Generalization Guarantees through Low Rank Jacobian

1 Generalization Guarantees For Neural Nets Via Harnessing the Low-Rankness of Jacobian

Definitions and notations.

- n: number of samples.
- d: dimension of training data.
- K: Number of classes, dimension of the output.
- One hidden layer neural network with the form

  $$x \mapsto f(x; W) := V\phi(Wx).$$

  where $x \in \mathbb{R}^d$, $W \in \mathbb{R}^{k\times d}$, $V \in \mathbb{R}^{K\times k}$ and ϕ is an activation function that acts component-wise. Only W is trained for simplicity in this work (but it is outlined how results can be generalized to the case in which V is also trained). We use the shorthand

  $$f(W) = [f(x_1; W)^\top, \ldots, f(x_n; W)^\top]^\top \in \mathbb{R}^{nK}.$$

- $(x_i, y_i) \in \mathbb{R}^d \times \mathbb{R}^K, 1 \leq i \leq n$: training data and corresponding labels (one-hot encodings).
- η: learning rate for gradient descent.
- $\theta \in \mathbb{R}^{kd}$: vectorized parameters of the neural network. We will denote $p = kd$.
- $\tilde{\theta} \in \mathbb{R}^{\max(Kn,p)}$: parameters of the linearized problem (more on this below).
- $\bar{\theta} \in \mathbb{R}^{\max(Kn,p)}$: θ (possibly) padded with x zeroes so it has the same length as $\tilde{\theta}$.
- $y = (y_1^\top, \ldots, y_n^\top)^\top \in \mathbb{R}^{nK}$: concatenation of labels.
- The loss function used in the optimization is the ℓ_2 loss:

  $$\mathcal{L}(W) = \frac{1}{2}\|\mathrm{f(W) - y}\|_2^2.$$

- The optimization algorithm is gradient descent, starting from an initialization W_0:

  $$W_{\tau+1} = W_\tau - \eta\nabla\mathcal{L}\left(W_\tau\right).$$

- (Remember we use $\theta \in \mathbb{R}^p$ for the vectorization of W). We use

  $$\mathcal{J}(\theta) = \frac{\partial f(\theta)}{\partial \theta} \in \mathbb{R}^{Kn\times p} \text{ so that } \theta_{\tau+1} = \theta_\tau - \eta\nabla\mathcal{L}(\theta_\tau) \text{ and } \nabla\mathcal{L}(\theta) = \mathcal{J}(\theta)^\top r(\theta).$$

  where we define the residual $r(\theta)$ as $f(\theta) - y$.

- **Information and Nuisance spaces**: For a matrix $J \in \mathbb{R}^{nK\times p}$ (that will typically be a Jacobian), consider its singular value decomposition

$$J = \sum_{s=1}^{nK} \lambda_s u_s v_s^T = U \operatorname{diag}(\lambda_1, \lambda_2, \ldots, \lambda_{nK}) V^T$$

with $\lambda_1 \geq \lambda_2 \geq \ldots \geq \lambda_{nK}$ and $u_s \in \mathbb{R}^{Kn}$, $v_s \in \mathbb{R}^p$ being the left and right singular vectors respectivelyi For a spectrum cutoff $0 < \alpha < \lambda_1$ let $c = c(\alpha)$ denote the index of the smallest singular value above α. Then the information and nuisance space associated with J are defined as

$$\mathcal{I} := \operatorname{span}\left(\{\boldsymbol{u}_s\}_{s=1}^{c}\right) \text{ and } \mathcal{N} := \operatorname{span}\left(\{\boldsymbol{u}_s\}_{s=c+1}^{Kn}\right).$$

- Multiclass Neural Tangent Kernel (M-NTK). Let $w \sim \mathcal{N}(0, I_d)$. Consider n input data points $x_1, \ldots, x_n \in \mathbb{R}^d$ aggregated in $X \in \mathbb{R}^{n\times d}$ and activation ϕ (it is assumed to be Lipschitz and smooth but the authors argue that they assume it for simplicity and outline how the result could be extended to use relu as activation). We define the multiclass kernel

$$\Sigma(X) := I_K \otimes \mathbb{E}\left[\left(\phi'(Xw)\phi'(Xw)^T\right) \odot \left(XX^T\right)\right],$$

where $\otimes$ is the Kronecker product and $\odot$ is the Hadamard product. This kernel is closely related to the Jacobian, it is known that $\mathbb{E}\left[\mathcal{J}(W_0)\mathcal{J}(W_0)^T\right] = \nu^2\Sigma(X)$ if V has i.i.d zero-mean entries with $\frac{\nu^2}{K}$ variance and W_0 has i.i.d. $\mathcal{N}(0,1)$ entries.

