Then,*

$$\begin{aligned} \lambda_{\min}(\mathbf{A} \odot \mathbf{B}) &\geq \left( \min_i B_{ii} \right) \lambda_{\min}(\mathbf{A}), \\ \lambda_{\max}(\mathbf{A} \odot \mathbf{B}) &\leq \left( \max_i B_{ii} \right) \lambda_{\max}(\mathbf{A}). \end{aligned}$$

Next we state a lemma regarding concentration of the Jacobian matrix at initialization.

**Lemma 6.13 (Concentration of the Jacobian at initialization)** *Consider a one-hidden layer neural network model of the form  $\mathbf{x} \mapsto \mathbf{V}\phi(\mathbf{W}\mathbf{x})$  where the activation  $\phi$  obeys  $|\phi(0)| \leq B$  and  $|\phi'(z)| \leq B$  for all  $z$ . Also assume we have  $n \geq K$  data points  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$  with unit euclidean norm ( $\|\mathbf{x}_i\|_{\ell_2} = 1$ ) aggregated as the rows of a matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$ . Furthermore, suppose  $\mathbf{V}$  has i.i.d.  $\nu$ -scaled Rademacher entries (i.e.  $\pm\nu$  equally-likely). Then, the Jacobian matrix at a random point  $\mathbf{W}_0 \in \mathbb{R}^{k \times d}$  with i.i.d.  $\mathcal{N}(0, 1)$  entries obeys*

$$\|\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T - \mathbb{E}[\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T]\| \leq 30K\sqrt{k}\nu^2 B^2 \|\mathbf{X}\|^2 \log(n).$$

with probability at least  $1 - 1/n^{100}$ . In particular, as long as

$$k \geq \frac{1000K^2 B^4 \|\mathbf{X}\|^4 \log(n)}{\delta^2},$$

with the same probability, we have that

$$\left\| \frac{1}{k\nu^2} \mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T - \Sigma(\mathbf{X}) \right\| \leq \delta.$$

**Proof** Define  $\mathbf{C} = \mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T$ . We begin by showing that the diagonal blocks of  $\mathbf{C}$  are concentrated. To do this first for  $1 \leq s \leq k$  define the random matrices

$$\mathbf{A}_s = \left( \phi'(\mathbf{X}\mathbf{w}_s) \phi'(\mathbf{X}\mathbf{w}_s)^T \right) \odot (\mathbf{X}\mathbf{X}^T).$$

Now consider  $n \times n$  diagonal blocks of  $\mathbf{C}$  (denoted by  $\mathbf{C}[\ell, \ell]$ ) and note that we have

$$\begin{aligned} \mathbf{C}[\ell, \ell] &= \left( \phi'(\mathbf{X}\mathbf{W}^T) \text{diag}(\mathbf{v}_\ell) \text{diag}(\mathbf{v}_\ell) \phi'(\mathbf{W}\mathbf{X}^T) \right) \odot (\mathbf{X}\mathbf{X}^T) \\ &= \sum_{s=1}^k \mathbf{V}_{\ell,s}^2 \mathbf{A}_s \\ &= \nu^2 \sum_{s=1}^k \mathbf{A}_s. \end{aligned}$$

Furthermore, using Lemma 6.12

$$\|\mathbf{A}_s\| \leq \left( \max_i \left( \phi'(\mathbf{x}_i^T \mathbf{w}_s) \right)^2 \right) \|\mathbf{X}\|^2 \leq B^2 \|\mathbf{X}\|^2.$$

Also, using Jensen's inequality

$$\|\mathbb{E}[\mathbf{A}_s]\| \leq \mathbb{E} \|\mathbf{A}_s\| \leq B^2 \|\mathbf{X}\|^2.$$

Combining the latter two identities via the triangular inequality we conclude that

$$\|(\mathbf{A}_s - \mathbb{E}[\mathbf{A}_s])^2\| = \|\mathbf{A}_s - \mathbb{E}[\mathbf{A}_s]\|^2 \leq (\|\mathbf{A}_s\| + \|\mathbb{E}[\mathbf{A}_s]\|)^2 \leq (2B^2 \|\mathbf{X}\|^2)^2. \quad (6.33)$$

To proceed, we will bound the weighted sum

$$\mathbf{S} = \sum_{s=1}^k \nu^2 (\mathbf{A}_s - \mathbb{E}[\mathbf{A}_s])$$

in spectral norm. To this aim we utilize the Matrix Hoeffding inequality which states that

$$\mathbb{P}(\|\mathbf{S}\| \geq t) \leq 2ne^{-\frac{t^2}{2\Delta^2}},$$

where  $\Delta^2$  is an upper bound on  $\left\| \sum_{s=1}^k \nu^4 (\mathbf{A}_s - \mathbb{E}[\mathbf{A}_s])^2 \right\|$ . Using (6.33) we can pick  $\Delta^2 = \sum_{s=1}^k (2\nu^2 B^2 \|\mathbf{X}\|^2)^2 = 4k\nu^4 B^4 \|\mathbf{X}\|^4$ . Setting  $t = 30\sqrt{k}\nu^2 B^2 \|\mathbf{X}\|^2 \sqrt{\log(n)}$ , we conclude that

$$\mathbb{P}\left\{ \|C[\ell, \ell] - \mathbb{E}[C[\ell, \ell]]\| \geq t \right\} = \mathbb{P}(\|\mathbf{S}\| \geq t) \leq n^{-102}$$

concluding the proof of concentration of the diagonal blocks of  $\mathbf{C}$ .

For the off-diagonal blocks  $\mathbf{C}[\ell, \tilde{\ell}]$  using (6.32) from the proof of Lemma 6.11 we have that

$$\mathbf{C}[\ell, \tilde{\ell}] = \sum_{s=1}^k \mathbf{V}_{\ell,s} \mathbf{V}_{\tilde{\ell},s}^T \mathbf{A}_s.$$

Note that by construction  $\{\mathbf{V}_{\ell,s} \mathbf{V}_{\tilde{\ell},s}^T\}_{s=1}^k$  are i.i.d.  $\pm\nu^2$  Rademacher variables and thus  $\mathbf{C}[\ell, \tilde{\ell}]$  is sum of zero-mean i.i.d. matrices and we are again in the position to apply Hoeffding's inequality. To this aim note that

$$\left\| \sum_{s=1}^k \mathbf{V}_{\ell,s}^2 \mathbf{V}_{\tilde{\ell},s}^2 \mathbf{A}_s^2 \right\| = \nu^4 \left\| \sum_{s=1}^k \mathbf{A}_s^2 \right\| \leq \nu^4 \sum_{s=1}^k \|\mathbf{A}_s\|^2 \leq \nu^4 k B^4 \|\mathbf{X}\|^4,$$

so that we can take  $\Delta^2 = \nu^4 k B^4 \|\mathbf{X}\|^4$  and again conclude that for  $t = 30\sqrt{k}\nu^2 B^2 \|\mathbf{X}\|^2 \log(n)$  we have

$$\mathbb{P}\left\{ \|\mathbf{C}[\ell, \tilde{\ell}]\| \geq t \right\} \leq n^{-102}$$

Using the fact that  $\mathbb{E}[\mathbf{C}[\ell, \tilde{\ell}]] = 0$  and  $K \leq n$ , combined with a union bound over all sub-matrices  $1 \leq \ell, \tilde{\ell} \leq K$  we conclude that

$$\mathbb{P}\left\{ \|\mathbf{C}[\ell, \tilde{\ell}] - \mathbb{E}[\mathbf{C}[\ell, \tilde{\ell}]]\| \geq t \right\} \leq K^2 n^{-102} \leq n^{-100}.$$

All that remains is to combine the concentration results for the sub-matrices to arrive at the complete bound. In mathematical terms we need to bound  $\mathbf{D} := \|\mathbf{C} - \mathbb{E}[\mathbf{C}]\|$ . To this aim define  $\mathbf{D}[\ell, :]$  to denote the  $\ell$ th block row of  $\mathbf{D}$ . Standard bounds on spectral norm in terms of sub-matrices allow us to conclude that

$$\begin{aligned} \|\mathbf{D}[\ell, :]\| &\leq \sqrt{K} \sup_{1 \leq \tilde{\ell} \leq K} \|\mathbf{D}[\ell, \tilde{\ell}]\| \leq \sqrt{K} t \Rightarrow \\ \|\mathbf{D}\| &\leq \sqrt{K} \sup_{1 \leq \ell \leq K} \|\mathbf{D}[\ell, :]\| \leq \sqrt{K} \sqrt{K} t = K t = 30K \sqrt{k} \nu^2 B^2 \|\mathbf{X}\|^2 \log(n), \end{aligned}$$

concluding the proof. The result in terms of  $\delta$  is obtained by using the population covariance Lemma 6.11. ■

#### 6.2.4 Upper bound on initial residual

In this section we prove a lemma concerning the size of the initial misfit. The proof of this lemma (stated below) follows from a similar argument in the proof of [48, Lemma 6.12].

**Lemma 6.14 (Upper bound on initial residual)** *Consider a one-hidden layer neural network model of the form  $\mathbf{x} \mapsto \mathbf{V} \phi(\mathbf{W} \mathbf{x})$  where the activation  $\phi$  has bounded derivatives obeying  $|\phi(0)|, |\phi'(z)| \leq B$ . Also assume we have  $n$  data points  $\mathbf{x}_1, \mathbf{x}_2, \dots, \mathbf{x}_n \in \mathbb{R}^d$  with unit euclidean norm ( $\|\mathbf{x}_i\|_{\ell_2} = 1$ ) aggregated as rows of a matrix  $\mathbf{X} \in \mathbb{R}^{n \times d}$  and the corresponding labels given by  $\mathbf{y} \in \mathbb{R}^{K^n}$ . Furthermore, assume the entries of  $\mathbf{V}$  are i.i.d. Rademacher variables scaled by  $\frac{\nu \|\mathbf{y}\|_{\ell_2}}{50B\sqrt{K \log(2K)kn}}$  and the entries of  $\mathbf{W} \in \mathbb{R}^{k \times d}$  are i.i.d.  $\mathcal{N}(0, 1)$ . Then,*

$$\|\mathbf{V} \phi(\mathbf{W} \mathbf{X}^T)\|_F \leq \nu \|\mathbf{y}\|_{\ell_2},$$

holds with probability at least  $1 - (2K)^{-100}$ .

**Proof** We begin the proof by noting that

$$\|\mathbf{V}\phi(\mathbf{W}\mathbf{X}^T)\|_F^2 = \sum_{\ell=1}^K \|\mathbf{v}_\ell^T \phi(\mathbf{W}\mathbf{X}^T)\|_{\ell_2}^2.$$

We will show that for any row  $\mathbf{v}$  of  $\mathbf{V}$ , with probability at least  $1 - (2K)^{-101}$ ,

$$\|\mathbf{v}_\ell^T \phi(\mathbf{W}\mathbf{X}^T)\|_{\ell_2} \leq \frac{\nu}{\sqrt{K}} \|\mathbf{y}\|_{\ell_2}. \quad (6.34)$$

so that a simple union bound can conclude the proof. Therefore, all that remains is to show (6.34) holds. To prove the latter, note that for any two matrices  $\widetilde{\mathbf{W}}, \mathbf{W} \in \mathbb{R}^{k \times d}$  we have

$$\begin{aligned} \left| \|\phi(\mathbf{X}\widetilde{\mathbf{W}}^T)\mathbf{v}\|_{\ell_2} - \|\phi(\mathbf{X}\mathbf{W}^T)\mathbf{v}\|_{\ell_2} \right| &\leq \|\phi(\mathbf{X}\widetilde{\mathbf{W}}^T)\mathbf{v} - \phi(\mathbf{X}\mathbf{W}^T)\mathbf{v}\|_{\ell_2} \\ &\leq \|\phi(\mathbf{X}\widetilde{\mathbf{W}}^T) - \phi(\mathbf{X}\mathbf{W}^T)\| \|\mathbf{v}\|_{\ell_2} \\ &\leq \|\phi(\mathbf{X}\widetilde{\mathbf{W}}^T) - \phi(\mathbf{X}\mathbf{W}^T)\|_F \|\mathbf{v}\|_{\ell_2} \\ &\stackrel{(a)}{=} \|(\phi'(\mathbf{S} \odot \mathbf{X}\widetilde{\mathbf{W}}^T + (\mathbf{1}_{k \times n} - \mathbf{S}) \odot \mathbf{X}\mathbf{W}^T)) \odot (\mathbf{X}(\widetilde{\mathbf{W}} - \mathbf{W})^T)\|_F \|\mathbf{v}\|_{\ell_2} \\ &\leq B \|\mathbf{X}(\widetilde{\mathbf{W}} - \mathbf{W})^T\|_F \|\mathbf{v}\|_{\ell_2} \\ &\leq B \|\mathbf{X}\| \|\mathbf{v}\|_{\ell_2} \|\widetilde{\mathbf{W}} - \mathbf{W}\|_F, \end{aligned}$$

where in (a) we used the mean value theorem with  $\mathbf{S}$  a matrix with entries obeying  $0 \leq \mathbf{S}_{i,j} \leq 1$  and  $\mathbf{1}_{k \times n}$  the matrix of all ones. Thus,  $\|\phi(\mathbf{X}\mathbf{W}^T)\mathbf{v}\|_{\ell_2}$  is a  $B\|\mathbf{X}\|\|\mathbf{v}\|_{\ell_2}$ -Lipschitz function of  $\mathbf{W}$ . Thus, fixing  $\mathbf{v}$ , for a matrix  $\mathbf{W}$  with i.i.d. Gaussian entries

$$\|\phi(\mathbf{X}\mathbf{W}^T)\mathbf{v}\|_{\ell_2} \leq \mathbb{E}[\|\phi(\mathbf{X}\mathbf{W}^T)\mathbf{v}\|_{\ell_2}] + t, \quad (6.35)$$

holds with probability at least  $1 - e^{-\frac{t^2}{2B^2\|\mathbf{v}\|_{\ell_2}^2\|\mathbf{X}\|^2}}$ . Next given  $g \sim \mathcal{N}(0, 1)$ , we have

$$|\mathbb{E}[\phi(g)]| \leq |\mathbb{E}[\phi(0)]| + |\mathbb{E}[\phi(g) - \phi(0)]| \leq B + B\mathbb{E}[|g|] \leq 2B \quad \text{and} \quad \text{Var}(\phi(g)) \leq B^2. \quad (6.36)$$

where the latter follows from Poincare inequality (e.g. see [35, p. 49]). Furthermore, since  $\mathbf{v}$  has i.i.d. Rademacher entries, applying Bernstein bound, event

$$E_{\mathbf{v}} := \{|\mathbf{1}^T \mathbf{v}|^2 \leq 250 \log K \|\mathbf{v}\|_{\ell_2}^2\} \quad (6.37)$$

holds with probability  $1 - (2K)^{-102}$ . Conditioned on  $E_{\mathbf{v}}$ , we now upper bound the expectation via

$$\begin{aligned} \mathbb{E}[\|\phi(\mathbf{X}\mathbf{W}^T)\mathbf{v}\|_{\ell_2}] &\stackrel{(a)}{\leq} \sqrt{\mathbb{E}[\|\phi(\mathbf{X}\mathbf{W}^T)\mathbf{v}\|_{\ell_2}^2]} \\ &= \sqrt{\sum_{i=1}^n \mathbb{E}[(\mathbf{v}^T \phi(\mathbf{W}\mathbf{x}_i))^2]} \\ &\stackrel{(b)}{=} \sqrt{n} \sqrt{\mathbb{E}_{g \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)}[(\mathbf{v}^T \phi(g))^2]} \\ &\stackrel{(c)}{=} \sqrt{n} \sqrt{\|\mathbf{v}\|_{\ell_2}^2 \mathbb{E}_{g \sim \mathcal{N}(0,1)}[(\phi(g) - \mathbb{E}[\phi(g)])^2] + (\mathbf{1}^T \mathbf{v})^2 (\mathbb{E}_{g \sim \mathcal{N}(0,1)}[\phi(g)])^2} \\ &\stackrel{(d)}{\leq} \sqrt{n} \|\mathbf{v}\|_{\ell_2} \sqrt{250 \times 4B^2 \log(2K) + B^2} \\ &\leq 32\sqrt{n \log(2K)} B \|\mathbf{v}\|_{\ell_2}. \end{aligned}$$

Here, (a) follows from Jensen's inequality, (b) from linearity of expectation and the fact that for  $\mathbf{x}_i$  with unit Euclidean norm  $\mathbf{W}\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_k)$ , (c) from simple algebraic manipulations, (d) from the inequalities (6.37) and (6.36). Thus using  $t = 18\sqrt{n \log(2K)}B \|\mathbf{v}\|_{\ell_2}$  in (6.35), conditioned on  $E_{\mathbf{v}}$  we conclude that

$$\|\phi(\mathbf{X}\mathbf{W}^T)\mathbf{v}\|_{\ell_2} \leq 50\sqrt{n \log(2K)}B \|\mathbf{v}\|_{\ell_2} = 50\sqrt{n \log(2K)}B\sqrt{k} \frac{\nu \|\mathbf{y}\|_{\ell_2}}{50B\sqrt{K \log(2K)kn}} = \frac{\nu \|\mathbf{y}\|_{\ell_2}}{\sqrt{K}}, \quad (6.38)$$

holds with probability at least  $1 - \exp(-102 \log(2K) \frac{n}{\|\mathbf{X}\|^2}) \geq 1 - (2K)^{-102}$  where we used  $n \geq \|\mathbf{X}\|^2$ . Using a union bound over  $E_{\mathbf{v}}$  and the conditional concentration over  $\mathbf{W}$ , the overall probability of success in (6.38) is at least  $1 - (2K)^{-101}$  concluding the proof of (6.34) and the Lemma.  $\blacksquare$

### 6.3 Rademacher complexity and generalization bounds

In this section we state and prove some Rademacher complexity results that will be used in our generalization bounds. We begin with some basic notation regarding Rademacher complexity. Let  $\mathcal{F}$  be a function class. Suppose  $f \in \mathcal{F}$  maps  $\mathbb{R}^d$  to  $\mathbb{R}^K$ . Let  $\{\varepsilon_i\}_{i=1}^n$  be i.i.d. vectors in  $\mathbb{R}^K$  with i.i.d. Rademacher variables. Given i.i.d. samples  $\mathcal{S} = \{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \sim \mathcal{D}$ , we define the empirical Rademacher complexity to be

$$\mathcal{R}_{\mathcal{S}}(\mathcal{F}) = \frac{1}{n} \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i^T f(\mathbf{x}_i) \right].$$

We begin by stating a vector contraction inequality by Maurer [41]. This is obtained by setting  $h_i(f(\mathbf{x}_i)) = h(\mathbf{y}_i, f(\mathbf{x}_i))$  in Corollary 4 of [41].