2 Overview

This work is along the lines of previous works that work with the NTK. In particular, using overparametrization, they will prove that the problem is close to its linearization $f(\theta) \approx f_{\text{lin}}(\theta) = f(\theta_0) + \mathcal{J}(\theta_0)(\theta - \bar{\theta}_0)$ (since we will find solutions close to θ_0) and that will allow them to state their optimization theorem for neural networks, to be explained later. The main trait of this work is that they remove the main assumption on the data (but the unit length assumption) made by other works and as a result their results incur some bias. In particular instead of assuming that two data points are not parallel, proving using this that the NTK is positive semi-definite and having a dependence (in terms of overparametrization and number of iterations needed) on the inverse minimum eigenvalue of the NTK (e.g. [2]) or instead of assuming that any two data points satisfy $\|\mathrm{x_i} - \mathrm{x_j}\| \geq \delta$ and having a dependence (in terms of overparametrization and number of iterations needed) on the inverse of δ (e.g. [1]), they allow the NTK to have 0 or very small eigenvalues and split the space into the information space (span of the first top left singular vectors of the jacobian at initialization or equivalently, first top eigenvectors of the NTK) and the nuisance space, proving that now the dependence (in terms of overparametrization and number of optimization time steps needed) is on the inverse of the lowest eigenvalue of the information space and the projection of the residual on the information space decreases exponentially while the projection of the residual on the nuisance space increases by a constant factor. They also work with the setting of arbitrary initialization in which under some assumptions, they can follow similar arguments to those made in the NTK, so they obtain an optimization guarantee in such a case. Also, the optimization guarantee translates to a generalization guarantee via the use of standard Rademacher complexity arguments.

We outline now the main approach followed to prove their main (meta-)theorem:

- Provided that our network has enough overparametrization, we can **relate the training of the neural network with gradient descent with a linear method**. This is in the sense that both the trajectory and the residuals of the linear method and the residuals will be close. Given an initial point $\theta_0 \in \mathbb{R}^p$, define an (ε_0, β) reference Jacobian $J \in \mathbb{R}^{Kn\times \max(Kn,p)}$ a matrix satisfying:

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

where $\overline{\mathcal{J}}(\theta_0) \in \mathbb{R}^{Kn\times\max(Kn,p)}$ is a matrix obtained by appending $\max(0, Kn-p)$ zero columns to $\mathcal{J}(\theta_0)$ (note $\mathcal{J}(\theta_0) \in \mathbb{R}^{Kn\times p}$).

In the random initialization setting, the reference Jacobian will be the NTK. In the arbitrary initialization setting, the reference Jacobian will be the Jacobian at that initialization.

The bounded spectra of J will be an assumption in the arbitrary initialization and a consequence of the properties of the NTK in the other case. The reason why the other two conditions are true for the random initialization regime is that in the overparametrization regime the NTK of the finite net tends to the infinite width limit NTK.

- **Bounded perturbation.** Due to the overparametrization and the small choice of the learning rate we will have
$$\|\theta_0 - \theta_\tau\| < R,$$
for a constant R for all t between 0 and T, where T is picked later. This will along with overparametrization imply
$$\|\mathrm{J}(\theta_0) - \mathrm{J}(\theta_\tau)\| < \varepsilon.$$
for a constant ε.

- Now if we followed other works on the NTK, we would **analyze the linear case** and would see that the residual of the linearized problem, $\widetilde{r}_\tau$ evolves in a precise sense
$$\widetilde{r}_\tau = U\left(I - \eta\Lambda^2\right)^\tau a = \sum_{s=1}^{nK}\left(1 - \eta\lambda_s^2\right)^\tau a_s u_s$$
where we are using the matrices U and Λ that come from the singular value decomposition of the reference Jacobian $J = U\Lambda V$. Also, λ_s are the diagonal entries of Λ, u_s are the rows of U and a is a vector whose value is the projection of the initial residual via U, i.e. $a = U^\top \tilde{r}_0 = U^\top r_0$. Previous approaches used that λ_{nK}^2, (the smallest one) is positive, and set a good value of the overparametrization and learning rate parameters (high and low respectively) to show that the corresponding eigenvalue for the Jacobian at initialization is positive too and finally, they used the bounded perturbation property to conclude that the residual also decreases with time. In this work, we follow this approach only for the information space, and since there is no assumption on λ_{nK}^2 being > 0, the approximation error incurred by the linearization could mean that the projection of the residual on the nuisance space is increasing. However, if it increases it does it at a slow pace, since the approximation error is low in the overparametrization regime. In particular, we have, for the linearized regime
$$\|\widetilde{r}_\tau\|_{\ell_2} \leq \left(1 - \eta\alpha^2\right)^\tau \|\Pi_{\mathcal{I}}(r_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(r_0)\|_{\ell_2}.$$
and if we define $e_{\tau+1} = r_{\tau+1} - \widetilde{r}_{\tau+1}$ then it obeys (assuming small learning rate, in particular $\eta < \beta^2$):
$$\|e_{\tau+1}\|_{\ell_2} \leq \eta\left(\varepsilon_0^2 + \varepsilon\beta\right)\|\widetilde{r}_\tau\|_{\ell_2} + \left(1 + \eta\varepsilon^2\right)\|e_\tau\|_{\ell_2}$$
which intuitively means that the error increases by a summand that is of the order of the residual plus a multiplicative expansion with respect to the previous error, due to the nuisance space. However, the rate of increase is small enough so that after T iterations the error will be controlled. Once we have this, we can proceed to the next step.

- Use overparametrization (and bounded perturbation) to prove that in particular one has
$$\|r_\tau - \widetilde{r}_\tau\|_{\ell_2} \leq \frac{3}{5}\frac{\delta\alpha}{\beta}\|r_0\|_{\ell_2} \quad \text{and} \quad \left\|\bar{\theta}_\tau - \widetilde{\theta}_\tau\right\|_{\ell_2} \leq \delta\frac{\Gamma}{\alpha}\|r_0\|_{\ell_2},$$
where δ is a hyperparameter, Γ is another hyperparameter that modulates the total number of time steps (that is chosen to be $T = \frac{\Gamma}{\eta\alpha^2}$). Finally, $\bar{\theta}$ is equal to $\theta \in \mathbb{R}^p$ padded with zeroes till size $\max(Kn, p)$.

- Prove bounded initial residual. In the random initialization regime, this will be a property one can prove about the NTK. In the arbitrary initialization regime, it is an assumption.

- **Put all together** to conclude
$$\|r_T\|_{\ell_2} \leq e^{-\Gamma}\|\Pi_{\mathcal{I}}(r_0)\|_{\ell_2} + \|\Pi_{\mathcal{N}}(r_0)\|_{\ell_2} + \frac{\delta\alpha}{\beta}\|r_0\|_{\ell_2}.$$

2.1 Some Proofs

In this section, we give more precise statements and prove some of the things we talked about in the previous section.

We will first make the following two assumptions about the jacobians of our non-linear models. We will see that these assumptions hold when these non-linear models are two-layered neural networks with smooth activation functions.

Assumption 1 (β-Bounded spectrum)**.** *The non-linear function $f : \mathbb{R}^p \to \mathbb{R}^n$ satisfies the β-bounded spectrum assumption, when the jacobian associated with f satisfies the following for all $\theta \in \mathbb{R}^p$*

$$\|\mathcal{J}(\theta)\| \leq \beta \tag{1}$$

Assumption 2 ($(\varepsilon, R, \theta_0)$-bounded jacobian perturbation)**.** *The non-linear function $f : \mathbb{R}^p \to \mathbb{R}^n$ satisfies the $(\varepsilon, R, \theta_0)$-bounded jacobian perturbation when the following is satisfied for all $\theta \in \mathbb{R}^p$ such that $\|\theta - \theta_{\mathrm{p}}\| \leq R$*

$$\|\mathcal{J}(\theta) - \mathcal{J}(\theta_0)\| \leq \frac{\varepsilon}{2} \tag{2}$$

We will be looking at the following meta theorem.

Theorem 1. *Consider a non-linear least squares problem of the form $\mathcal{L}(\theta) = \frac{1}{2}\|\mathrm{f}(\theta) - \mathrm{y}\|_2^2$ with $f : \mathbb{R}^p \to \mathbb{R}^{nK}$ the multi-class non-linear mapping, $\theta \in \mathbb{R}^p$ the parameters of the model, and $\mathbf{y} \in \mathbb{R}^{nK}$ the concatenated labels. Let $\bar{\theta}$ be a zero-padded θ till size $\max(Kn, p)$, Also consider a point $\theta_0 \in \mathbb{R}^p$ with $\mathbf{J}$ an (ε_0, β) reference Jacobian associated with $\mathcal{J}(\theta_0)$, and fitting the linearized problem $f_{\mathrm{lin}}\left(\widetilde{\theta}\right) = f(\theta_0) + \mathbf{J}\left(\widetilde{\theta} - \bar{\theta_0}\right)$ via the loss $\mathcal{L}_{\mathrm{lin}}(\theta) = \frac{1}{2}\|\mathrm{f}_{\mathrm{lin}}(\theta) - \mathrm{y}\|_2^2$.*

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 α.