**Lemma 6.15** *Let  $f(\cdot) : \mathbb{R}^d \rightarrow \mathbb{R}^K$  and let  $\ell : \mathbb{R}^K \times \mathbb{R}^K \rightarrow \mathbb{R}$  be a 1 Lipschitz loss function with respect to second variable. Let  $\{\varepsilon_i\}_{i=1}^n$  be i.i.d. Rademacher variables. Given i.i.d. samples  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ , define*

$$\mathcal{R}_{\mathcal{S}}(\ell, \mathcal{F}) = \mathbb{E} \left[ \sup_{f \in \mathcal{F}} \sum_{i=1}^n \varepsilon_i \ell(\mathbf{y}_i, f(\mathbf{x}_i)) \right].$$

We have that

$$\mathcal{R}_{\mathcal{S}}(\ell, \mathcal{F}) \leq \sqrt{2} \mathcal{R}_{\mathcal{S}}(\mathcal{F}).$$

Combining the above result with standard generalization bounds based on Rademacher complexity [9] allows us to prove the following result.

**Lemma 6.16** *Let  $\ell(\cdot, \cdot) : \mathbb{R}^K \times \mathbb{R}^K \rightarrow [0, 1]$  be a 1 Lipschitz loss function. Given i.i.d. samples  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$ , consider the empirical loss*

$$\mathcal{L}(f, \ell) = \frac{1}{n} \sum_{i=1}^n \ell(\mathbf{y}_i, f(\mathbf{x}_i)).$$

With probability  $1 - \delta$  over the samples, for all  $f \in \mathcal{F}$ , we have that

$$\mathbb{E}[\mathcal{L}(f, \ell)] \leq \mathcal{L}(f, \ell) + 2\sqrt{2} \mathcal{R}_{\mathcal{S}}(\mathcal{F}) + \sqrt{\frac{5 \log(2/\delta)}{n}}$$

**Proof** Based on [9],

$$\mathbb{E}[\mathcal{L}(f, \ell)] \leq \mathcal{L}(f, \ell) + 2\mathcal{R}_{\mathcal{S}}(\ell, \mathcal{F}) + \sqrt{\frac{5 \log(2/\delta)}{n}}$$

holds with  $1 - \delta$  probability. Combining the latter with Lemma 6.15 completes the proof.  $\blacksquare$



**Lemma 6.17** Consider a neural network model of the form  $\mathbf{x} \mapsto f(\mathbf{x}; \mathbf{V}, \mathbf{W}) = \mathbf{V}\phi(\mathbf{W}\mathbf{x})$  with  $\mathbf{W} \in \mathbb{R}^{k \times d}$  and  $\mathbf{V} \in \mathbb{R}^{K \times k}$  denoting the input and output weight matrices. Suppose  $\mathbf{V}_0 \in \mathbb{R}^{K \times k}$  is a matrix obeying  $\|\mathbf{V}_0\|_{\ell_\infty} \leq \nu/\sqrt{kK}$ . Also let  $\mathbf{W}_0 \in \mathbb{R}^{k \times d}$  be a reference input weight matrix. Furthermore, we define the neural network function space parameterized by the weights as follows

$$\mathcal{F}_{\mathcal{V}, \mathcal{W}} = \left\{ f(\mathbf{x}; \mathbf{V}, \mathbf{W}) \text{ such that } \mathbf{V} \in \mathcal{V} \text{ and } \mathbf{W} \in \mathcal{W} \right\} \text{ with } \mathcal{V} = \left\{ \mathbf{V} : \|\mathbf{V} - \mathbf{V}_0\|_F \leq \frac{\nu M_{\mathcal{V}}}{\sqrt{Kk}} \right\}$$

$$\text{and } \mathcal{W} = \left\{ \mathbf{W} : \|\mathbf{W} - \mathbf{W}_0\|_F \leq M_{\mathcal{W}} \text{ and } \|\mathbf{W} - \mathbf{W}_0\|_{2, \infty} \leq \frac{R}{\sqrt{k}} \right\}. \quad (6.39)$$

Additionally, assume the training data  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$  are generated i.i.d. with the input data points of unit Euclidean norm (i.e.  $\|\mathbf{x}_i\|_{\ell_2} = 1$ ). Also, define the average energy at  $\mathbf{W}_0$  as

$$E = \left( \frac{1}{kn} \sum_{i=1}^n \|\phi(\mathbf{W}_0 \mathbf{x}_i)\|_{\ell_2}^2 \right)^{1/2}.$$

Also let  $\{\boldsymbol{\xi}_i\}_{i=1}^n \in \mathbb{R}^K$  be i.i.d. vectors with i.i.d. Rademacher entries and define the empirical Rademacher complexity

$$\mathcal{R}_{\mathcal{S}}(\mathcal{F}_{\mathcal{V}, \mathcal{W}}) := \frac{1}{n} \mathbb{E} \left[ \sup_{f \in \mathcal{F}_{\mathcal{V}, \mathcal{W}}} \sum_{i=1}^n \boldsymbol{\xi}_i^T f(\mathbf{x}_i) \right].$$

Then,

$$\mathcal{R}_{\mathcal{S}}(\mathcal{F}_{\mathcal{V}, \mathcal{W}}) \leq \nu B \left( \frac{M_{\mathcal{W}} + EM_{\mathcal{V}}}{\sqrt{n}} + \frac{R^2 + M_{\mathcal{W}}M_{\mathcal{V}}}{\sqrt{k}} \right). \quad (6.40)$$

**Proof** We shall use  $\mathbf{w}_\ell$  to denote the rows of  $\mathbf{W}$  (same for  $\mathbf{W}_0, \mathbf{V}, \mathbf{V}_0$ ). We will approximate  $\phi(\langle \mathbf{w}_\ell, \mathbf{x}_i \rangle)$  by its linear approximation  $\phi(\langle \mathbf{w}_\ell^0, \mathbf{x}_i \rangle) + \phi'(\langle \mathbf{w}_\ell^0, \mathbf{x}_i \rangle) (\langle \mathbf{w}_\ell - \mathbf{w}_\ell^0, \mathbf{x}_i \rangle)$  via the second order Taylor's mean value theorem. We thus have

$$\begin{aligned} \mathcal{R}_{\mathcal{S}}(\mathcal{F}_{\mathcal{V}, \mathcal{W}}) &\leq \frac{1}{n} \mathbb{E} \left[ \sum_{i=1}^n \boldsymbol{\xi}_i^T \mathbf{V}_0 \phi(\mathbf{W}_0 \mathbf{x}_i) \right] \\ &\quad + \underbrace{\frac{1}{n} \mathbb{E} \left[ \sup_{\mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \boldsymbol{\xi}_i^T \mathbf{V}_0 \text{diag}(\phi'(\mathbf{W}_0 \mathbf{x}_i)) (\mathbf{W} - \mathbf{W}_0) \mathbf{x}_i \right]}_{\mathcal{R}_1} \\ &\quad + \underbrace{\frac{1}{2n} \mathbb{E}_{\xi_{i,j} \stackrel{\text{i.i.d.}}{\sim} \pm 1} \left[ \sup_{\mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \sum_{\ell=1}^k \sum_{j=1}^K \xi_{i,j}^T \mathbf{V}_{j,\ell}^0 \phi''((1-t_{i\ell})\langle \mathbf{w}_\ell^0, \mathbf{x}_i \rangle + t_{i\ell}\langle \mathbf{w}_\ell, \mathbf{x}_i \rangle) (\langle \mathbf{w}_\ell - \mathbf{w}_\ell^0, \mathbf{x}_i \rangle)^2 \right]}_{\mathcal{R}_2} \\ &\quad + \underbrace{\frac{1}{n} \mathbb{E} \left[ \sup_{\mathbf{V} \in \mathcal{V}, \mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \boldsymbol{\xi}_i^T (\mathbf{V} - \mathbf{V}_0) (\phi(\mathbf{W} \mathbf{x}_i) - \phi(\mathbf{W}_0 \mathbf{x}_i)) \right]}_{\mathcal{R}_3} \\ &\quad + \underbrace{\frac{1}{n} \mathbb{E} \left[ \sup_{\mathbf{V} \in \mathcal{V}} \sum_{i=1}^n \boldsymbol{\xi}_i^T (\mathbf{V} - \mathbf{V}_0) \phi(\mathbf{W}_0 \mathbf{x}_i) \right]}_{\mathcal{R}_4} \end{aligned}$$

We proceed by bounding each of these four terms. For the first term note that

$$\begin{aligned}
\mathcal{R}_1 &\leq \frac{1}{n} \mathbb{E} \left[ \sup_{\|\mathbf{W} - \mathbf{W}_0\|_F \leq M_{\mathcal{W}}} \sum_{i=1}^n \boldsymbol{\xi}_i^T \mathbf{V}_0 \text{diag}(\phi'(\mathbf{W}_0 \mathbf{x}_i)) (\mathbf{W} - \mathbf{W}_0) \mathbf{x}_i \right] \\
&\leq \frac{1}{n} \mathbb{E} \left[ \sup_{\|\mathbf{W} - \mathbf{W}_0\|_F \leq M_{\mathcal{W}}} \left\langle \sum_{i=1}^n \text{diag}(\phi'(\mathbf{W}_0 \mathbf{x}_i)) \mathbf{V}_0^T \boldsymbol{\xi}_i \mathbf{x}_i^T, \mathbf{W} - \mathbf{W}_0 \right\rangle \right] \\
&\leq \frac{M_{\mathcal{W}}}{n} \mathbb{E} \left[ \left\| \left( \sum_{i=1}^n \text{diag}(\mathbf{V}_0^T \boldsymbol{\xi}_i) \phi'(\mathbf{W}_0 \mathbf{x}_i) \mathbf{x}_i^T \right) \right\|_F \right] \\
&\leq \frac{M_{\mathcal{W}}}{n} \mathbb{E} \left[ \left\| \left( \sum_{i=1}^n \text{diag}(\mathbf{V}_0^T \boldsymbol{\xi}_i) \phi'(\mathbf{W}_0 \mathbf{x}_i) \mathbf{x}_i^T \right) \right\|_F^2 \right]^{1/2} \\
&= \frac{M_{\mathcal{W}}}{n} \left[ \sum_{i=1}^n \mathbb{E} \left\| \text{diag}(\mathbf{V}_0^T \boldsymbol{\xi}_i) \phi'(\mathbf{W}_0 \mathbf{x}_i) \mathbf{x}_i^T \right\|_F^2 \right]^{1/2} \\
&= \frac{M_{\mathcal{W}}}{n} \left[ \sum_{i=1}^n \mathbb{E} \left\| \text{diag}(\mathbf{V}_0^T \boldsymbol{\xi}_i) \phi'(\mathbf{W}_0 \mathbf{x}_i) \right\|_{\ell_2}^2 \right]^{1/2} \\
&\leq \frac{BM_{\mathcal{W}}}{n} \left[ \sum_{i=1}^n \mathbb{E} \left\| \mathbf{V}_0^T \boldsymbol{\xi}_i \right\|_{\ell_2}^2 \right]^{1/2} \\
&\leq \frac{BM_{\mathcal{W}}}{n} \|\mathbf{V}_0\|_F \\
&\leq \frac{BM_{\mathcal{W}}\nu}{\sqrt{n}},
\end{aligned}$$

where in the last inequality we used the fact that  $\|\mathbf{V}_0\|_F \leq \nu$ . For the second term note that

$$\begin{aligned}
\mathcal{R}_2 &\leq \frac{1}{2n} \mathbb{E} \left[ \sup_{\|\mathbf{W} - \mathbf{W}_0\|_{2,\infty} \leq R} \sum_{i=1}^n \sum_{\ell=1}^k \sum_{j=1}^K \xi_{i,j} \mathbf{v}_{0,j,\ell} \phi''((1-t_{i\ell})\langle \mathbf{w}_\ell^0, \mathbf{x}_i \rangle + t_{i\ell}\langle \mathbf{w}_\ell, \mathbf{x}_i \rangle) (\langle \mathbf{w}_\ell - \mathbf{w}_\ell^0, \mathbf{x}_i \rangle)^2 \right] \\
&\leq \frac{1}{2n} \sum_{\ell=1}^k \mathbb{E} \left[ \sup_{\|\mathbf{w}_\ell - \mathbf{w}_\ell^0\|_{\ell_2} \leq R} \sum_{i=1}^n \left| \sum_{j=1}^K \xi_{i,j} \mathbf{v}_{0,j,\ell} \right| |\phi''((1-t_{i\ell})\langle \mathbf{w}_\ell^0, \mathbf{x}_i \rangle + t_{i\ell}\langle \mathbf{w}_\ell, \mathbf{x}_i \rangle)| (\langle \mathbf{w}_\ell - \mathbf{w}_\ell^0, \mathbf{x}_i \rangle)^2 \right] \\
&\leq \frac{1}{2kn} \sum_{\ell=1}^k \sum_{i=1}^n \mathbb{E} \left[ \left| \sum_{j=1}^K \xi_{i,j} \mathbf{v}_{0,j,\ell} \right| \right] R^2 B \\
&\leq \frac{BR^2}{2k} \|\mathbf{V}_0^T\|_{2,1} \\
&\leq \frac{BR^2\nu}{2\sqrt{k}}.
\end{aligned}$$

In the above we used  $\|\mathbf{M}\|_{2,1}$  for a matrix  $\mathbf{M}$  to denote the sum of the Euclidean norm of the rows of  $\mathbf{M}$ .

We also used the fact that  $\|\mathbf{V}_0^T\|_{2,1} \leq \nu\sqrt{k}$ . To bound the third term note that

$$\begin{aligned}
\mathcal{R}_3 &= \frac{1}{n} \mathbb{E} \left[ \sup_{\mathbf{V} \in \mathcal{V}, \mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \boldsymbol{\xi}_i^T (\mathbf{V} - \mathbf{V}_0) (\phi(\mathbf{W} \mathbf{x}_i) - \phi(\mathbf{W}_0 \mathbf{x}_i)) \right] \\
&\leq \frac{1}{n} \mathbb{E} \left[ \sup_{\mathbf{V} \in \mathcal{V}, \mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \|\mathbf{V} - \mathbf{V}_0\|_F \|\boldsymbol{\xi}_i\|_{\ell_2} \|\phi(\mathbf{W} \mathbf{x}_i) - \phi(\mathbf{W}_0 \mathbf{x}_i)\|_{\ell_2} \right] \\
&\leq \frac{\nu M_{\mathcal{V}}}{n\sqrt{kK}} \mathbb{E} \left[ \sup_{\mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \|\boldsymbol{\xi}_i\|_{\ell_2} \|\phi(\mathbf{W} \mathbf{x}_i) - \phi(\mathbf{W}_0 \mathbf{x}_i)\|_{\ell_2} \right] \\
&\leq \frac{\nu M_{\mathcal{V}}}{n\sqrt{k}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \|\phi(\mathbf{W} \mathbf{x}_i) - \phi(\mathbf{W}_0 \mathbf{x}_i)\|_{\ell_2} \\
&\leq \frac{\nu B M_{\mathcal{V}}}{n\sqrt{k}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \|(\mathbf{W} - \mathbf{W}_0) \mathbf{x}_i\|_{\ell_2} \\
&\leq \frac{\nu B M_{\mathcal{V}}}{n\sqrt{k}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \sum_{i=1}^n \|(\mathbf{W} - \mathbf{W}_0)\|_F \\
&= \frac{\nu B M_{\mathcal{V}}}{\sqrt{k}} \cdot \sup_{\mathbf{W} \in \mathcal{W}} \|(\mathbf{W} - \mathbf{W}_0)\|_F \\
&= \frac{\nu B M_{\mathcal{V}} M_{\mathcal{W}}}{\sqrt{k}}.
\end{aligned}$$

Finally, to bound the fourth term note that we have

$$\begin{aligned}
\mathcal{R}_4 &= \frac{1}{n} \mathbb{E} \left[ \sup_{\mathbf{V} \in \mathcal{V}} \sum_{i=1}^n \boldsymbol{\xi}_i^T (\mathbf{V} - \mathbf{V}_0) \phi(\mathbf{W}_0 \mathbf{x}_i) \right] \\
&= \frac{1}{n} \mathbb{E} \left[ \sup_{\|\mathbf{V} - \mathbf{V}_0\|_F \leq \frac{\nu M_{\mathcal{V}}}{\sqrt{kK}}} \left\langle \sum_{i=1}^n \boldsymbol{\xi}_i \phi(\mathbf{W}_0 \mathbf{x}_i)^T, (\mathbf{V} - \mathbf{V}_0) \right\rangle \right] \\
&= \frac{\nu M_{\mathcal{V}}}{n\sqrt{kK}} \mathbb{E} \left[ \left\| \sum_{i=1}^n \phi(\mathbf{W}_0 \mathbf{x}_i) \boldsymbol{\xi}_i^T \right\|_F \right] \\
&\leq \frac{\nu M_{\mathcal{V}}}{n\sqrt{kK}} \left( \mathbb{E} \left[ \left\| \sum_{i=1}^n \phi(\mathbf{W}_0 \mathbf{x}_i) \boldsymbol{\xi}_i^T \right\|_F^2 \right] \right)^{1/2} \\
&= \frac{\nu M_{\mathcal{V}}}{n\sqrt{kK}} \left( \sum_{i=1}^n \mathbb{E} \left[ \|\phi(\mathbf{W}_0 \mathbf{x}_i) \boldsymbol{\xi}_i^T\|_F^2 \right] \right)^{1/2} \\
&= \frac{\nu M_{\mathcal{V}}}{\sqrt{n}} \left( \frac{1}{kn} \sum_{i=1}^n \mathbb{E} \left[ \|\phi(\mathbf{W}_0 \mathbf{x}_i)\|_F^2 \right] \right)^{1/2} \\
&= \frac{\nu E M_{\mathcal{V}}}{\sqrt{n}} \\
&\leq \frac{\nu B E M_{\mathcal{V}}}{\sqrt{n}}
\end{aligned}$$

Combining these four bounds we conclude that

$$\mathcal{R}_{\mathcal{S}}(\mathcal{F}_{\mathcal{V}, \mathcal{W}}) \leq \nu B \left( \frac{M_{\mathcal{W}}}{\sqrt{n}} + \frac{R^2}{\sqrt{k}} + \frac{M_{\mathcal{V}} M_{\mathcal{W}}}{\sqrt{k}} + \frac{E M_{\mathcal{V}}}{\sqrt{n}} \right),$$

concluding the proof of (6.40). ■

Next we state a crucial lemma that connects the test error measured by any Lipschitz loss to that of the quadratic loss on the training data.

**Lemma 6.18** Consider a one-hidden layer neural network with input to output mapping of the form  $\mathbf{x} \in \mathbb{R}^d \mapsto f(\mathbf{x}; \mathbf{V}, \mathbf{W}) = \mathbf{V} \phi(\mathbf{W} \mathbf{x}) \in \mathbb{R}^K$  with  $\mathbf{W} \in \mathbb{R}^{k \times d}$  denoting the input-to-hidden weights and  $\mathbf{V} \in \mathbb{R}^{K \times K}$  the hidden-to-output weights. Suppose  $\mathbf{V}_0 \in \mathbb{R}^{K \times K}$  is a matrix obeying  $\|\mathbf{V}_0\|_{\ell_\infty} \leq \nu/\sqrt{kK}$ . Also let  $\mathbf{W}_0 \in \mathbb{R}^{k \times d}$  be a reference input weight matrix. Also define the empirical losses

$$\mathcal{L}(\mathbf{V}, \mathbf{W}) = \frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i - f(\mathbf{x}_i; \mathbf{V}, \mathbf{W})\|_{\ell_2}^2,$$

and

$$\mathcal{L}(f, \ell) = \frac{1}{n} \sum_{i=1}^n \ell(f(\mathbf{x}_i; \mathbf{V}, \mathbf{W}), \mathbf{y}_i),$$

with  $\ell: \mathbb{R}^K \times \mathbb{R}^K \rightarrow [0, 1]$  a one Lipschitz loss function obeying  $\ell(\mathbf{y}, \mathbf{y}) = 0$ . Additionally, assume the training data  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n$  are generated i.i.d. according to a distribution  $\mathcal{D}$  with the input data points of unit Euclidean norm (i.e.  $\|\mathbf{x}_i\|_{\ell_2} = 1$ ). Also, define the average energy at  $\mathbf{W}_0$  as

$$E = \left( \frac{1}{kn} \sum_{i=1}^n \|\phi(\mathbf{W}_0 \mathbf{x}_i)\|_{\ell_2}^2 \right)^{1/2}.$$

Then for all  $f$  in the function class  $\mathcal{F}_{\mathcal{W}}$  given by (6.39)

$$\mathbb{E}[\mathcal{L}(f, \ell)] \leq \sqrt{\mathcal{L}(\mathbf{V}, \mathbf{W})} + 2\sqrt{2}\nu B \left( \frac{M_{\mathcal{W}}}{\sqrt{n}} + \frac{R^2}{\sqrt{k}} + \frac{\nu B M_{\mathcal{V}} M_{\mathcal{W}}}{\sqrt{k}} + \frac{E M_{\mathcal{V}}}{\sqrt{n}} \right) + \sqrt{\frac{5 \log(2/\delta)}{n}}, \quad (6.41)$$

holds with probability at least  $1 - \delta$ . Furthermore, Suppose labels are one-hot encoded and thus unit Euclidean norm. Given a sample  $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^K$  generated according to the distribution  $\mathcal{D}$ , define the population classification error

$$\text{Err}_{\mathcal{D}}(\mathbf{W}) = \mathbb{P}(\arg \max_{1 \leq \ell \leq K} \mathbf{y}_i \neq \arg \max_{1 \leq \ell \leq K} f_i(\mathbf{x}, \mathbf{W})).$$

Then, we also have

$$\text{Err}_{\mathcal{D}}(\mathbf{W}) \leq 2 \left[ \sqrt{\mathcal{L}(\mathbf{V}, \mathbf{W})} + 2\sqrt{2}\nu B \left( \frac{M_{\mathcal{W}}}{\sqrt{n}} + \frac{R^2}{\sqrt{k}} + \frac{\nu B M_{\mathcal{V}} M_{\mathcal{W}}}{\sqrt{k}} + \frac{E M_{\mathcal{V}}}{\sqrt{n}} \right) + \sqrt{\frac{5 \log(2/\delta)}{n}} \right]. \quad (6.42)$$

**Proof** To begin first note that any 1-Lipschitz  $\ell$  with  $\ell(\mathbf{y}, \mathbf{y}) = 0$  obeys  $\ell(\mathbf{y}, \hat{\mathbf{y}}) \leq \|\mathbf{y} - \hat{\mathbf{y}}\|_{\ell_2}$ . Thus, we have

$$\mathcal{L}(f, \ell) \leq \frac{1}{n} \sum_{i=1}^n \|\mathbf{y}_i - f(\mathbf{x}_i; \mathbf{V}, \mathbf{W})\|_{\ell_2} \leq \sqrt{\mathcal{L}(\mathbf{V}, \mathbf{W})},$$

where the last inequality follows from Cauchy-Schwarz. Consequently, applying Lemmas 6.16 and 6.17 we conclude that

$$\begin{aligned} \mathbb{E}[\mathcal{L}(f, \ell)] &\leq \mathcal{L}(f, \ell) + 2\sqrt{2} \cdot \mathcal{R}_{\mathcal{S}}(\mathcal{F}) + \sqrt{\frac{5 \log(2/\delta)}{n}}, \\ &\leq \sqrt{\mathcal{L}(\mathbf{W})} + 2\sqrt{2}\nu B \left( \frac{M_{\mathcal{W}}}{\sqrt{n}} + \frac{R^2}{\sqrt{k}} + \frac{\nu B M_{\mathcal{V}} M_{\mathcal{W}}}{\sqrt{k}} + \frac{E M_{\mathcal{V}}}{\sqrt{n}} \right) + \sqrt{\frac{5 \log(2/\delta)}{n}}, \end{aligned}$$

which yields the first statement.

To prove the second statement on classification accuracy, we pick the  $\ell$  function as follows

$$\ell(\mathbf{y}, \hat{\mathbf{y}}) = \min(1, \|\mathbf{y} - \hat{\mathbf{y}}\|_{\ell_2}).$$

Note that, given a sample  $(\mathbf{x}, \mathbf{y}) \in \mathbb{R}^d \times \mathbb{R}^K$  with one-hot encoded labels, if

$$\arg \max_{1 \leq \ell \leq K} \mathbf{y}_\ell \neq \arg \max_{1 \leq \ell \leq K} f_\ell(\mathbf{x}; \mathbf{V}, \mathbf{W}),$$

this implies

$$\ell(\mathbf{y}, f(\mathbf{x}; \mathbf{V}, \mathbf{W})) \geq 0.5.$$

Combining the latter with Markov inequality we arrive at

$$\text{Err}_{\mathcal{D}}(\mathbf{W}) \leq 2 \mathbb{E}_{(\mathbf{x}, \mathbf{y}) \sim \mathcal{D}}[\ell(\mathbf{y}, f(\mathbf{x}; \mathbf{V}, \mathbf{W}))] = 2 \mathbb{E}[\mathcal{L}(\ell, \mathbf{W})].$$

Now since  $\ell$  is 1 Lipschitz and bounded, it obeys (6.41), which combined with the above identity yields (6.42), completing the proof.  $\blacksquare$

## 6.4 Proofs for neural nets with arbitrary initialization (Proof of Theorem 3.3)

In this section we prove Theorem 3.3. We first discuss a preliminary optimization result in Section 6.4.1. Next, in Section 6.4.2 we build upon this result to prove our main optimization result. Finally, in Section 6.4.3 we use these optimization results to prove our main generalization result, completing the proof of Theorem 3.3.

### 6.4.1 Preliminary Optimization Result

**Lemma 6.19 (Deterministic convergence guarantee)** *Consider a one-hidden layer neural net of the form  $\mathbf{x} \mapsto f(\mathbf{x}; \mathbf{W}) := \mathbf{V}\phi(\mathbf{W}\mathbf{x})$  with input weights  $\mathbf{W} \in \mathbb{R}^{k \times d}$  and output weights  $\mathbf{V} \in \mathbb{R}^{K \times k}$  and an activation  $\phi$  obeying  $|\phi(0)| \leq B$ ,  $|\phi'(z)| \leq B$ , and  $|\phi''(z)| \leq B$  for all  $z$ . Also assume  $\mathbf{V}$  is fixed with all entries bounded by  $\|\mathbf{V}\|_{\ell_\infty} \leq \frac{\nu}{\sqrt{kK}}$  and we train over  $\mathbf{W}$  based on the loss*

$$\mathcal{L}(\mathbf{W}) = \frac{1}{2} \sum_{i=1}^n \|f(\mathbf{x}_i; \mathbf{W}) - \mathbf{y}_i\|_{\ell_2}^2.$$

*Also, consider a point  $\mathbf{W}_0 \in \mathbb{R}^{k \times d}$  with  $\mathbf{J}$  an  $(\epsilon_0, \nu B \|\mathbf{X}\|)$  reference Jacobian associated with  $\mathcal{J}(\mathbf{W}_0)$  per Definition 5.1. Furthermore, define the information  $\mathcal{I}$  and nuisance  $\mathcal{N}$  subspaces and the truncated Jacobian  $\mathbf{J}_{\mathcal{I}}$  associated with the reference Jacobian  $\mathbf{J}$  based on a cut-off spectrum value of  $\alpha$  per Definition 5.2. Let the initial residual vector be  $\mathbf{r}_0 = \mathbf{y} - f(\mathbf{W}_0) \in \mathbb{R}^{nK}$ . Furthermore, assume*

$$\epsilon_0 \leq \frac{\alpha}{5} \min\left(\delta, \sqrt{\frac{\delta\alpha}{\Gamma\nu B \|\mathbf{X}\|}}\right) \quad (6.43)$$

and

$$k \geq 400 \frac{\nu^6 B^6 \|\mathbf{X}\|^6 \Gamma^2}{\delta^2 \alpha^8} (\mathcal{B}_{\alpha, \Gamma} + \delta \Gamma \|\mathbf{r}_0\|_{\ell_2})^2, \quad (6.44)$$

*with  $0 \leq \delta \leq 1$  and  $\Gamma \geq 1$ . We run gradient descent iterations of the form  $\mathbf{W}_{\tau+1} = \mathbf{W}_\tau - \eta \nabla \mathcal{L}(\mathbf{W}_\tau)$  starting from  $\mathbf{W}_0$  with step size  $\eta$  obeying  $\eta \leq \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ . Then for all iterates  $\tau$  obeying  $0 \leq \tau \leq T := \frac{\Gamma}{\eta \alpha^2}$*

$$\|\mathbf{W}_\tau - \mathbf{W}_0\|_F \leq \frac{\mathcal{B}_{\alpha, \Gamma}}{\alpha} + \delta \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2}. \quad (6.45)$$

$$\|\mathbf{W}_\tau - \mathbf{W}_0\|_{2, \infty} \leq \frac{2\nu B \Gamma \|\mathbf{X}\|}{\sqrt{k} \alpha^2} \|\mathbf{r}_0\|_{\ell_2}. \quad (6.46)$$

*Furthermore, after  $\tau = T$  iteration we have*

$$\|\mathbf{r}_T\|_{\ell_2} \leq e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \frac{\delta\alpha}{\nu B \|\mathbf{X}\|} \|\mathbf{r}_0\|_{\ell_2}. \quad (6.47)$$

**Proof** To prove this lemma we wish to apply Theorem 5.3. We thus need to ensure that the assumptions of this theorem are satisfied. To do this note that by Lemma 6.10 Assumption 1 holds with  $\beta = \nu B \|\mathbf{X}\|$ . Furthermore, we pick  $\varepsilon = \frac{\delta \alpha^3}{5\Gamma \nu^2 B^2 \|\mathbf{X}\|^2} = \frac{\delta \alpha^3}{5\Gamma \beta^2}$  which together with (6.43) guarantees (5.11) holds. We now turn our attention to verifying Assumption 2. To this aim note that for all  $\mathbf{W} \in \mathbb{R}^{k \times d}$  obeying

$$\|\mathbf{W} - \mathbf{W}_0\|_F \leq R := 2 \left( \frac{\mathcal{B}_{\alpha, \Gamma}}{\alpha} + \delta \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \right)$$

as long as (6.44) holds by Lemma 6.10 we have

$$\begin{aligned} \|\mathcal{J}(\mathbf{W}) - \mathcal{J}(\mathbf{W}_0)\| &\leq B\sqrt{K} \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\| R \\ &\leq \frac{\nu}{\sqrt{k}} B \|\mathbf{X}\| R \\ &= \frac{\delta \alpha^3}{10\Gamma \nu^2 B^2 \|\mathbf{X}\|^2} \frac{\frac{20\Gamma \nu^3 B^3 \|\mathbf{X}\|^3}{\delta \alpha^4} (\mathcal{B}_{\alpha, \Gamma} + \delta \Gamma \|\mathbf{r}_0\|_{\ell_2})}{\sqrt{k}} \\ &\leq \frac{\delta \alpha^3}{10\Gamma \nu^2 B^2 \|\mathbf{X}\|^2} \\ &= \frac{\varepsilon}{2}. \end{aligned}$$

Thus, Assumption 2 holds with  $\|\mathbf{W} - \mathbf{W}_0\|_F \leq R := 2 \left( \frac{\mathcal{B}_{\alpha, \Gamma}}{\alpha} + \delta \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \right)$ . Now that we have verified that the assumptions of Theorem 5.3 hold so do its conclusions and thus (6.45) and (6.47) hold.