Furthermore consider a point $\theta_0 \in \mathbb{R}^p$, a tolerance level $0 < \delta \leq 1$, stopping time $\Gamma \geq 1$ and assume the Jacobian mapping $\mathcal{J}(\theta) \in \mathbb{R}^{nK \times p}$ associated with f obeys β-Bounded spectrum assumption (Assumption 1) and $(\varepsilon, R, \theta_0)$-bounded jacobian perturbation assumption (Assumption 2) for

$$R := 2\left(\left\|\mathbf{J}_{\mathcal{I}}^{\dagger}\mathbf{r}_0\right\|_2 + \frac{\Gamma}{\alpha}\|\Pi_{\mathcal{N}}(\mathbf{r}_0)\| + \delta\frac{\Gamma}{\alpha}\|\mathbf{r}_0\|_2\right) \tag{3}$$

and

$$\varepsilon \leq \frac{\delta\alpha^3}{5\Gamma\beta^2} \tag{4}$$

Finally assume the following in regards to the reference Jacobian.

$$\varepsilon_0 \leq \frac{\min\left(\delta\alpha, \sqrt{\frac{\delta\alpha^3}{\Gamma\beta}}\right)}{5} \tag{5}$$

We run gradient descent iterations of the form a) Original Problem: $\theta_{\tau+1} = \theta_\tau - \eta\nabla\mathcal{L}(\theta_\tau)$ and b) Linearized Problem: $\widetilde{\theta}_{\tau+1} = \widetilde{\theta}_\tau - \eta\nabla\mathcal{L}_{\mathrm{lin}}\left(\widetilde{\theta}_\tau\right)$ starting from θ_0 with step size η obeying $\eta \leq \frac{1}{\beta^2}$.

Then for all iterates obeying $0 \leq \tau \leq T := \frac{\Gamma}{\eta\alpha^2}$ iterations of the original (θ_τ) and linearized $\left(\widetilde{\theta}_\tau\right)$ problems and the corresponding residuals $\mathbf{r}_\tau := f(\theta_\tau) - \mathbf{y}$ and $\widetilde{\mathbf{r}}_\tau := f_{\mathrm{lin}}\left(\widetilde{\theta}_\tau\right) - \mathbf{y}$ closely track each other.

That is

- ***Original and linear residuals are close:***

$$\|\mathbf{r}_\tau - \widetilde{\mathbf{r}}_\tau\| \leq \frac{3}{5}\frac{\delta\alpha}{\beta}\|\mathbf{r}_0\| \tag{6}$$

- ***Original and linearized paramaters are close:***

$$\left\|\bar{\theta}_\tau - \widetilde{\theta}_\tau\right\| \leq \delta\frac{\Gamma}{\alpha}\|\mathbf{r}_0\| \tag{7}$$

- ***Original iterates are close to initialization****: Furthermore, for all iterates $0 \le \tau \le T := \frac{\Gamma}{\eta\alpha^2}$, we have that the original parameters θ_τ is close to the inital parameters.*

$$\|\theta_\tau - \theta_0\| \le \frac{R}{2} = \left\|\mathbf{J}_{\mathcal{I}}^{\dagger}\mathbf{r}_0\right\|_2 + \frac{\Gamma}{\alpha}\|\Pi_{\mathcal{N}}\left(\mathbf{r}_0\right)\|_2 + \delta\frac{\Gamma}{\alpha}\|\mathbf{r}_0\|_2 \tag{8}$$

- ***Final non-linear residual is bounded****: and at $\tau = T$, we have that*

$$\|\mathbf{r}_{\mathrm{T}}\|_2 \le e^{-\Gamma}\|\Pi_{\mathcal{I}}\left(\mathbf{r}_0\right)\|_2 + \|\Pi_{\mathcal{N}}\left(\mathbf{r}_0\right)\| + \frac{\delta\alpha}{\beta}\|\mathbf{r}_0\|_2 \tag{9}$$

First, we will show that the difference of the non-linear and the linear residual in the τ^{th} time step is of the order of the difference in the previous time step added to a term linear in the residual of the linearized problem. Precisely, it is stated as follows.

Lemma 2 (Lemma 6.7)**.** *Assume Assumption 1 (with β) and Assumption 2 (with $(\varepsilon, R, \theta_0)$) hold and θ_τ and $\theta_{\tau+1}$ are within an R neighbourhood of the initilization θ_0, i.e. $\|\theta_\tau - \theta_0\| \le R$ and $\|\theta_{\tau+1} - \theta_0\| \le R$.*

Then, on running gradient descent with $\eta \le \frac{1}{\beta^2}$. the difference between the non-linear and the linear residual $\mathbf{e}_{\tau+1} = \mathbf{r}_{\tau+1} - \widetilde{\mathbf{r}}_{\tau+1}$ follow

$$\|\mathbf{r}_{\tau+1}\|_2 \le \eta\left(\varepsilon_0^2 + \varepsilon\beta\right)\|\widetilde{\mathbf{r}}_\tau\|_2 + \left(1 + \eta\varepsilon^2\right)\|\mathbf{e}_\tau\|_2 \tag{10}$$

Proof. Let $\mathbf{A} = \mathcal{J}\left(\theta_0\right), \mathbf{B}_2 = \mathcal{J}\left(\theta_\tau\right)$ and

$$\mathbf{B}_1 = \mathcal{J}\left(\theta_\tau, \theta_{\tau+1}\right) = \int_0^1 \mathcal{J}\left(t\theta_{\tau+1} + (1-t)\,\theta_\tau\right)dt$$