We now turn our attention to proving the row-wise bound (6.46). To this aim let  $\mathbf{w}_\ell^{(\tau)}$  denote the  $\ell$ th row of  $\mathbf{W}_\tau$ . Also note that

$$\nabla \mathcal{L}(\mathbf{w}_\ell) = \ell\text{th row of } \text{mat}(\mathcal{J}(\mathbf{W})^T \mathbf{r}_\tau).$$

Hence, using Lemma 6.10 equation (6.31) we conclude that

$$\|\nabla \mathcal{L}(\mathbf{w}_\ell^{(\tau)})\|_{\ell_2} \leq B\sqrt{K} \|\mathbf{V}\|_{\ell_\infty} \|\mathbf{X}\| \|\mathbf{r}_\tau\|_{\ell_2} \leq \frac{\nu B \|\mathbf{X}\|}{\sqrt{k}} \|\mathbf{r}_\tau\|_{\ell_2}.$$

Consequently, for any row  $1 \leq \ell \leq k$ , we have

$$\|\mathbf{w}_\ell^{(\tau)} - \mathbf{w}_\ell^{(0)}\|_{\ell_2} \leq \eta \frac{\nu B \|\mathbf{X}\|}{\sqrt{k}} \sum_{t=0}^{\tau-1} \|\mathbf{r}_t\|_{\ell_2}. \quad (6.48)$$

To bound the right-hand side we use the triangular inequality combined with (6.25) and (6.26) to conclude that

$$\begin{aligned} \eta \sum_{t=0}^{\tau-1} \|\mathbf{r}_t\|_{\ell_2} &\leq \eta \sum_{t=0}^{\tau-1} \|\tilde{\mathbf{r}}_t\|_{\ell_2} + \eta \sum_{t=0}^{\tau-1} \|\mathbf{r}_t - \tilde{\mathbf{r}}_t\|_{\ell_2} \\ &\leq \frac{\Gamma}{\alpha^2} \|\mathbf{r}_0\|_{\ell_2} + \frac{2\Gamma^2(\varepsilon\beta + \varepsilon_0^2)}{\alpha^4} \|\mathbf{r}_0\|_{\ell_2} \\ &= \frac{2\Gamma(\varepsilon_0^2 + \varepsilon\beta) + \alpha^2}{\alpha^4} \Gamma \|\mathbf{r}_0\|_{\ell_2} \\ &\leq 2 \frac{\Gamma}{\alpha^2} \|\mathbf{r}_0\|_{\ell_2}, \end{aligned} \quad (6.49)$$

where in the last inequality we used the fact that  $\varepsilon_0^2 \leq \frac{\alpha^2}{25\Gamma}$  per (6.43) and  $\varepsilon\beta = \frac{\delta \alpha^3}{5\Gamma \beta} \leq \frac{\alpha^2}{5\Gamma}$  per our choice of  $\varepsilon$ . Combining (6.48) and (6.49), we obtain

$$\|\mathbf{w}_\ell^{(\tau)} - \mathbf{w}_\ell^{(0)}\|_{\ell_2} \leq \frac{2\nu B \|\mathbf{X}\| \Gamma}{\sqrt{k} \alpha^2} \|\mathbf{r}_0\|_{\ell_2},$$

completing the proof of (6.46) and the theorem. ■

### 6.4.2 Main Optimization Result

**Lemma 6.20 (Deterministic optimization guarantee)** *Consider the setting and assumptions of Lemma 6.19. Also assume  $\|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} \geq c\|\mathbf{r}_0\|_{\ell_2}$  for a constant  $c > 0$  if  $\varepsilon_0 > 0$ . Furthermore, assume*

$$\varepsilon_0^2 \leq \frac{\alpha^2}{25} \min\left(c \frac{\mathcal{B}_{\alpha,\Gamma}\alpha}{\nu B \Gamma^2 \|\mathbf{r}_0\|_{\ell_2} \|\mathbf{X}\|}, \frac{\zeta^2 \nu^2 B^2 \|\mathbf{X}\|^2}{\alpha^2}, \frac{\zeta}{\Gamma}\right), \quad (6.50)$$

and

$$k \geq 1600 \left( \frac{\alpha}{\zeta \nu B \|\mathbf{X}\|} + \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\mathcal{B}_{\alpha,\Gamma}} \right)^2 \frac{\nu^6 B^6 \|\mathbf{X}\|^6 \Gamma^2 \mathcal{B}_{\alpha,\Gamma}^2}{\alpha^8}, \quad (6.51)$$

and  $\Gamma \geq 1$ . We run gradient descent iterations of the form  $\mathbf{W}_{\tau+1} = \mathbf{W}_{\tau} - \eta \nabla \mathcal{L}(\mathbf{W}_{\tau})$  starting from  $\mathbf{W}_0$  with step size  $\eta$  obeying  $\eta \leq \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ . Then for all iterates  $\tau$  obeying  $0 \leq \tau \leq T := \frac{\Gamma}{\eta \alpha^2}$

$$\|\mathbf{W}_{\tau} - \mathbf{W}_0\|_F \leq \frac{2\mathcal{B}_{\alpha,\Gamma}}{\alpha}. \quad (6.52)$$

$$\|\mathbf{W}_{\tau} - \mathbf{W}_0\|_{2,\infty} \leq \frac{2\nu B \Gamma \|\mathbf{X}\|}{\sqrt{k} \alpha^2} \|\mathbf{r}_0\|_{\ell_2}. \quad (6.53)$$

Furthermore, after  $\tau = T$  iteration we have

$$\|f(\mathbf{W}_T) - \mathbf{y}\|_{\ell_2} \leq e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \zeta \|\mathbf{r}_0\|_{\ell_2}. \quad (6.54)$$

**Proof** To prove this lemma we aim to substitute

$$\delta = \min\left(\frac{\zeta \nu B \|\mathbf{X}\|}{\alpha}, \frac{\mathcal{B}_{\alpha,\Gamma}}{\Gamma \|\mathbf{r}_0\|_{\ell_2}}\right) \leq 1, \quad (6.55)$$

in Theorem 6.19. To do this we need to verify the assumptions of Theorem 6.19. To this aim note that the choice of  $\delta$  from (6.55) combined with (6.51) ensures that

$$\begin{aligned} k &\geq 1600 \left( \frac{\alpha}{\zeta \nu B \|\mathbf{X}\|} + \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\mathcal{B}_{\alpha,\Gamma}} \right)^2 \frac{\nu^6 B^6 \|\mathbf{X}\|^6 \Gamma^2 \mathcal{B}_{\alpha,\Gamma}^2}{\alpha^8} \\ &\geq \max\left(\frac{\alpha}{\zeta \nu B \|\mathbf{X}\|}, \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\mathcal{B}_{\alpha,\Gamma}}\right)^2 \frac{1600 \nu^6 B^6 \|\mathbf{X}\|^6 \Gamma^2 \mathcal{B}_{\alpha,\Gamma}^2}{\alpha^8} \\ &\geq \frac{1}{\min\left(\frac{\zeta \nu B \|\mathbf{X}\|}{\alpha}, \frac{\mathcal{B}_{\alpha,\Gamma}}{\Gamma \|\mathbf{r}_0\|_{\ell_2}}\right)^2} \frac{1600 \nu^6 B^6 \|\mathbf{X}\|^6 \Gamma^2 \mathcal{B}_{\alpha,\Gamma}^2}{\alpha^8} \\ &= \frac{1600 \Gamma^2 \nu^6 B^6 \|\mathbf{X}\|^6 \mathcal{B}_{\alpha,\Gamma}^2}{\delta^2 \alpha^8} \\ &= 400 \frac{\Gamma^2 \nu^6 B^6 \|\mathbf{X}\|^6 (\mathcal{B}_{\alpha,\Gamma} + \mathcal{B}_{\alpha,\Gamma})^2}{\delta^2 \alpha^8} \\ &\geq 400 \frac{\Gamma^2 \nu^6 B^6 \|\mathbf{X}\|^6 (\mathcal{B}_{\alpha,\Gamma} + \delta \Gamma \|\mathbf{r}_0\|_{\ell_2})^2}{\delta^2 \alpha^8}, \end{aligned}$$

so that (6.44) holds. We thus turn our attention to proving (6.43). If  $\varepsilon_0 = 0$ , the statement already holds. Otherwise, note that based on Lemma 6.2 equation (6.3) we have

$$\mathcal{B}_{\alpha,\Gamma} \geq \frac{\alpha \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2}}{\lambda_1} \geq \frac{\alpha \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2}}{\nu B \|\mathbf{X}\|} \geq \frac{c \alpha \|\mathbf{r}_0\|_{\ell_2}}{\nu B \|\mathbf{X}\|} \Rightarrow \frac{\mathcal{B}_{\alpha,\Gamma}}{c \|\mathbf{r}_0\|_{\ell_2}} \geq \frac{\alpha}{\nu B \|\mathbf{X}\|}. \quad (6.56)$$

Recall that  $\alpha = \frac{\nu \alpha_0}{\sqrt{K}}$ , which implies that

- If  $\delta = \frac{\zeta \nu B \|\mathbf{X}\|}{\alpha}$ : For (6.43) to hold it suffices to have  $\varepsilon_0 \leq \frac{\alpha}{5} \min\left(\frac{\zeta \nu B \|\mathbf{X}\|}{\alpha}, \sqrt{\frac{\zeta}{\Gamma}}\right)$ .
- If  $\delta = \frac{\mathcal{B}_{\alpha, \Gamma}}{\Gamma \|\mathbf{r}_0\|_{\ell_2}}$ : For (6.43) to hold it suffices to have  $\varepsilon_0 \leq \frac{\alpha}{5} \sqrt{c \frac{\mathcal{B}_{\alpha, \Gamma} \alpha}{\nu B \Gamma^2 \|\mathbf{r}_0\|_{\ell_2} \|\mathbf{X}\|}}$  as based on (6.56) we have  $\sqrt{\frac{c\alpha}{\beta\Gamma}} = \sqrt{\frac{c\alpha}{\nu B \Gamma \|\mathbf{X}\|}} \leq \sqrt{\delta}$

$$\begin{aligned}
\varepsilon_0 &\leq \frac{\alpha}{5} \sqrt{\frac{c \mathcal{B}_{\alpha, \Gamma} \alpha}{\nu B \Gamma^2 \|\mathbf{r}_0\|_{\ell_2} \|\mathbf{X}\|}} \\
&= \frac{\alpha}{5} \sqrt{\delta} \sqrt{\frac{c\alpha}{\Gamma\beta}} \\
&= \frac{\alpha}{5} \sqrt{\delta} \cdot \min\left(\sqrt{\delta}, \sqrt{\frac{c\alpha}{\Gamma\beta}}\right) \\
&= \frac{\alpha}{5} \min\left(\delta, \sqrt{\frac{\delta\alpha}{\Gamma\beta}}\right)
\end{aligned}$$

Combining the latter two cases as long as

$$\varepsilon_0^2 \leq \frac{\alpha^2}{25} \min\left(c \frac{\mathcal{B}_{\alpha, \Gamma} \alpha}{\nu B \Gamma^2 \|\mathbf{r}_0\|_{\ell_2} \|\mathbf{X}\|}, \frac{\zeta^2 \nu^2 B^2 \|\mathbf{X}\|^2}{\alpha^2}, \frac{\zeta}{\Gamma}\right) \Leftrightarrow (6.50),$$

then (6.43) holds. As a result when (6.50) and (6.51) hold with  $\delta = \min\left(\frac{\zeta \nu B \|\mathbf{X}\|}{\alpha}, \frac{\mathcal{B}_{\alpha, \Gamma}}{\Gamma \|\mathbf{r}_0\|_{\ell_2}}\right)$  both assumptions of Theorem 6.19 also hold and so do its conclusions. In particular, (6.52) follows from (6.45) by noting that based on our choice of  $\delta$  we have  $\delta \frac{\Gamma}{\alpha} \|\mathbf{r}_0\|_{\ell_2} \leq \frac{\mathcal{B}_{\alpha, \Gamma}}{\alpha}$ , (6.53) follows immediately from (6.46), and (6.54) follows from (6.47) by noting that based on our choice of  $\delta$  we have  $\frac{\delta \alpha}{\nu B \|\mathbf{X}\|} \leq \zeta$ . ■

#### 6.4.3 Main generalization result (completing the proof of Theorem 3.3)

Theorem 3.3 immediately follows from Theorem 6.21 below by upper bounding  $\mathcal{D}_{\alpha, \Gamma}$  (see Definition 6.1) using Lemma 6.2 equation (6.2).

**Theorem 6.21** Consider a training data set  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^K$  generated i.i.d. according to a distribution  $\mathcal{D}$  where the input samples have unit Euclidean norm. Also consider a neural net with  $k$  hidden layers as described in (1.1) parameterized by  $\mathbf{W}$  where the activation function  $\phi$  obeys  $|\phi'(z)|, |\phi''(z)| \leq B$ . Let  $\mathbf{W}_0$  be the initial weight matrix with i.i.d.  $\mathcal{N}(0, 1)$  entries. Also assume the output matrix has bounded entries obeying  $\|\mathbf{V}\|_{\ell_\infty} \leq \frac{\nu}{\sqrt{kK}}$ . Furthermore, set  $\mathbf{J} := \mathcal{J}(\mathbf{W}_0)$  and define the information  $\mathcal{I}$  and nuisance  $\mathcal{N}$  subspaces and the truncated Jacobian  $\mathbf{J}_{\mathcal{I}}$  associated with the reference/initial Jacobian  $\mathbf{J}$  based on a cut-off spectrum value  $\alpha = \nu B \bar{\alpha}^4 \sqrt{n} \sqrt{\|\mathbf{X}\|}$ . Also define the initial residual  $\mathbf{r}_0 = f(\mathbf{W}_0) - \mathbf{y} \in \mathbb{R}^{nK}$  and pick  $C_r > 0$  so that  $\frac{\|\mathbf{r}_0\|_{\ell_2}}{\sqrt{n}} \leq C_r$ . Also assume, the number of hidden nodes  $k$  obeys

$$k \geq 25600 \frac{C_r^2 \Gamma^4}{\bar{\alpha}^8 \nu^2 B^2 \zeta^2}, \tag{6.57}$$

with  $\Gamma \geq 1$  and tolerance level  $\zeta \leq 2$ . Run gradient descent updates (1.5) with learning rate  $\eta \leq \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ . Then, after  $T = \frac{\Gamma}{\eta \alpha^2}$  iterations, with probability at least  $1 - \delta$ , the generalization error obeys

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \leq \underbrace{\frac{2\|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2}}{\sqrt{n}}}_{\text{bias term}} + \underbrace{\frac{12\nu B \mathcal{D}_{\alpha, \Gamma}}{\sqrt{n}}}_{\text{variance term}} + 5\sqrt{\frac{\log(2/\delta)}{n}} + 2C_r(e^{-\Gamma} + \zeta), \tag{6.58}$$

where  $\mathcal{D}_{\alpha, \Gamma}$  is the early stopping distance as in Def. (6.1).



**Proof** First, note that using  $\alpha \leq \beta = \nu B \|\mathbf{X}\|$ ,  $\frac{\mathcal{B}_{\alpha,\Gamma}}{\Gamma} \leq \|\mathbf{r}_0\|_{\ell_2}$  per (6.2), and  $\|\mathbf{r}_0\|_{\ell_2} \leq C_r \sqrt{n}$  we have

$$\frac{\alpha}{\nu B \|\mathbf{X}\|} \leq 1 \leq \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\mathcal{B}_{\alpha,\Gamma}} \leq C_r \frac{\Gamma \sqrt{n}}{\mathcal{B}_{\alpha,\Gamma}}.$$

This together with  $\zeta \leq 2$  implies that

$$\begin{aligned} \frac{\alpha}{\frac{\zeta}{2} \nu B \|\mathbf{X}\|} + \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\mathcal{B}_{\alpha,\Gamma}} &\leq \frac{\alpha}{\frac{\zeta}{2} \nu B \|\mathbf{X}\|} + \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\frac{\zeta}{2} \mathcal{B}_{\alpha,\Gamma}} \\ &\leq 2 \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\frac{\zeta}{2} \mathcal{B}_{\alpha,\Gamma}} \\ &\leq 2 \frac{C_r}{\frac{\zeta}{2}} \frac{\Gamma}{\mathcal{B}_{\alpha,\Gamma}} \sqrt{n}. \end{aligned} \tag{6.59}$$

Thus when

$$\begin{aligned} k &\geq 25600 \frac{C_r^2 \Gamma^4}{\bar{\alpha}^8 \nu^2 B^2 \zeta^2} \\ &= 6400 \left( \frac{C_r \Gamma}{\frac{\zeta}{2}} \right)^2 \frac{n^2 \Gamma^2 \nu^6 B^6 \|\mathbf{X}\|^4}{\alpha^8} \\ &\stackrel{\sqrt{n} \geq \|\mathbf{X}\|}{\geq} 6400 \left( \frac{C_r \Gamma}{\frac{\zeta}{2}} \right)^2 \frac{n \nu^6 B^6 \|\mathbf{X}\|^6}{\alpha^8} \\ &= 1600 \left( 2 \frac{C_r}{\frac{\zeta}{2}} \frac{\Gamma}{\mathcal{B}_{\alpha,\Gamma}} \sqrt{n} \right)^2 \frac{\mathcal{B}_{\alpha,\Gamma}^2 \nu^6 B^6 \|\mathbf{X}\|^6}{\alpha^8} \\ &\stackrel{(6.59)}{\geq} 1600 \left( \frac{\alpha}{\frac{\zeta}{2} \nu B \|\mathbf{X}\|} + \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{\mathcal{B}_{\alpha,\Gamma}} \right)^2 \frac{\mathcal{B}_{\alpha,\Gamma}^2 \nu^6 B^6 \|\mathbf{X}\|^6}{\alpha^8} \end{aligned}$$

Thus, (6.51) holds. Also (6.50) trivially holds for  $\varepsilon_0$ . Thus applying Theorem 6.20 with  $\varepsilon_0 = 0$  the following three conclusions hold

$$\|\mathbf{W}_\tau - \mathbf{W}_0\|_F \leq \frac{2\mathcal{B}_{\alpha,\Gamma}}{\alpha} = 2\mathcal{D}_{\alpha,\Gamma}. \tag{6.60}$$

and

$$\begin{aligned} \|\mathbf{W}_\tau - \mathbf{W}_0\|_{2,\infty} &\leq \frac{2\nu B \Gamma \|\mathbf{X}\|}{\sqrt{k} \alpha^2} \|\mathbf{r}_0\|_{\ell_2} \\ &\stackrel{\|\mathbf{r}_0\|_{\ell_2} \leq C_r \sqrt{n}}{\leq} \frac{2\sqrt{n} C_r \nu B \Gamma \|\mathbf{X}\|}{\sqrt{k} \alpha^2}. \end{aligned} \tag{6.61}$$

and

$$\begin{aligned} \|f(\mathbf{W}_T) - \mathbf{y}\|_{\ell_2} &\leq e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \frac{\zeta}{2} \|\mathbf{r}_0\|_{\ell_2} \\ &\leq \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + \left( e^{-\Gamma} + \frac{\zeta}{2} \right) \|\mathbf{r}_0\|_{\ell_2} \\ &\stackrel{\|\mathbf{r}_0\|_{\ell_2} \leq C_r \sqrt{n}}{\leq} \|\Pi_{\mathcal{N}}(\mathbf{r}_0)\|_{\ell_2} + C_r \left( e^{-\Gamma} + \frac{\zeta}{2} \right) \sqrt{n}. \end{aligned} \tag{6.62}$$

Furthermore, using the assumption that  $\|\mathbf{V}\|_{\ell_\infty} \leq \frac{\nu}{\sqrt{kK}}$  Lemma 6.18 applies and hence equation (6.42) with  $\mathbf{W} = \mathbf{W}_T$ ,  $\sqrt{\mathcal{L}(\mathbf{W}_T)} = \frac{\|f(\mathbf{W}_T) - \mathbf{y}\|_{\ell_2}}{\sqrt{n}}$ ,  $M_{\mathcal{W}} = 2\mathcal{D}_{\alpha, \Gamma}$ ,  $M_{\mathcal{Y}} = 0$ , and  $R = \frac{2\sqrt{n}C_r\nu B\Gamma\|\mathbf{X}\|}{\alpha^2}$  implies that

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \leq 2 \left[ \frac{\|f(\mathbf{W}_T) - \mathbf{y}\|_{\ell_2}}{\sqrt{n}} + 3\nu B \left( \frac{2\mathcal{D}_{\alpha, \Gamma}}{\sqrt{n}} + \frac{R^2}{\sqrt{k}} \right) + \sqrt{\frac{5 \log(2/\delta)}{n}} \right] \quad (6.63)$$

Also note that using (6.57) we have

$$\frac{3\nu B R^2}{\sqrt{k}} \leq \frac{12C_r^2\Gamma^2\nu^3 B^3 n \|\mathbf{X}\|^2}{\sqrt{k}\alpha^4} \leq C_r \frac{\zeta}{2} \quad (6.64)$$

Plugging (6.62) and (6.64) into (6.63) completes the proof.  $\blacksquare$

## 6.5 Proofs for neural network with random initialization (proof of Theorem 3.2)

In this section we prove Theorem 3.2. We first discuss and prove an optimization result in Section 6.5.1. Next, in Section 6.5.2 we build upon this result to complete the proof of Theorem 3.2.

### 6.5.1 Optimization result

**Theorem 6.22 (Optimization guarantee for random initialization)** *Consider a training data set  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^K$  generated i.i.d. according to a distribution  $\mathcal{D}$  where the input samples have unit Euclidean norm and the concatenated label vector obeys  $\|\mathbf{y}\|_{\ell_2} = \sqrt{n}$  (e.g. one-hot encoding). Consider a neural net with  $k$  hidden layers as described in (1.1) parameterized by  $\mathbf{W}$  where the activation function  $\phi$  obeys  $|\phi'(z)|, |\phi''(z)| \leq B$ . Let  $\mathbf{W}_0$  be the initial weight matrix with i.i.d.  $\mathcal{N}(0, 1)$  entries. Fix a precision level  $\zeta$  and set*

$$\nu = \frac{\zeta}{50B\sqrt{\log(2K)}}. \quad (6.65)$$

Also assume the output layer  $\mathbf{V}$  has i.i.d. Rademacher entries scaled by  $\frac{\nu}{\sqrt{kK}}$ . Furthermore, set  $\mathbf{J} := \Sigma(\mathbf{X})^{1/2}$  and define the information  $\mathcal{I}$  and nuisance  $\mathcal{N}$  spaces and the truncated Jacobian  $\mathbf{J}_{\mathcal{I}}$  associated with the reference Jacobian  $\mathbf{J}$  based on a cut-off spectrum value of  $\alpha_0 = \bar{\alpha}\sqrt{n}\sqrt{K}\|\mathbf{X}\|B \leq B\sqrt{K}\|\mathbf{X}\|$  per Definition 2.1 so as to ensure  $\|\Pi_{\mathcal{I}}(\mathbf{y})\|_{\ell_2} \geq c\|\mathbf{y}\|_{\ell_2}$  for some constant  $c$ . Assume

$$k \geq 12 \times 10^7 \frac{\Gamma^4 K^4 B^8 \|\mathbf{X}\|^6 n \log(n)}{c^4 \zeta^4 \alpha_0^8} \quad (6.66)$$

with  $\Gamma \geq 1$  and  $\zeta \leq \frac{c}{2}$ . We run gradient descent iterations of the form (1.5) with a learning rate  $\eta \leq \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ . Then, after  $T = \frac{\Gamma K}{\eta \nu^2 \alpha_0^2}$  iterations, the following identities

$$\|f(\mathbf{W}_{\tau_0}) - \mathbf{y}\|_{\ell_2} \leq \|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2} + e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{y})\|_{\ell_2} + 4\zeta\sqrt{n}, \quad (6.67)$$

$$\|\mathbf{W}_{\tau} - \mathbf{W}_0\|_{\ell_2} \leq \frac{2\sqrt{K}(\mathcal{B}_{\alpha_0, \Gamma}(\mathbf{y}) + \Gamma\zeta\sqrt{n})}{\nu\alpha_0}, \quad (6.68)$$

$$\|\mathbf{W}_{\tau} - \mathbf{W}_0\|_{2, \infty} \leq \frac{4\Gamma B K \|\mathbf{X}\|}{\nu\alpha_0^2} \sqrt{\frac{n}{k}}, \quad (6.69)$$

hold with probability at least  $1 - (2K)^{-100}$ .

**Proof** To prove this result we wish to apply Theorem 6.20. To do this we need to verify the assumptions of this theorem. To start with, using Lemma 6.14 with probability at least  $1 - (2K)^{-100}$ , the initial prediction vector  $f(\mathbf{W}_0)$  obeys

$$\|f(\mathbf{W}_0)\|_{\ell_2} \leq \zeta \|\mathbf{y}\|_{\ell_2} = \zeta \sqrt{n} \leq \frac{\sqrt{n}}{2}. \quad (6.70)$$

Hence the initial residual obeys  $\|\mathbf{r}_0\|_{\ell_2} \leq 2\sqrt{n}$ . Furthermore, using  $\zeta \leq c/2$

$$\|\mathbf{r}_0 + \mathbf{y}\|_{\ell_2} \leq \zeta \|\mathbf{y}\|_{\ell_2} \implies \|\Pi_{\mathcal{I}}(\mathbf{r}_0 + \mathbf{y})\|_{\ell_2} \leq \zeta \|\mathbf{y}\|_{\ell_2}. \quad (6.71)$$

Thus,

$$\begin{aligned} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} &\geq \|\Pi_{\mathcal{I}}(\mathbf{y})\|_{\ell_2} - \|\Pi_{\mathcal{I}}(\mathbf{r}_0 + \mathbf{y})\|_{\ell_2} \\ &\geq \|\Pi_{\mathcal{I}}(\mathbf{y})\|_{\ell_2} - \zeta \|\mathbf{y}\|_{\ell_2} \\ &\geq (c - \zeta) \|\mathbf{y}\|_{\ell_2} \\ &\geq \frac{c}{2} \|\mathbf{y}\|_{\ell_2} \end{aligned} \quad (6.72)$$

$$\geq \frac{c}{4} \|\mathbf{r}_0\|_{\ell_2}. \quad (6.73)$$

Thus the assumption on the ratio of information to total energy of residual holds and we can replace  $c$  with  $\frac{c}{4}$  in Theorem 6.20. Furthermore, since  $\mathcal{B}_{\alpha_0, \Gamma}(\cdot)$  is  $\Gamma$ -Lipschitz function of its input vector in  $\ell_2$  norm hence we also have

$$\mathcal{B}_{\alpha_0, \Gamma}(\mathbf{r}_0) \leq \mathcal{B}_{\alpha_0, \Gamma}(\mathbf{y}) + \Gamma \|\mathbf{r}_0 + \mathbf{y}\|_{\ell_2} \leq \mathcal{B}_{\alpha_0, \Gamma}(\mathbf{y}) + \Gamma \zeta \|\mathbf{y}\|_{\ell_2}. \quad (6.74)$$

Next we wish to show that (6.50) holds. In particular we will show that there exists an  $\varepsilon_0$ -reference Jacobian  $\mathbf{J}$  for  $\mathcal{J}(\mathbf{W}_0)$  satisfying  $\mathbf{J}\mathbf{J}^T = \mathbb{E}[\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T]$ . Note that, such a  $\mathbf{J}$  will have exactly same information/nuisance spaces as the square-root of the multiclass kernel matrix i.e.  $(\mathbb{E}[\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T])^{\frac{1}{2}}$  since these subspaces are governed by the left eigenvectors. Applying Lemmas 6.13 (with a scaling of the Jacobian by  $1/\sqrt{kK}$  due to the different scaling of  $\mathbf{V}$ ), we find that if

$$k \geq \frac{1000K^2B^4\|\mathbf{X}\|^4 \log(n)}{\delta^2} \quad (6.75)$$

then,

$$\|\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T - \mathbb{E}[\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T]\| \leq \frac{\delta\nu^2}{K}. \quad (6.76)$$

Let  $\bar{\mathcal{J}}(\mathbf{W})$  be obtained by adding  $\max(Kn - p, 0)$  zero columns to  $\mathcal{J}(\mathbf{W})$ . Then, using (6.76) and Lemma 6.4, there exists  $\mathbf{J}$  satisfying  $\mathbf{J}\mathbf{J}^T = \mathbb{E}[\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T]$  and

$$\|\bar{\mathcal{J}}(\mathbf{W}_0) - \mathbf{J}\| \leq 2\sqrt{\frac{\delta\nu^2}{K}}.$$

Therefore,  $\mathbf{J}$  is an  $\varepsilon_0^2 = 4\frac{\delta\nu^2}{K}$  reference Jacobian. Now set

$$\Theta = \min\left(c \frac{\mathcal{B}_{\alpha_0, \Gamma}\alpha_0}{B\Gamma^2 \|\mathbf{X}\| \sqrt{nK}}, \left(\frac{\zeta B\sqrt{K}\|\mathbf{X}\|}{\alpha_0}\right)^2, \frac{\zeta}{\Gamma}\right)$$

and note that using  $\alpha = \frac{\nu}{\sqrt{K}}\alpha_0$  and  $\|\mathbf{r}_0\|_{\ell_2} \leq 2\sqrt{n}$

$$\begin{aligned}
\Theta &= \min \left( c \frac{\mathcal{B}_{\alpha_0, \Gamma} \alpha_0}{B \Gamma^2 \|\mathbf{X}\| \sqrt{nK}}, \left( \frac{\zeta B \sqrt{K} \|\mathbf{X}\|}{\alpha_0} \right)^2, \frac{\zeta}{\Gamma} \right) \\
&= \min \left( c \frac{\mathcal{B}_{\alpha_0, \Gamma} \alpha}{\nu B \Gamma^2 \|\mathbf{X}\| \sqrt{n}}, \left( \frac{\nu \zeta B \|\mathbf{X}\|}{\alpha} \right)^2, \frac{\zeta}{\Gamma} \right) \\
&\leq 2 \cdot \min \left( c \frac{\mathcal{B}_{\alpha, \Gamma} \alpha}{\nu B \Gamma^2 \|\mathbf{r}_0\|_{\ell_2} \|\mathbf{X}\|}, \frac{\zeta^2 \nu^2 B^2 \|\mathbf{X}\|^2}{\alpha^2}, \frac{\zeta}{\Gamma} \right)
\end{aligned} \tag{6.77}$$

To continue further, note that  $\mathcal{B}_{\alpha_0, \Gamma}$  calculated with respect to  $\Sigma(\mathbf{X})^{1/2}$  with cutoff  $\alpha_0$  is exactly same as  $\mathcal{B}_{\alpha, \Gamma}$  calculated with respect to  $\mathbf{J}$  with cutoff  $\alpha = \frac{\nu \alpha_0}{\sqrt{K}}$  which is a square-root of  $\mathbb{E}[\mathcal{J}(\mathbf{W}_0)\mathcal{J}(\mathbf{W}_0)^T]$ . Thus, using (6.77) to ensure (6.50) holds it suffices to show

$$\varepsilon_0^2 = 4 \frac{\delta \nu^2}{K} \leq \frac{\alpha^2}{25} \frac{\Theta}{2} = \frac{\nu^2 \alpha_0^2}{50K} \Theta.$$

Hence, to ensure (6.50) holds we need to ensure that  $\delta$  obeys

$$\delta \leq \frac{\alpha_0^2}{200} \Theta.$$

Thus using  $\delta = \frac{\alpha_0^2}{200} \Theta$  to ensure (6.50) we need to make sure  $k$  is sufficiently large so that (6.75) holds with this value of  $\delta$ . Thus it suffices to have

$$\begin{aligned}
k &\geq 12 \times 10^7 \frac{\Gamma^4 K^4 B^8 \|\mathbf{X}\|^6 n \log(n)}{c^4 \zeta^4 \alpha_0^8} \\
&\geq 12 \times 10^7 \frac{\Gamma^4 K^4 B^8 \|\mathbf{X}\|^8 \log(n)}{c^4 \zeta^4 \alpha_0^8} \\
&\geq 4 \times 10^7 \cdot \left( \frac{4K^2 B^4 \Gamma^4 \|\mathbf{X}\|^4}{c^4 \alpha_0^4} + \frac{1}{\zeta^4} + \frac{\Gamma^2}{\zeta^2} \right) \frac{K^2 B^4 \|\mathbf{X}\|^4 \log(n)}{\alpha_0^4} \\
&\geq 4 \times 10^7 \cdot \left( \frac{4K B^4 \Gamma^4 \|\mathbf{X}\|^4}{c^4 \alpha_0^4} + \frac{1}{\zeta^4} + \frac{\Gamma^2}{\zeta^2} \right) \frac{K^2 B^4 \|\mathbf{X}\|^4 \log(n)}{\alpha_0^4} \\
&\geq 4 \times 10^7 \cdot \left( \frac{4K B^4 \Gamma^4 \|\mathbf{X}\|^4}{c^4 \alpha_0^4} + \frac{\alpha_0^4}{\zeta^4 B^4 K^2 \|\mathbf{X}\|^4} + \frac{\Gamma^2}{\zeta^2} \right) \frac{K^2 B^4 \|\mathbf{X}\|^4 \log(n)}{\alpha_0^4} \\
&\stackrel{(a)}{\geq} 4 \times 10^7 \cdot \left( \frac{n K B^2 \Gamma^4 \|\mathbf{X}\|^2}{c^2 \mathcal{B}_{\alpha_0, \Gamma}^2 \alpha_0^2} + \frac{\alpha_0^4}{\zeta^4 B^4 K^2 \|\mathbf{X}\|^4} + \frac{\Gamma^2}{\zeta^2} \right) \frac{K^2 B^4 \|\mathbf{X}\|^4 \log(n)}{\alpha_0^4} \\
&\geq 4 \times 10^7 \cdot \max \left( \frac{n K B^2 \Gamma^4 \|\mathbf{X}\|^2}{c^2 \mathcal{B}_{\alpha_0, \Gamma}^2 \alpha_0^2}, \frac{\alpha_0^4}{\zeta^4 B^4 K^2 \|\mathbf{X}\|^4}, \frac{\Gamma^2}{\zeta^2} \right) \frac{K^2 B^4 \|\mathbf{X}\|^4 \log(n)}{\alpha_0^4} \\
&= 4 \times 10^7 \frac{K^2 B^4 \|\mathbf{X}\|^4 \log(n)}{\alpha_0^4 \cdot \min \left( \left( \frac{c \mathcal{B}_{\alpha_0, \Gamma} \alpha_0}{B \Gamma^2 \|\mathbf{X}\| \sqrt{nK}} \right)^2, \left( \frac{\zeta B \sqrt{K} \|\mathbf{X}\|}{\alpha_0} \right)^4, \frac{\zeta^2}{\Gamma^2} \right)} \\
&= \frac{1000 K^2 B^4 \|\mathbf{X}\|^4 \log(n)}{\delta^2}
\end{aligned} \tag{6.79}$$