By 0^{th} order Taylor expansion with remainder term, we can write

$$\begin{aligned} f\left(\theta_{\tau+1}\right) = f\left(\theta_\tau - \eta\nabla\mathcal{L}\left(\theta_\tau\right)\right) &= f\left(\theta_\tau\right) + \eta\mathbf{B}_1\nabla\mathcal{L}\left(\theta_\tau\right) \\ &= f\left(\theta_\tau\right) + \eta\mathbf{B}_1\mathbf{B}_2^\top\left(f\left(\theta_\tau\right) - \mathbf{y}\right) \\ \mathbf{r}_{\tau+1} = f\left(\theta_{\tau+1}\right) - \mathbf{y} &= \left(\mathbf{I} - \eta\mathbf{B}_1\mathbf{B}_2^\top\right)\mathbf{r}_\tau \end{aligned}$$

For the linear problem, we have

$$\widetilde{\mathbf{r}}_{\tau+1} = \left(\mathbf{I} - \mathbf{J}\mathbf{J}^\top\right)\widetilde{\mathbf{r}}_\tau$$

Thus

$$\begin{aligned} \|\mathbf{e}_{\tau+1}\| = \|\mathbf{r}_{\tau+1} - \widetilde{\mathbf{r}}_{\tau+1}\| &= \left\|\left(\mathbf{I} - \eta\mathbf{B}_1\mathbf{B}_2^\top\right)\mathbf{r}_\tau - \left(\mathbf{I} - \mathbf{J}\mathbf{J}^\top\right)\widetilde{\mathbf{r}}_\tau\right\| \\ &= \left\|\left(\mathbf{I} - \mathbf{B}_1\mathbf{B}_2^\top\right)\mathbf{e}_\tau - \eta\left(\mathbf{B}_1\mathbf{B}_2^\top - \mathbf{J}\mathbf{J}^\top\right)\widetilde{\mathbf{r}}_\tau\right\| \\ &\le \left\|\left(\mathbf{I} - \mathbf{B}_1\mathbf{B}_2^\top\right)\mathbf{e}_\tau\right\| + \eta\left\|\left(\mathbf{B}_1\mathbf{B}_2^\top - \mathbf{J}\mathbf{J}^\top\right)\right\|\|\widetilde{\mathbf{r}}_\tau\| \end{aligned}$$

First, we bound $\left\|\left(\mathbf{I} - \mathbf{B}_1\mathbf{B}_2^\top\right)\mathbf{e}_\tau\right\|$ using the fact (Lemma 6.3) that if $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times p}$ are matrices obeying $\|\mathbf{A}\|, \|\mathbf{B}\| \le \beta, \|\mathbf{B} - \mathbf{A}\| \le \varepsilon$, then $\forall\ \mathbf{z} \in \mathbb{R}^n, \eta \le \frac{1}{\beta^2}$ we have that $\left\|\left(\mathbf{I} - \mathbf{B}_1\mathbf{B}_2^\top\right)\mathbf{z}\right\| \le \left(1 + \eta\varepsilon^2\right)\|\mathbf{z}\|_2$. Next, we bound $\left\|\left(\mathbf{B}_1\mathbf{B}_2^\top - \mathbf{J}\mathbf{J}^\top\right)\right\|$ as follows.

$$\begin{aligned} \left\|\left(\mathbf{B}_1\mathbf{B}_2^\top - \mathbf{J}\mathbf{J}^\top\right)\right\| &= \left\|\left(\mathbf{B}_1\mathbf{B}_2^\top - \mathbf{A}\mathbf{B}_2^\top + \mathbf{A}\mathbf{B}_2^\top - \mathbf{A}\mathbf{A}^\top + \mathbf{A}\mathbf{A}^\top - \mathbf{J}\mathbf{J}^\top\right)\right\| \\ &\le \left\|\left(\mathbf{B}_1 - \mathbf{A}\right)\mathbf{B}_2^\top\right\| + \left\|\mathbf{A}\left(\mathbf{B}_2^\top - \mathbf{A}^\top\right)\right\| + \left\|\mathbf{A}\mathbf{A}^\top - \mathbf{J}\mathbf{J}^\top\right\| \\ &\beta\frac{\varepsilon}{2} + \beta\frac{\varepsilon}{2} + \varepsilon_0^2 \end{aligned}$$

□

Next, we will prove a lemma that will finally allow us to control the growth of the difference between the linear and the non-linear residuals.