Here, (a) follows from the fact that  $\|\Sigma(\mathbf{X})^{1/2}\| := \lambda_1 \leq B \|\mathbf{X}\|$ , equation (6.3), and  $\|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} \geq \frac{c}{2} \|\mathbf{y}\|_{\ell_2} = \frac{c}{2} \sqrt{n}$  which combined imply

$$\mathcal{B}_{\alpha_0, \Gamma} \geq \frac{\alpha}{\lambda_1} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} \geq \frac{\alpha}{B \|\mathbf{X}\|} \|\Pi_{\mathcal{I}}(\mathbf{r}_0)\|_{\ell_2} \geq \frac{\alpha_0 c}{2} \frac{\|\mathbf{y}\|_{\ell_2}}{B \|\mathbf{X}\|} = \frac{\alpha_0 c}{2} \frac{\sqrt{n}}{B \|\mathbf{X}\|}. \tag{6.80}$$

To be able to apply Theorem 6.20 we must also ensure (6.51) holds. Therefore, it suffices to have

$$k \geq 64 \times 10^6 \frac{K^4 B^8 \|\mathbf{X}\|^6 \Gamma^4 n \log(n)}{\zeta^4 \alpha_0^8} \quad (6.81)$$

$$\stackrel{(a)}{\geq} 25600 \frac{K^4 B^6 \|\mathbf{X}\|^6 \Gamma^4 n}{\zeta^2 \nu^2 \alpha_0^8} \quad (6.82)$$

$$\begin{aligned} k &\geq 12800 \left( \frac{\alpha_0^2}{\zeta^2 B^2 K \|\mathbf{X}\|^2} + 1 \right) \frac{K^4 B^6 \|\mathbf{X}\|^6 \Gamma^4 n}{\nu^2 \alpha_0^8} \\ &\stackrel{(b)}{\geq} 3200 \left( \frac{\alpha_0^2}{\zeta^2 B^2 K \|\mathbf{X}\|^2} + \frac{4\Gamma^2 n}{B_{\alpha,\Gamma}^2} \right) \frac{K^4 B^6 \|\mathbf{X}\|^6 \Gamma^2 \mathcal{B}_{\alpha,\Gamma}^2}{\nu^2 \alpha_0^8} \\ &\geq 3200 \left( \frac{\alpha_0^2}{\zeta^2 B^2 K \|\mathbf{X}\|^2} + \frac{\Gamma^2 \|\mathbf{r}_0\|_{\ell_2}^2}{B_{\alpha,\Gamma}^2} \right) \frac{K^4 B^6 \|\mathbf{X}\|^6 \Gamma^2 \mathcal{B}_{\alpha,\Gamma}^2}{\nu^2 \alpha_0^8} \\ &\geq 1600 \left( \frac{\alpha_0}{\zeta \sqrt{K} B \|\mathbf{X}\|} + \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{B_{\alpha,\Gamma}} \right)^2 \frac{K^4 B^6 \|\mathbf{X}\|^6 \Gamma^2 \mathcal{B}_{\alpha,\Gamma}^2}{\nu^2 \alpha_0^8} \\ &= 1600 \left( \frac{\alpha}{\zeta \nu B \|\mathbf{X}\|} + \frac{\Gamma \|\mathbf{r}_0\|_{\ell_2}}{B_{\alpha,\Gamma}} \right)^2 \frac{\nu^6 B^6 \|\mathbf{X}\|^6 \Gamma^2 \mathcal{B}_{\alpha,\Gamma}^2}{\alpha^8} \end{aligned} \quad (6.83)$$

Here, (a) follows from the fact that  $n \geq K$  and the relationship between  $\zeta$  and  $\nu$  per (6.85) and (b) follows from the fact that per equation (6.2) we have

$$\mathcal{B}_{\alpha,\Gamma} \leq \Gamma \|\mathbf{r}_0\|_{\ell_2} \leq 2\Gamma \sqrt{n}$$

Note that (6.78) and (6.82) are implied by

$$k \geq 12 \times 10^7 \frac{\Gamma^4 K^4 B^8 \|\mathbf{X}\|^6 n \log(n)}{c^4 \zeta^4 \alpha_0^8}, \quad (6.84)$$

which is the same as (6.66). What remains is stating the optimization bounds in terms of the labels  $\mathbf{y}$ . This follows by substituting (6.70), (6.74), and the fact that  $\|\mathbf{r}_0\|_{\ell_2} \leq 2\sqrt{n}$  into (6.54), (6.52), and (6.53), respectively. ■

### 6.5.2 Generalization result (completing the proof of Theorem 3.2)

Theorem below is a restatement of Theorem 3.2 after substituting the upper bound on the early stopping distance  $\mathcal{D}_{\alpha_0,\Gamma}$  of Def. (6.1).

**Theorem 6.23 (Neural Net – Generalization)** *Consider a training data set  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^K$  generated i.i.d. according to a distribution  $\mathcal{D}$  where the input samples have unit Euclidean norm and the concatenated label vector obeys  $\|\mathbf{y}\|_{\ell_2} = \sqrt{n}$  (e.g. one-hot encoding). Consider a neural net with  $k$  hidden layers as described in (1.1) parameterized by  $\mathbf{W}$  where the activation function  $\phi$  obeys  $|\phi'(z)|, |\phi''(z)| \leq B$ . Let  $\mathbf{W}_0$  be the initial weight matrix with i.i.d.  $\mathcal{N}(0,1)$  entries. Fix a precision level  $\zeta \leq \frac{\epsilon}{2}$  and set*

$$\nu = \frac{\zeta}{50B\sqrt{\log(2K)}}. \quad (6.85)$$

*Also assume the output layer  $\mathbf{V}$  has i.i.d. Rademacher entries scaled by  $\frac{\nu}{\sqrt{kK}}$ . Furthermore, set  $\mathbf{J} := \Sigma(\mathbf{X})^{1/2}$  and define the information  $\mathcal{I}$  and nuisance  $\mathcal{N}$  spaces and the truncated Jacobian  $\mathbf{J}_{\mathcal{I}}$  associated with the Jacobian  $\mathbf{J}$  based on a cut-off spectrum value of  $\alpha_0 = \bar{\alpha} \sqrt[4]{n} \sqrt{K} \|\mathbf{X}\| B \leq B \|\mathbf{X}\|$  per Definition 2.1 chosen to ensure  $\|\Pi_{\mathcal{I}}(\mathbf{y})\|_{\ell_2} \geq c \|\mathbf{y}\|_{\ell_2}$  for some constant  $c > 0$ . Assume*

$$k \geq 12 \times 10^7 \frac{\Gamma^4 K^4 B^8 \|\mathbf{X}\|^4 n^2 \log(n)}{c^4 \zeta^4 \alpha_0^8} \quad (6.86)$$

with  $\Gamma \geq 1$ . We run gradient descent iterations of the form (1.5) with a learning rate  $\eta \leq \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ . Then, after  $T = \frac{\Gamma K}{\eta \nu^2 \alpha_0^2}$  iterations, classification error  $\text{Err}_{\mathcal{D}}(\mathbf{W}_T)$  is upper bounded by

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \leq 2 \frac{\|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2} + e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{y})\|_{\ell_2}}{\sqrt{n}} + \frac{12B\sqrt{K}}{\sqrt{n}} \mathcal{D}_{\alpha_0, \Gamma} + 12 \left(1 + \frac{\Gamma}{\bar{\alpha}^4 \sqrt{n} \|\mathbf{X}\|^2}\right) \zeta + 10 \sqrt{\frac{\log(2/\delta)}{n}},$$

holds with probability at least  $1 - (2K)^{-100} - \delta$ .

**Proof** Under the stated assumptions, Theorem 6.22 holds with probability  $1 - (2K)^{-100}$ . The proof will condition on outcomes of the Theorem 6.22. Specifically, we shall apply (6.42) of Lemma 6.18 with  $M_{\mathcal{W}}$  and  $R$  dictated by Theorem 6.22 where the output layer  $\mathbf{V}$  is fixed. Observe that  $\|\mathbf{V}\|_F = \sqrt{Kk} \|\mathbf{V}\|_{\ell_\infty} = \nu$ , we have

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \leq 2 \left[ \frac{\|f(\mathbf{W}_T) - \mathbf{y}\|_{\ell_2}}{\sqrt{n}} + 3\nu B \left( \frac{M_{\mathcal{W}}}{\sqrt{n}} + \frac{R^2}{\sqrt{k}} \right) + \sqrt{\frac{5 \log(2/\delta)}{n}} \right]. \quad (6.87)$$

Theorem 6.22 yields

$$\frac{\|f(\mathbf{W}_T) - \mathbf{y}\|_{\ell_2}}{\sqrt{n}} \leq \frac{\|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2} + e^{-\Gamma} \|\Pi_{\mathcal{I}}(\mathbf{y})\|_{\ell_2}}{\sqrt{n}} + 4\zeta. \quad (6.88)$$

Using (6.68) for  $M_{\mathcal{W}}$

$$\frac{\nu B M_{\mathcal{W}}}{\sqrt{n}} \leq \frac{2B\sqrt{K} \mathcal{D}_{\alpha_0, \Gamma}(\mathbf{y})}{\sqrt{n}} + \frac{2B\sqrt{K} \Gamma \zeta}{\alpha_0}. \quad (6.89)$$

Using (6.69) on row bound  $R$  and lower bound on  $k$

$$\begin{aligned} 3\nu B \frac{R^2}{\sqrt{k}} &= \frac{48n\Gamma^2 B^3 K^2 \|\mathbf{X}\|^2}{\nu \alpha_0^4 \sqrt{k}} \\ &\leq \frac{c^2 \zeta^2}{230\nu B \log(n)} \\ &\leq \zeta. \end{aligned} \quad (6.90)$$

Plugging in (6.88), (6.89), and (6.90) into (6.87) concludes the proof.  $\blacksquare$

## Acknowledgements

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

## References

- [1] ALLEN-ZHU, Z., LI, Y., AND LIANG, Y. Learning and generalization in overparameterized neural networks, going beyond two layers. *arXiv preprint arXiv:1811.04918* (2018).

- [2] ALLEN-ZHU, Z., LI, Y., AND SONG, Z. A convergence theory for deep learning via over-parameterization. *arXiv preprint arXiv:1811.03962* (2018).
- [3] ANTHONY, M., AND BARTLETT, P. L. *Neural network learning: Theoretical foundations*. cambridge university press, 2009.
- [4] ARORA, S., DU, S. S., HU, W., LI, Z., AND WANG, R. Fine-grained analysis of optimization and generalization for overparameterized two-layer neural networks. *arXiv preprint arXiv:1901.08584* (2019).
- [5] ARORA, S., GE, R., NEYSHABUR, B., AND ZHANG, Y. Stronger generalization bounds for deep nets via a compression approach. *arXiv preprint arXiv:1802.05296* (2018).
- [6] BARTLETT, P., FOSTER, D. J., AND TELGARSKY, M. Spectrally-normalized margin bounds for neural networks.
- [7] BARTLETT, P. L. The sample complexity of pattern classification with neural networks: the size of the weights is more important than the size of the network. *IEEE transactions on Information Theory* 44, 2 (1998), 525–536.
- [8] BARTLETT, P. L., MAIOROV, V., AND MEIR, R. Almost linear vc dimension bounds for piecewise polynomial networks. In *Advances in Neural Information Processing Systems* (1999), pp. 190–196.
- [9] BARTLETT, P. L., AND MENDELSON, S. Rademacher and gaussian complexities: Risk bounds and structural results. *Journal of Machine Learning Research* 3, Nov (2002), 463–482.
- [10] BELKIN, M., HSU, D., MA, S., AND MANDAL, S. Reconciling modern machine learning and the bias-variance trade-off. *arXiv preprint arXiv:1812.11118* (2018).
- [11] BELKIN, M., HSU, D., AND MITRA, P. Overfitting or perfect fitting? risk bounds for classification and regression rules that interpolate.
- [12] BELKIN, M., HSU, D., AND XU, J. Two models of double descent for weak features. *arXiv preprint arXiv:1903.07571* (2019).
- [13] BELKIN, M., RAKHLIN, A., AND TSYBAKOV, A. B. Does data interpolation contradict statistical optimality?
- [14] BRUTZKUS, A., GLOBERSON, A., MALACH, E., AND SHALEV-SHWARTZ, S. Sgd learns over-parameterized networks that provably generalize on linearly separable data. *arXiv preprint arXiv:1710.10174* (2017).
- [15] CAO, Y., AND GU, Q. A generalization theory of gradient descent for learning over-parameterized deep relu networks. *arXiv preprint arXiv:1902.01384* (2019).
- [16] CHAUDHARI, P., CHOROMANSKA, A., SOATTO, S., LECUN, Y., BALDASSI, C., BORGS, C., CHAYES, J., SAGUN, L., AND ZECCHINA, R. Entropy-sgd: Biasing gradient descent into wide valleys. *arXiv preprint arXiv:1611.01838* (2016).
- [17] CHIZAT, L., AND BACH, F. A note on lazy training in supervised differentiable programming. *arXiv preprint arXiv:1812.07956* (2018).
- [18] CHIZAT, L., AND BACH, F. On the global convergence of gradient descent for over-parameterized models using optimal transport. *arXiv preprint arXiv:1805.09545* (2018).
- [19] DOU, X., AND LIANG, T. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. *arXiv preprint arXiv:1901.07114* (2019).

- [20] DU, S. S., LEE, J. D., LI, H., WANG, L., AND ZHAI, X. Gradient descent finds global minima of deep neural networks. *arXiv preprint arXiv:1811.03804* (2018).
- [21] DU, S. S., ZHAI, X., PO CZOS, B., AND SINGH, A. Gradient descent provably optimizes over-parameterized neural networks. *arXiv preprint arXiv:1810.02054* (2018).
- [22] DZIUGAITE, G. K., AND ROY, D. M. Computing nonvacuous generalization bounds for deep (stochastic) neural networks with many more parameters than training data. *arXiv preprint arXiv:1703.11008* (2017).
- [23] GHORBANI, B., MEI, S., MISIAKIEWICZ, T., AND MONTANARI, A. Linearized two-layers neural networks in high dimension. *arXiv preprint arXiv:1904.12191* (2019).
- [24] GOEL, S., AND KLIVANS, A. Learning neural networks with two nonlinear layers in polynomial time. *arXiv preprint arXiv:1709.06010* (2017).
- [25] GOEL, S., KLIVANS, A., AND MEKA, R. Learning one convolutional layer with overlapping patches. *arXiv preprint arXiv:1802.02547* (2018).
- [26] GOLOWICH, N., RAKHLIN, A., AND SHAMIR, O. Size-independent sample complexity of neural networks. *arXiv preprint arXiv:1712.06541* (2017).
- [27] GUNASEKAR, S., LEE, J. D., SOUDRY, D., AND SREBRO, N. Implicit bias of gradient descent on linear convolutional networks. In *Advances in Neural Information Processing Systems* (2018), pp. 9461–9471.
- [28] HARDT, M., RECHT, B., AND SINGER, Y. Train faster, generalize better: Stability of stochastic gradient descent. *arXiv preprint arXiv:1509.01240* (2015).
- [29] HOFFER, E., HUBARA, I., AND SOUDRY, D. Train longer, generalize better: closing the generalization gap in large batch training of neural networks. In *Advances in Neural Information Processing Systems* (2017), pp. 1731–1741.
- [30] HUANG, W. R., EMAM, Z., GOLDBLUM, M., FOWL, L., TERRY, J. K., HUANG, F., AND GOLDSTEIN, T. Understanding generalization through visualizations.
- [31] JACOT, A., GABRIEL, F., AND HONGLER, C. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems* (2018), pp. 8571–8580.
- [32] JI, Z., AND TELGARSKY, M. Gradient descent aligns the layers of deep linear networks. *arXiv preprint arXiv:1810.02032* (2018).
- [33] JI, Z., AND TELGARSKY, M. Risk and parameter convergence of logistic regression. *arXiv preprint arXiv:1803.07300* (2018).
- [34] KESKAR, N. S., MUDIGERE, D., NOCEDAL, J., SMELYANSKIY, M., AND TANG, P. T. P. On large-batch training for deep learning: Generalization gap and sharp minima. *arXiv preprint arXiv:1609.04836* (2016).
- [35] LEDOUX, M. The concentration of measure phenomenon. *volume 89 of Mathematical Surveys and Monographs. American Mathematical Society, Providence, RI* (2001).
- [36] LI, H., XU, Z., TAYLOR, G., STUDER, C., AND GOLDSTEIN, T. Visualizing the loss landscape of neural nets. In *Advances in Neural Information Processing Systems* (2018), pp. 6389–6399.
- [37] LI, M., SOLTANOLKOTABI, M., AND OYMAK, S. Gradient descent with early stopping is provably robust to label noise for overparameterized neural networks. *arXiv preprint arXiv:1903.11680* (2019).
- [38] LIANG, T., AND RAKHLIN, A. Just interpolate: Kernel "ridgeless" regression can generalize.