Lemma 3 (Lemma 6.8)**.** *Consider positive scalars $\Gamma, \alpha, \varepsilon, \eta > 0$. Also assume $\eta \leq \frac{1}{\alpha^2}$ and $\alpha \geq \sqrt{2\Gamma\varepsilon}$ and set $T = \frac{\Gamma}{\eta\alpha^2}$. For $0 \leq \tau \leq T$, non-negative entries $\rho_, \rho_+ \geq 0$, assume that the scalar sequences e_τ and $\widetilde{r}_\tau$ obey the following*

- $e_0 = 0$
- $\widetilde{r}_\tau \leq \left(1 - \eta\alpha^2\right)^\tau \rho_+ + \rho_-$
- $e_\tau \leq \left(1 + \eta\varepsilon^2\right) e_{\tau-1} + \eta\Theta\widetilde{r}_{\tau-1}$

Let $\Lambda = \dfrac{2\left(\Gamma\rho_- + \rho_+\right)}{\alpha^2}$. Then for all $0 \leq \tau \leq T$, the following holds

$$e_\tau \leq \Theta\Lambda$$

Proof. We will prove this by induction. Note that $e_0 = 0$ satisfies the base condition. Suppose, $e_t \leq \Theta\Lambda$ holds for all $t < \tau$.

Thus for all $0 \leq t \leq \tau$,

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

□

We will combine this to provide a rough proof of 6 using induction.

Proof of Theorem 1. We will prove this by induction. We will only provide a rough proof sketch to keep it simple and ignore the computations. We will assume that for all $0 \leq t \leq \tau$ the induction hypothesis holds true i.e. $\|\theta_0 - \theta_\mathrm{t}\| \leq R$ 3, 6, 7, 8, and 9 holds true. We will show that they all hold true for $t = \tau + 1$

- **Proving $\|\theta_0 - \theta_\mathrm{t}\| \leq R$ for all $t = \tau + 1$:** We know by (8) that $\|\theta_0 - \theta_\tau\| \leq \frac{R}{2}$. We need to show that $\|\theta_\tau - \theta_{\tau+1}\| \leq \frac{R}{2}$.

$$\begin{aligned}
\|\theta_\tau - \theta_{\tau+1}\| &= \eta\|\nabla\mathcal{L}\left(\theta_\tau\right)\| = \eta\left\|\mathcal{J}^\top\left(\theta_\tau\right)\mathbf{r}_\tau\right\| \\
&\leq \eta\left\|\mathbf{J}^\top\widetilde{\mathbf{r}}_\tau\right\| + \eta\left\|\left(\mathcal{J}\left(\theta_\tau\right) - \mathbf{J}\right)^\top\right\|\|\widetilde{\mathbf{r}}_\tau\| + \eta\|\mathcal{J}\left(\theta_\tau\right)\|\|\widetilde{\mathbf{r}}_\tau - \mathbf{r}_\tau\|
\end{aligned}$$

We can bound the first term as (Page 25) as

$$\eta\left\|\mathbf{J}^\top\widetilde{\mathbf{r}}_\tau\right\| \leq \left\|\mathbf{J}_\mathcal{I}^\dagger\mathbf{r}_0\right\|_2 + \frac{\Gamma}{\alpha}\|\Pi_\mathcal{N}\left(\mathbf{r}_0\right)\|$$

the second term as

$$\eta\left\|\left(\mathcal{J}\left(\theta_\tau\right)-\mathbf{J}\right)^\top\right\|\|\widetilde{\mathbf{r}}_\tau\| \leq \eta\left\|\left(\mathcal{J}\left(\theta_\tau\right)-\mathcal{J}\left(\theta_0\right)\right)+\left(\mathcal{J}\left(\theta_0\right)-\mathbf{J}\right)^\top\right\|\|\widetilde{\mathbf{r}}_0\| \leq \left(\varepsilon+\varepsilon_0\right)\|\widetilde{\mathbf{r}}_0\| \leq \frac{2\delta\alpha}{5\beta^2}\|\widetilde{\mathbf{r}}_0\|$$

and the third term as (Using Eq (6))

$$\eta\|\mathcal{J}\left(\theta_\tau\right)\|\|\widetilde{\mathbf{r}}_\tau-\mathbf{r}_\tau\| \leq \frac{3\delta\alpha}{5\beta^2}\|\widetilde{\mathbf{r}}_0\|$$

Combining them we get

$$\|\theta_\tau-\theta_{\tau+1}\| \leq \left\|\mathbf{J}_{\mathcal{I}}^\dagger\mathbf{r}_0\right\|_2+\frac{\Gamma}{\alpha}\|\Pi_{\mathcal{N}}\left(\mathbf{r}_0\right)\|+\frac{\delta\alpha}{\beta^2}\|\widetilde{\mathbf{r}}_0\| = \frac{R}{2}$$