- [39] LONG, P. M., AND SEDGHI, H. Size-free generalization bounds for convolutional neural networks. *arXiv preprint arXiv:1905.12600* (2019).
- [40] MA, C., WU, L., ET AL. A comparative analysis of the optimization and generalization property of two-layer neural network and random feature models under gradient descent dynamics. *arXiv preprint arXiv:1904.04326* (2019).
- [41] MAURER, A. A vector-contraction inequality for rademacher complexities. In *International Conference on Algorithmic Learning Theory* (2016), Springer, pp. 3–17.
- [42] MEI, S., MONTANARI, A., AND NGUYEN, P.-M. A mean field view of the landscape of two-layers neural networks. *arXiv preprint arXiv:1804.06561* (2018).
- [43] NAGARAJAN, V., AND KOLTER, J. Z. Deterministic pac-bayesian generalization bounds for deep networks via generalizing noise-resilience. *arXiv preprint arXiv:1905.13344* (2019).
- [44] NEYSHABUR, B., BHOJANAPALLI, S., MCALLESTER, D., AND SREBRO, N. Exploring generalization in deep learning. In *Advances in Neural Information Processing Systems* (2017), pp. 5947–5956.
- [45] NEYSHABUR, B., BHOJANAPALLI, S., MCALLESTER, D., AND SREBRO, N. A pac-bayesian approach to spectrally-normalized margin bounds for neural networks. *arXiv preprint arXiv:1707.09564* (2017).
- [46] NITANDA, A., AND SUZUKI, T. Refined generalization analysis of gradient descent for over-parameterized two-layer neural networks with smooth activations on classification problems. *arXiv preprint arXiv:1905.09870* (2019).
- [47] OYMAK, S., AND SOLTANOLKOTABI, M. Overparameterized nonlinear learning: Gradient descent takes the shortest path?
- [48] OYMAK, S., AND SOLTANOLKOTABI, M. Towards moderate overparameterization: global convergence guarantees for training shallow neural networks. *arXiv preprint arXiv:1902.04674* (2019).
- [49] ROTSKOFF, G. M., AND VANDEN-EIJNDEN, E. Neural networks as interacting particle systems: Asymptotic convexity of the loss landscape and universal scaling of the approximation error.
- [50] SAGUN, L., EVCI, U., GUNEY, V. U., DAUPHIN, Y., AND BOTTOU, L. Empirical analysis of the hessian of over-parametrized neural networks. *arXiv preprint arXiv:1706.04454* (2017).
- [51] SCHMITT, B. A. Perturbation bounds for matrix square roots and pythagorean sums. *Linear algebra and its applications* 174 (1992), 215–227.
- [52] SCHUR, J. Bemerkungen zur theorie der beschränkten bilinearformen mit unendlich vielen veränderlichen. *Journal für die reine und angewandte Mathematik* 140 (1911), 1–28.
- [53] SIRIGNANO, J., AND SPILIOPOULOS, K. Mean field analysis of neural networks: A central limit theorem.
- [54] SONG, M., MONTANARI, A., AND NGUYEN, P. A mean field view of the landscape of two-layers neural networks. In *Proceedings of the National Academy of Sciences* (2018), vol. 115, pp. E7665–E7671.
- [55] SOUDRY, D., HOFFER, E., NACSON, M. S., GUNASEKAR, S., AND SREBRO, N. The implicit bias of gradient descent on separable data. *The Journal of Machine Learning Research* 19, 1 (2018), 2822–2878.
- [56] YEHUDAI, G., AND SHAMIR, O. On the power and limitations of random features for understanding neural networks. *arXiv preprint arXiv:1904.00687* (2019).
- [57] YIN, D., RAMCHANDRAN, K., AND BARTLETT, P. Rademacher complexity for adversarially robust generalization. *arXiv preprint arXiv:1810.11914* (2018).

- [58] YU, Y., WANG, T., AND SAMWORTH, R. J. A useful variant of the davis–kahan theorem for statisticians. *Biometrika* 102, 2 (2014), 315–323.
- [59] ZHANG, C., BENGIO, S., HARDT, M., RECHT, B., AND VINYALS, O. Understanding deep learning requires rethinking generalization. *arXiv preprint arXiv:1611.03530* (2016).
- [60] ZHANG, H., YU, D., CHEN, W., AND LIU, T.-Y. Training over-parameterized deep resnet is almost as easy as training a two-layer network. *arXiv preprint arXiv:1903.07120* (2019).
- [61] ZOU, D., CAO, Y., ZHOU, D., AND GU, Q. Stochastic gradient descent optimizes over-parameterized deep relu networks. *arXiv preprint arXiv:1811.08888* (2018).

## A The Jacobian of the Mixture Model is low-rank (Proofs for Section 3.3)

The following theorem considers a simple noiseless mixture model and proves that its Jacobian is low-rank and the concatenated multiclass label vectors lie on a rank  $K^2C$  information space associated with this Jacobian.

**Theorem A.1** *Consider a data set of size  $n$  consisting of input/label pairs  $\{(\mathbf{x}_i, \mathbf{y}_i)\}_{i=1}^n \in \mathbb{R}^d \times \mathbb{R}^K$  generated according to the Gaussian mixture model of Definition 3.4 with  $K$  classes each consisting of  $C$  clusters with the cluster centers given by  $\{\boldsymbol{\mu}_{\ell, \tilde{\ell}}\}_{(\ell, \tilde{\ell})=(1,1)}^{(K,C)}$  and  $\sigma = 0$ . Let  $\boldsymbol{\Sigma}(\mathbf{X})$  be the multiclass neural tangent kernel matrix associated with input matrix  $\mathbf{X} = [\mathbf{x}_1 \dots \mathbf{x}_n]^T$  with the standard deviation of the output layer set to  $\nu = \frac{1}{\sqrt{k}}$ . Also define the information space  $\mathcal{I}$  to be the range space of  $\boldsymbol{\Sigma}(\mathbf{X})$ . Also let  $\mathbf{M} = [\boldsymbol{\mu}_{1,1} \dots \boldsymbol{\mu}_{K,C}]^T$  be the matrix obtained by aggregating all the cluster centers as rows and let  $\mathbf{g}$  be a Gaussian random vector with distribution  $\mathcal{N}(\mathbf{0}, \mathbf{I}_d)$ . Define the neural tangent kernel matrix associated with the cluster centers as*

$$\tilde{\boldsymbol{\Sigma}}(\mathbf{M}) = (\mathbf{M}\mathbf{M}^T) \odot \mathbb{E}_{\mathbf{g} \sim \mathcal{N}(\mathbf{0}, \mathbf{I}_d)}[\phi'(\mathbf{M}\mathbf{g})\phi'(\mathbf{M}\mathbf{g})^T] \in \mathbb{R}^{KC \times KC},$$

and assume that  $\tilde{\boldsymbol{\Sigma}}(\mathbf{M})$  is full rank. Then, the following properties hold with probability  $1 - KC \exp(-\frac{n}{8KC})$

- $\mathcal{I}$  is a  $K^2C$  dimensional subspace.
- The concatenated label vector  $\mathbf{y} = [\mathbf{y}_1^T \quad \mathbf{y}_2^T \quad \dots \quad \mathbf{y}_n^T]^T$  lies on  $\mathcal{I}$ .
- The nonzero eigenvalues (top  $K^2C$  eigenvalues) of  $\boldsymbol{\Sigma}(\mathbf{X})$  are between  $\frac{n}{2KC} s_{\min}(\boldsymbol{\Sigma}(\mathbf{X}))$  and  $\frac{2n}{KC} \|\boldsymbol{\Sigma}(\mathbf{X})\|$ . Hence the eigenvalues of the information space grow with  $\frac{n}{KC}$ .

**Proof** First, we establish that each cluster has around the same size. Applying Chernoff bound and a union bound, we find that with probability  $1 - KC \exp(-\frac{n}{8KC})$

$$0.5\tilde{n} \leq \tilde{n}_{\ell, \tilde{\ell}} \leq 2\tilde{n}.$$

Note that based on Lemma 6.11, the multiclass covariance is given by

$$\boldsymbol{\Sigma}(\mathbf{X}) = k\nu^2 \mathbf{I}_K \otimes \tilde{\boldsymbol{\Sigma}}(\mathbf{X}).$$

where  $\tilde{\boldsymbol{\Sigma}}(\mathbf{X}) = (\mathbf{X}\mathbf{X}^T) \odot \mathbb{E}_{\mathbf{g} \sim \text{i.i.d. } \mathcal{N}(0,1)}[\phi'(\mathbf{X}\mathbf{g})\phi'(\mathbf{X}\mathbf{g})^T]$ . Due to this Kronecker product representation, the range space of  $\boldsymbol{\Sigma}(\mathbf{X})$  is separable. In particular, note that with

$$\tilde{\mathcal{I}} = \text{Range}(\tilde{\boldsymbol{\Sigma}}(\mathbf{X}))$$

we have  $\mathcal{I} = \mathbf{I}_K \otimes \tilde{\mathcal{I}}$  which also implies  $\text{rank}(\mathcal{I}) = K \cdot \text{rank}(\tilde{\mathcal{I}})$ . Hence, this identity allows us to reduce the problem to a single output network. To complete the proof we will prove the following three identities:

- $\tilde{\mathcal{I}}$  has rank  $KC$ .
- The nonzero eigenvalues of  $\tilde{\Sigma}(\mathbf{X})$  are between  $0.5\tilde{n}s_{\min}(\tilde{\Sigma}(\mathbf{M}))$  to  $2\tilde{n}\|\tilde{\Sigma}(\mathbf{M})\|$ .
- The portion of the label vector associated with class  $\ell$  i.e.  $\mathbf{y}^{(\ell)} \in \mathbb{R}^n$  (see (5.2)) lies on  $\mathcal{I}$ . Hence, the concatenated vector  $\mathbf{y}$  lies on  $\mathcal{I} = \mathbf{I}_K \otimes \tilde{\mathcal{I}}$ .

To prove these statements let  $\mathcal{J}_\ell(\mathbf{X}; \mathbf{W}_0)$  and  $\mathcal{J}_\ell(\mathbf{M}; \mathbf{W}_0)$  be the Jacobian associated with the  $\ell$ th output of the neural net (see (5.2)) for data matrices  $\mathbf{X}$  and  $\mathbf{M}$ . Observe that the columns of  $\mathcal{J}_\ell(\mathbf{X}; \mathbf{W}_0)$  are chosen from  $\mathcal{J}_\ell(\mathbf{M}; \mathbf{W}_0)$  and in particular each column of  $\mathcal{J}_\ell(\mathbf{M}; \mathbf{W}_0)$  is repeated between  $0.5\tilde{n}$  to  $2\tilde{n}$  times. To mathematically relate this, define the  $KC$  dimensional subspace  $\mathcal{S}$  of  $\mathbb{R}^n$  where for any  $\mathbf{v} \in \mathcal{S}$ , entries  $v_i$  and  $v_j$  of  $\mathbf{v}$  are equal iff data point  $\mathbf{x}_i$  and  $\mathbf{x}_j$  are equal (i.e. belong to the same class/cluster pair). Now, we define the orthonormal matrix  $\mathbf{U}_\mathcal{S} \in \mathbb{R}^{n \times KC}$  as the 0-1 matrix with orthogonal rows that map  $\mathbb{R}^{KC}$  to  $\mathcal{S}$  as follows. Assume the  $i$ th data point  $\mathbf{x}_i$  belongs to the class/cluster pair  $(\ell_i, \tilde{\ell}_i)$ . We then set the  $i$ th row of  $\mathbf{U}_\mathcal{S}$  as  $\text{vect}(\mathbf{e}_{\ell_i} \mathbf{e}_{\tilde{\ell}_i}^T)$ . Using  $\mathbf{U}_\mathcal{S}$  we have

$$\mathbf{U}_\mathcal{S} \mathcal{J}_\ell(\mathbf{M}; \mathbf{W}_0) = \mathcal{J}_\ell(\mathbf{X}; \mathbf{W}_0).$$

Now note that using the above identity we have

$$\mathbf{U}_\mathcal{S} \tilde{\Sigma}(\mathbf{M}) \mathbf{U}_\mathcal{S}^T = \tilde{\Sigma}(\mathbf{X}).$$

Since  $\mathbf{U}_\mathcal{S}$  is tall and orthogonal, the range of  $\tilde{\Sigma}(\mathbf{X})$  is exactly the range of  $\mathbf{U}_\mathcal{S}$  hence  $\tilde{\mathcal{I}} = \mathcal{S}$  which is  $KC$  dimensional. Furthermore, nonzero eigenvectors of  $\tilde{\Sigma}(\mathbf{X})$  lie on  $\mathcal{S}$  and any eigenvector  $\mathbf{v}$  satisfies

$$\mathbf{v}^T \tilde{\Sigma}(\mathbf{X}) \mathbf{v} \geq s_{\min}(\mathbf{U}_\mathcal{S})^2 s_{\min}(\tilde{\Sigma}(\mathbf{M})) \geq 0.5\tilde{n}s_{\min}(\tilde{\Sigma}(\mathbf{M}))$$

and similarly

$$\mathbf{v}^T \tilde{\Sigma}(\mathbf{X}) \mathbf{v} \leq 2\tilde{n}\|\tilde{\Sigma}(\mathbf{M})\|$$

which follows from the fact that  $\ell_2$ -norm-squared of columns of  $\mathbf{U}$  are between  $0.5\tilde{n}$  to  $2\tilde{n}$ . Finally, we will argue that label vector  $\mathbf{y}^{(\ell)}$  lies on  $\mathcal{S}$ . Note that for all samples  $i$  that belong to the same cluster  $\mathbf{y}_i^{(\ell)}$  will be the same (either zero or one), thus  $\mathbf{y}^{(\ell)} \in \mathcal{S}$ .  $\blacksquare$

Next lemma provides a perturbation analysis when there is noise.

**Lemma A.2** Consider the single-output NTK kernel given by

$$\tilde{\Sigma}(\mathbf{X}) = \mathbb{E} \left[ \phi'(\mathbf{X}\mathbf{w}) \phi'(\mathbf{X}\mathbf{w})^T \right] \odot (\mathbf{X}\mathbf{X}^T),$$

and assume that this matrix has rank  $r$  so that  $\lambda_{r+1}(\tilde{\Sigma}(\mathbf{X})) = \lambda_{r+2}(\tilde{\Sigma}(\mathbf{X})) = \dots = \lambda_n(\tilde{\Sigma}(\mathbf{X})) = 0$ . Also assume a noise corrupted version of  $\mathbf{X}$  given by

$$\tilde{\mathbf{X}} = \mathbf{X} + \frac{\sigma}{\sqrt{d}} \mathbf{Z}$$

with  $\mathbf{Z}$  a matrix consisting of i.i.d.  $\mathcal{N}(0, 1)$  entries. Then,  $\|\tilde{\Sigma}(\tilde{\mathbf{X}}) - \tilde{\Sigma}(\mathbf{X})\| \lesssim \Delta$  where

$$\Delta := \sigma^2 B^2 \log n \|\mathbf{X}\|^2 + \sigma^2 B^2 (n/d + 1) + \sqrt{\log n} \cdot \sigma B^2 \|\mathbf{X}\|^2 + \sigma B^2 \sqrt{n/d + 1} \|\mathbf{X}\| \quad (\text{A.1})$$

holds with probability at least  $1 - 2ne^{-\frac{d}{2}}$ . Whenever  $\sigma \leq \frac{1}{\sqrt{\log n}}$ ,  $\Delta$  is upper bounded as

$$\frac{\Delta}{n} \lesssim B^2 \sigma \sqrt{\log n}. \quad (\text{A.2})$$

Furthermore, let  $\tilde{\mathbf{V}}, \mathbf{V} \in \mathbb{R}^{n \times r}$  be orthonormal matrices corresponding to the top  $r$  eigenvalues of  $\tilde{\Sigma}(\tilde{\mathbf{X}})$  and  $\tilde{\Sigma}(\mathbf{X})$ . Then,

$$\|\tilde{\mathbf{V}} \tilde{\mathbf{V}}^T - \mathbf{V} \mathbf{V}^T\| \leq \frac{\Delta}{\lambda_r(\tilde{\Sigma}(\mathbf{X})) - \Delta}$$

**Proof** Note that

$$\begin{aligned} \text{diag}(\phi'(\tilde{\mathbf{X}}\mathbf{w}))\tilde{\mathbf{X}} - \text{diag}(\phi'(\mathbf{X}\mathbf{w}))\mathbf{X} &= \text{diag}(\phi'(\tilde{\mathbf{X}}\mathbf{w}))\tilde{\mathbf{X}} - \text{diag}(\phi'(\mathbf{X}\mathbf{w}))\mathbf{X} \\ &= \text{diag}(\phi'(\tilde{\mathbf{X}}\mathbf{w}) - \phi'(\mathbf{X}\mathbf{w}))\mathbf{X} \\ &\quad + \text{diag}(\phi'(\tilde{\mathbf{X}}\mathbf{w}))(\tilde{\mathbf{X}} - \mathbf{X}) \end{aligned}$$