- **Proving that** $\|\mathbf{e}_{\tau+1}\| \leq \frac{3}{5}\frac{\delta\alpha}{\beta}\|\mathbf{r}_0\|$**:** We have shown that $\|\theta_{\mathrm{t}}-\theta_0\| \leq R$, for $t \leq \tau+1$. Then we can use Lemma 2 to say that for all $0 < t \leq \tau+1$ the following holds

  $$\|\mathbf{e}_{\mathrm{t}}\| \leq \eta\left(\varepsilon_0^2+\varepsilon\beta\right)\|\widetilde{\mathbf{r}}_{\mathrm{t}-1}\|+\left(1+\eta\varepsilon^2\right)\|\mathbf{e}_{\mathrm{t}-1}\|_2$$

  We already know that the linear residuals satisfy the following for all $0 < t \leq \tau+1$

  $$\|\widetilde{\mathrm{r}}_{\mathrm{t}}\| \leq \left(1-\eta\alpha^2\right)^t\|\Pi_{\mathcal{I}}\left(\mathbf{r_0}\right)\|+\|\Pi_{\mathcal{N}}\left(\mathbf{r}_0\right)\|$$

  Finally, we can apply Lemma 3 with the following substitutions

  - $\Theta = \varepsilon_0^2+\varepsilon\beta$
  - $\rho_+ = \|\Pi_{\mathcal{I}}\|\left(\mathbf{r}_0\right), \quad \rho\_ = \|\Pi_{\mathcal{N}}\|\left(\mathbf{r}_0\right)$
  - $e_\tau = \|\mathbf{e}_{\tau+1}\|, \quad \widetilde{r}_\tau = \|\widetilde{\mathbf{r}}_{\tau+1}\|$

  Note that all the assumptions of Lemma 3 are also satisfied i.e. a) $\eta \leq \frac{1}{\beta^2} \leq \frac{1}{\alpha^2}$ as $\beta \geq \alpha$, the cutoff singular value of $\mathbf{J}$, and b) By definition of ε (eq. 4), $\frac{\alpha}{\varepsilon} \geq \frac{5\Gamma}{\delta}\frac{\beta^2}{\alpha^2} \geq \sqrt{2\Gamma}$. $\therefore \alpha \geq$ Thus,

  $$\begin{aligned}\|\mathbf{e}_{\tau+1}\| &\leq 2\left(\varepsilon_0^2+\varepsilon\beta\right)\frac{\Pi_{\mathcal{I}}\left(\mathbf{r}_0\right)+\Gamma\Pi_{\mathcal{N}}\left(\mathbf{r}_0\right)}{\alpha^2}\\ &\leq \frac{2\Gamma\left(\varepsilon_0^2+\varepsilon\beta\right)\|\mathbf{r}_0\|}{\alpha^2}\\ &\leq \left(\frac{2}{25}+\frac{2}{5}\right)\frac{\delta\alpha}{\beta}\|\mathbf{r}_0\| \leq \frac{3}{5}\frac{\delta\alpha}{\beta}\|\mathbf{r}_0\|\end{aligned}$$

Finally, we can prove Eq (9)

$$\begin{aligned}\|\mathbf{r}_{\mathrm{T}}\| &\leq \|\widetilde{\mathbf{r}}_{\mathrm{T}}\|+\|\mathbf{r}_{\mathrm{T}}-\widetilde{\mathbf{r}}_{\mathrm{T}}\|\\ &\leq \left(1-\eta\alpha^2\right)^t\|\Pi_{\mathcal{I}}\left(\mathbf{r_0}\right)\|+\|\Pi_{\mathcal{N}}\left(\mathbf{r}_0\right)\|+\frac{3}{5}\frac{\delta\alpha}{\beta}\|\mathbf{r}_0\|\\ &\leq e^{-\Gamma}\|\Pi_{\mathcal{I}}\left(\mathbf{r_0}\right)\|+\|\Pi_{\mathcal{N}}\left(\mathbf{r}_0\right)\|+\frac{\delta\alpha}{\beta}\|\mathbf{r}_0\|\end{aligned}$$

□

For a neural network, we just need to ensure that Assumption (1) and (2) are satisfied. We state the a simplified theorem for the generelization of multi-class neural networks below without providing a proof. But it follows from the above using a rademacher complexity argument.

Theorem 4 (Generelization of multi-output neural network)**.** *With $\Gamma > 0$, consider an i.i.d. dataset $\{(\mathbf{x}_i, y_i)\} \in \mathbb{R}^d \times \mathbb{R}^K$ where $\mathbf{x}_i$ are unit length data points and $\mathbf{y}_i s$ are one-hot encoded labels.*

Consider the neural network to be initialized with $\mathbf{W}_0 \sim \mathcal{N}(0, \mathbf{I})$ and $\mathbf{V}$ be properly scaled rademacher entries.

Consider the reference jacobian, with information and nuisance subspaces divided according to α, to be $\mathbf{J} = \Sigma(\mathbf{X})^{1/2}$ where

$$\Sigma(\mathbf{X}) = \mathbf{I}_K \otimes \mathbb{E}\left[\left(\phi'(\mathbf{X}\mathbf{w})\,\phi'(\mathbf{X}\mathbf{w})^\top\right) \odot \left(\mathbf{X}\mathbf{X}^\top\right)\right]$$

Assume the overparameterization to be

$$k \geq \frac{\Gamma^4 \log n}{\alpha^8}$$

Then after $T = \dfrac{\Gamma}{\eta\alpha^2}$ iterations, the generalization error obeys

$$\operatorname{Err}(\mathbf{W}_T) \leq \frac{\Pi_{\mathcal{N}}(\mathbf{y})}{\sqrt{n}} + e^{-\Gamma} + \frac{\Gamma}{\alpha\sqrt{n}}$$

Lemma 5. *For a neural network as defined above where the activation function ϕ is such that $|\phi'(\mathbf{z})| \leq B$ and $|\phi''(\mathbf{z})| \leq B$ for all $\mathbf{z}$, K is the number of classes, then for all $\mathbf{W} \in \mathbb{R}^{k \times d}$*

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

and if all data points are unit norm i.e. $\|\mathbf{x}_\mathrm{i}\| = 1$ then the jacobian is lipschitz with respect to the spectral norm for all $\mathbf{W}, \widetilde{\mathbf{W}} \in \mathbb{R}^{k \times d}$

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

Proof. Given two matrices $\mathbf{A} = \left[\mathbf{A}_1^\top, \cdots, \mathbf{A}_K^\top\right]$ and $\mathbf{B} = \left[\mathbf{B}_1^\top, \cdots, \mathbf{B}_K^\top\right]$, the following is true

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

We will first show that for a single output neural network i.e. for $K = 1$ we have $\|\mathcal{J}(\mathbf{W})\| \leq B\sqrt{k}\|\mathbf{V}\|_\infty \|\mathbf{X}\|$

$$\begin{aligned}
\mathcal{J}(\mathbf{W})\,\mathcal{J}^\top(\mathbf{W}) &= \left(\phi'\left(\mathbf{X}\mathbf{W}^\top\right) \operatorname{diag}(\mathbf{v}) \operatorname{diag}(\mathbf{v})\,\phi'\left(\mathbf{W}\mathbf{X}^\top\right)\right) \odot \left(\mathbf{X}\mathbf{X}^\top\right) \\
\|\mathcal{J}(\mathbf{W})\|^2 &\leq \left(\max_i \|\operatorname{diag}(\mathbf{v})\,\phi'(\mathbf{W}\mathbf{x}_\mathrm{i})\|^2\right) \|\mathbf{X}\|_2^2 \\
&\leq kB^2 \|\mathbf{v}\|_\infty^2 \|\mathbf{X}\|_2^2
\end{aligned}$$

Thus for multi-output neural networks, we have that

$$\|\mathcal{J}(\mathbf{W})\| \leq \sqrt{KkB^2}\|\mathbf{V}\|_\infty \|\mathbf{X}\|_2$$

We are omitting the proof of lipschitzness but will cover it if time permits. □

With this, we can apply the meta-theorem directly to multi-output neural networks by taking the NTK to be the reference Jacobian. One can prove that the NTK satisfies the assumptions of the reference Jacobian but we will omit the proof for simplicity and might discuss it if time permits.

3 Experiments

Authors use very recent methods to approximate the spectra of $J(\theta_\tau)J^\top(\theta_\tau)$. They perform experiments on CIFAR-10 and MNIST on ResNet20.

- The value of the top eigenvalues increases significantly, comparing the Jacobian at initialization with the Jacobian after training. In general, it is observed that the Jacobian is approximately low-rank, in the sense that it has a small set of large eigenvalues and then the rest are fairly small. This fits naturally with their theory, allowing to set a good cutoff for their bounds.

- They plot the norm of the projection of the residual onto the information and the nuisance space and observe that indeed, as predicted by the theory, the projection on the information space decreases much rapidly than the other one. Note that if training with some corrupted labels, and the residual corresponding to the noisy labels falls mostly into the nuisance space and analogously for the uncorrupted data and the information space, then the neural network would fit much faster the data that conveys information and this would have generalization implications at early stopping.

- They measured the norm of the projection of the labels and the residuals at initialization, but using the information and nuisance space of two Jacobians: the one given by the initialization and the one given by the trained network. For both the labels and the initial residual, the majority of the projection lies in the nuisance space for the Jacobian at initialization but for the other one the converse happens: the projected norm onto the information space is significantly bigger than the projection onto the nuisance space. Authors argue that this adaptation would in principle suggest a better generalization according to their theory (possibly using the arbitrary initialization theorem and initializing to the value of the Jacobian at time some iterations before stopping). It would also suggest that this adaptation of the Jacobian would speed up training. They also performed experiments with corrupted labels and saw that the projection onto the information space after training is not that large in that case. Also the normalized projection of the labels onto the nuisance space correlates with the test error.

References

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

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