Now define  $\tilde{\mathbf{M}} = \text{diag}(\phi'(\tilde{\mathbf{X}}\mathbf{w}))\tilde{\mathbf{X}}$  and  $\mathbf{M} = \text{diag}(\phi'(\mathbf{X}\mathbf{w}))\mathbf{X}$  and note that using the above we can conclude that

$$\begin{aligned} \|\tilde{\mathbf{M}} - \mathbf{M}\| &\leq \|\text{diag}(\phi'(\tilde{\mathbf{X}}\mathbf{w}) - \phi'(\mathbf{X}\mathbf{w}))\mathbf{X}\| \\ &\quad + \|\text{diag}(\phi'(\tilde{\mathbf{X}}\mathbf{w}))(\tilde{\mathbf{X}} - \mathbf{X})\| \\ &\leq B\|(\tilde{\mathbf{X}} - \mathbf{X})\mathbf{w}\|_{\ell_\infty}\|\mathbf{X}\| + B\|\tilde{\mathbf{X}} - \mathbf{X}\| \end{aligned}$$

Now using the fact that

$$\|\tilde{\mathbf{M}}\tilde{\mathbf{M}}^T - \mathbf{M}\mathbf{M}^T\| \leq \|\tilde{\mathbf{M}} - \mathbf{M}\|^2 + 2\|\tilde{\mathbf{M}} - \mathbf{M}\|\|\mathbf{M}\|,$$

we conclude that

$$\begin{aligned} \|\tilde{\Sigma}(\tilde{\mathbf{X}}) - \tilde{\Sigma}(\mathbf{X})\| &= \|\mathbb{E}[\tilde{\mathbf{M}}\tilde{\mathbf{M}}^T - \mathbf{M}\mathbf{M}^T]\| \\ &\leq \mathbb{E}\left[\left(B\|(\tilde{\mathbf{X}} - \mathbf{X})\mathbf{w}\|_{\ell_\infty}\|\mathbf{X}\| + B\|\tilde{\mathbf{X}} - \mathbf{X}\|\right)^2\right] \\ &\quad + 2B\|\mathbf{X}\|\left(B\|\mathbf{X}\|\mathbb{E}[\|(\tilde{\mathbf{X}} - \mathbf{X})\mathbf{w}\|_{\ell_\infty}] + B\|\tilde{\mathbf{X}} - \mathbf{X}\|\right) \\ &\leq 2B^2\|\mathbf{X}\|^2\mathbb{E}[\|(\tilde{\mathbf{X}} - \mathbf{X})\mathbf{w}\|_{\ell_\infty}^2] + 2B^2\|\tilde{\mathbf{X}} - \mathbf{X}\|^2 \\ &\quad + 2B^2\|\mathbf{X}\|^2\mathbb{E}[\|(\tilde{\mathbf{X}} - \mathbf{X})\mathbf{w}\|_{\ell_\infty}] + 2B^2\|\tilde{\mathbf{X}} - \mathbf{X}\|\|\mathbf{X}\| \end{aligned}$$

To proceed further, with probability  $1 - n\exp(-d/2)$ , each row of  $\tilde{\mathbf{X}} - \mathbf{X}$  is upper bounded by  $2\sigma$ . Hence, using a standard tail bound over supremum of  $n$  Gaussian random variables (which follows by union bounding) we have

$$\mathbb{E}[\|(\tilde{\mathbf{X}} - \mathbf{X})\mathbf{w}\|_{\ell_\infty}^2]^{1/2} \leq 2\sigma\sqrt{2\log n}$$

holds with the same probability. Furthermore, spectral norm bound on Gaussian random matrix implies that

$$\|\tilde{\mathbf{X}} - \mathbf{X}\|^2 \leq \left(2(\sqrt{n} + \sqrt{d})\right)^2 \frac{\sigma^2}{d} \leq 8(n/d + 1)\sigma^2,$$

holds with probability at least  $1 - e^{-\frac{1}{2}(n+d)}$ . Plugging these two probabilistic bounds into the chain of inequalities we conclude that

$$\|\tilde{\Sigma}(\tilde{\mathbf{X}}) - \tilde{\Sigma}(\mathbf{X})\| \lesssim \sigma^2 B^2 \log n \|\mathbf{X}\|^2 + \sigma^2 B^2 (n/d + 1) + \sqrt{\log n} \cdot \sigma B^2 \|\mathbf{X}\|^2 + \sigma B^2 \sqrt{n/d + 1} \|\mathbf{X}\|$$

To establish (A.2), observe that  $B^2\sigma\sqrt{\log n}\|\mathbf{X}\|^2$  dominates over other terms in the regime  $\sigma\sqrt{\log n}$  is small. The final bound is a standard application of Davis-Kahan Theorem [58] when we use the fact that  $\tilde{\Sigma}(\mathbf{X})$  is low-rank.  $\blacksquare$

The following lemma plugs in the critical quantities of Theorem 3.2 for our mixture model to obtain a generalization bound.

**Theorem A.3 (Generalization for Mixture Model)** *Consider a dataset  $\{\mathbf{x}_i, \mathbf{y}_i\}_{i=1}^n$  generated i.i.d. from the Gaussian mixture model in Definition 3.4. Let  $\lambda_{\mathbf{M}} = \lambda_{\min}(\Sigma(\mathbf{M}))$  where  $\mathbf{M} \in \mathbb{R}^{K \times d}$  is the matrix of cluster centers. Suppose input noise level  $\sigma$  obeys*

$$\sigma \lesssim \frac{\lambda_{\min}}{B^2 K C \sqrt{\log n}}$$

Consider the setup of Theorem 3.2 with quantities  $\zeta$  and  $\Gamma$ . Suppose network width obeys

$$k \gtrsim \frac{\Gamma^4 B^8 K^8 C^4 \log n}{\zeta^4 \lambda_{\min}^4}.$$

With probability  $1 - ne^{-d/2} - KC \exp(-\frac{n}{8KC}) - (2K)^{-100} - \delta$ , running gradient descent for  $T = \frac{2\Gamma K^2 C}{\eta \nu^2 n \lambda_{\min}}$  with learning rate  $\eta \leq \frac{1}{\nu^2 B^2 \|\mathbf{X}\|^2}$ , we have that

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \lesssim \sqrt{\frac{\sigma \sqrt{\log n} B^2 K C}{\lambda_{\min}}} + \frac{\Gamma B K \sqrt{C}}{\sqrt{n \lambda_{\min}}} + 12\zeta + 5\sqrt{\frac{\log(2/\delta)}{n}} + 2e^{-\Gamma}.$$

**Proof** The proof is an application of Lemma A.2 and Theorem A.1. Let  $\mathcal{I}'$  be the information space corresponding to noiseless dataset where input samples are identical to cluster centers. Let  $\mathbf{P}'$ ,  $\mathbf{P}$  correspond to the projection matrices to  $\mathcal{I}$  and  $\mathcal{I}'$ . First, using Lemma A.2 and the bound on  $\sigma$ , we have

$$\|\mathbf{P}' - \mathbf{P}\| \leq c \frac{\sigma \sqrt{\log n} B^2 K C}{\lambda_{\min}}$$

for some constant  $c > 0$ . Next we quantify  $\Pi_{\mathcal{I}}(\mathbf{y})$  using the fact that (i)  $\Pi_{\mathcal{I}'}(\mathbf{y}) = \mathbf{y}$  via Theorem A.1 as follows

$$\|\Pi_{\mathcal{I}}(\mathbf{y})\|_{\ell_2} \geq \|\Pi_{\mathcal{I}'}(\mathbf{y})\|_{\ell_2} - \|\Pi_{\mathcal{I}}(\mathbf{y}) - \Pi_{\mathcal{I}'}(\mathbf{y})\|_{\ell_2} \geq \sqrt{n}(1 - c \frac{\sigma \sqrt{\log n} B^2 K C}{\lambda_{\min}}). \quad (\text{A.3})$$

In return, this implies that

$$\|\Pi_{\mathcal{N}}(\mathbf{y})\|_{\ell_2} \lesssim \sqrt{\frac{n \sigma \sqrt{\log n} B^2 K C}{\lambda_{\min}}}$$

To proceed, we pick  $\alpha_0 = \sqrt{\frac{\lambda_{\min} n}{2KC}}$  and corresponding  $\bar{\alpha} = \frac{\alpha_0}{\sqrt[4]{n} \sqrt{K} \|\mathbf{X}\|_B} \geq \sqrt{\frac{\lambda_{\min}}{2B^2 K^2 C}}$  and apply (3.3) to find that, classification error is upper bounded by

$$\text{Err}_{\mathcal{D}}(\mathbf{W}_T) \lesssim \sqrt{\frac{\sigma \sqrt{\log n} B^2 K C}{\lambda_{\min}}} + \frac{\Gamma B K \sqrt{C}}{\sqrt{n \lambda_{\min}}} + 12\zeta + 5\sqrt{\frac{\log(2/\delta)}{n}} + 2e^{-\Gamma}.$$

■

## B Joint input-output optimization

In this section we wish to provide the ingredients necessary to prove a result for the case where both set of input and output weights  $\mathbf{W}$  and  $\mathbf{V}$  are trained. To this aim, we consider the combined neural net Jacobian associated with input and output layers given by

$$\mathbf{x} \mapsto f(\mathbf{x}; \mathbf{V}, \mathbf{W}) := \mathbf{V} \phi(\mathbf{W} \mathbf{x}). \quad (\text{B.1})$$

Denoting the Jacobian associated with (B.1) by  $\mathcal{J}(\mathbf{V}, \mathbf{W})$  we have that

$$\mathcal{J}(\mathbf{V}, \mathbf{W}) = [\mathcal{J}(\mathbf{V}) \ \mathcal{J}(\mathbf{W})] \in \mathbb{R}^{Kn \times k(K+d)}$$

Here,  $\mathcal{J}(\mathbf{W})$  is as before whereas  $\mathcal{J}(\mathbf{V})$  is the Jacobian with respect to  $\mathbf{V}$  and is given by

$$\mathcal{J}(\mathbf{V}) = [\mathcal{J}(\mathbf{v}_1) \ \mathcal{J}(\mathbf{v}_2) \ \dots \ \mathcal{J}(\mathbf{v}_K)]. \quad (\text{B.2})$$

where  $\mathcal{J}(\mathbf{v}_\ell) \in \mathbb{R}^{Kn \times k}$  is so that its  $\ell$ 'th block rows of size  $n \times k$  is nonzero for  $1 \leq \ell \leq K$  i.e.

$$\tilde{\ell}\text{th block row of } \mathcal{J}(\mathbf{v}_\ell) = \begin{cases} 0 & \text{if } \ell \neq \tilde{\ell} \\ \phi(\mathbf{X}\mathbf{W}^T) & \text{else} \end{cases}.$$

Hence,  $\mathcal{J}(\mathbf{V})$  is  $K \times K$  block diagonal with blocks equal to  $\phi(\mathbf{X}\mathbf{W}^T)$ . The following theorem summarizes the properties of the joint Jacobian.

**Theorem B.1 (Properties of the Combined Input/Output Jacobian)**  $\mathcal{J}(\mathbf{V}, \mathbf{W})$  satisfies the following properties.

- **Upper bound:**  $\|\mathcal{J}(\mathbf{V}, \mathbf{W})\| \leq B\|\mathbf{X}\|(\|\mathbf{W}\|_F + \sqrt{Kk}\|\mathbf{V}\|_{\ell_\infty})$ .
- **Row-bound:** For unit length  $\mathbf{u}$ :  $\|\text{mat}(\mathcal{J}^T(\mathbf{W})\mathbf{u})\|_{2,\infty} \leq B\sqrt{K}\|\mathbf{V}\|_{\ell_\infty}\|\mathbf{X}\|$ .
- **Entry-bound:** For unit length  $\mathbf{u}$ :  $\|\text{mat}(\mathcal{J}^T(\mathbf{V})\mathbf{u})\|_{\ell_\infty} \leq B\|\mathbf{W}\|_{2,\infty}\|\mathbf{X}\|$ .
- **Lipschitzness:** Given inputs  $\mathbf{V}, \mathbf{V}'$  and outputs  $\mathbf{W}, \mathbf{W}'$

$$\|\mathcal{J}(\mathbf{V}, \mathbf{W}) - \mathcal{J}(\mathbf{V}', \mathbf{W}')\| \leq B\|\mathbf{X}\|(\sqrt{Kk}\|\mathbf{V} - \mathbf{V}'\|_{\ell_\infty} + \sqrt{K}\|\mathbf{V}\|_{\ell_\infty}\|\mathbf{W}' - \mathbf{W}\|_F + \|\mathbf{W} - \mathbf{W}'\|_F).$$

**Proof** First, we prove results concerning  $\mathcal{J}(\mathbf{V})$ . First, note that

$$\|\mathcal{J}(\mathbf{V})\| \leq \|\phi(\mathbf{X}\mathbf{W}^T)\| \leq B\|\mathbf{X}\|\|\mathbf{W}\|_F.$$

Next, note that for  $\mathbf{u} = [\mathbf{u}_1 \dots \mathbf{u}_K] \in \mathbb{R}^{Kn}$  we have

$$\begin{aligned} \|\mathcal{J}^T(\mathbf{V})\mathbf{u}\|_{\ell_\infty} &= \max_{1 \leq \ell \leq K} \|\phi(\mathbf{W}\mathbf{X}^T)\mathbf{u}_\ell\|_{\ell_\infty} \\ &= \max_{1 \leq s \leq k} |\phi(\mathbf{w}_s\mathbf{X}^T)\mathbf{u}_\ell| \\ &= B\|\mathbf{W}\|_{2,\infty}\|\mathbf{X}\|. \end{aligned}$$

Let  $\mathcal{J}_1, \mathcal{J}_2$  be the Jacobian matrices restricted to  $\mathbf{V}$  and  $\mathbf{W}$  of  $\mathcal{J}(\mathbf{V}, \mathbf{W})$ . To prove Lipschitzness, first observe that

$$\|\mathcal{J}(\mathbf{V}, \mathbf{W}) - \mathcal{J}(\mathbf{V}', \mathbf{W}')\| \leq \|\mathcal{J}_1(\mathbf{V}, \mathbf{W}) - \mathcal{J}_1(\mathbf{V}', \mathbf{W}')\| + \|\mathcal{J}_2(\mathbf{V}, \mathbf{W}) - \mathcal{J}_2(\mathbf{V}', \mathbf{W}')\|.$$

Next, observe that

$$\|\mathcal{J}_1(\mathbf{V}, \mathbf{W}) - \mathcal{J}_1(\mathbf{V}', \mathbf{W}')\| \leq \|\phi(\mathbf{X}\mathbf{W}^T) - \phi(\mathbf{X}\mathbf{W}'^T)\| \leq B\|\mathbf{X}\|\|\mathbf{W} - \mathbf{W}'\|_F.$$

We decompose  $\mathcal{J}_2$  via

$$\begin{aligned} \|\mathcal{J}_2(\mathbf{V}, \mathbf{W}) - \mathcal{J}_2(\mathbf{V}', \mathbf{W}')\| &\leq \|\mathcal{J}_2(\mathbf{V}, \mathbf{W}) - \mathcal{J}_2(\mathbf{V}, \mathbf{W}')\| + \|\mathcal{J}_2(\mathbf{V}, \mathbf{W}') - \mathcal{J}_2(\mathbf{V}', \mathbf{W}')\| \\ &\leq B\sqrt{K}\|\mathbf{V}\|_{\ell_\infty}\|\mathbf{X}\|\|\mathbf{W}' - \mathbf{W}\|_F + \|\mathcal{J}_2(\mathbf{V}, \mathbf{W}') - \mathcal{J}_2(\mathbf{V}', \mathbf{W}')\|. \end{aligned}$$

To address the second term, note that, Jacobian is linear with respect to output layer hence

$$\|\mathcal{J}_2(\mathbf{V}, \mathbf{W}') - \mathcal{J}_2(\mathbf{V}', \mathbf{W}')\| = \|\mathcal{J}_2(\mathbf{V} - \mathbf{V}', \mathbf{W}')\| \leq B\sqrt{Kk}\|\mathbf{V} - \mathbf{V}'\|_{\ell_\infty}\|\mathbf{X}\|.$$

Combining the latter two identities we arrive at

$$\|\mathcal{J}_2(\mathbf{V}, \mathbf{W}) - \mathcal{J}_2(\mathbf{V}', \mathbf{W}')\| \leq B\|\mathbf{X}\|(\sqrt{Kk}\|\mathbf{V} - \mathbf{V}'\|_{\ell_\infty} + \sqrt{K}\|\mathbf{V}\|_{\ell_\infty}\|\mathbf{W}' - \mathbf{W}\|_F),$$

completing the proof. ■