

# Linearized Networks

Yilan Chen

August 26, 2020

This is some derivations in paper *Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent*[1].

## 1 Problem Setup

Training set:  $D \subseteq \mathbb{R}^{n_0 \times n_k}$ ,  $\mathcal{X} = \{x : (x, y) \in D\}$ ,  $\mathcal{Y} = \{y : (x, y) \in D\}$ .

Consider a fully-connected feed-forward network with  $L$  hidden layers with widths  $n_l$ , for  $l = 1, \dots, L$  and readout layer  $n_{L+1} = k$ .

$$\begin{cases} h^{l+1} = x^l W^{l+1} + b^{l+1} \\ x^{l+1} = \phi(x^l) \end{cases} \quad \begin{cases} W_{ij}^{l+1} = \frac{\sigma_{\omega}}{\sqrt{n_l}} w_{ij}^{l+1} \\ b_j^l = \sigma_b \beta_j^l \end{cases} \quad (1)$$

$$f_t(x) \equiv h^{L+1}(x) \in \mathbb{R}^k \quad (2)$$

where  $W^{l+1} \in \mathbb{R}^{n_l \times n_{l+1}}$ ,  $b^{l+1} \in \mathbb{R}^{n_{l+1}}$ ,  $w_{i,j}^l, \beta_j^l \sim \mathcal{N}(0, 1)$

$$\nabla_{\theta} f_t(\mathcal{X}) = \begin{bmatrix} \nabla_{\theta} f_1(x_1) \\ \vdots \\ \nabla_{\theta} f_k(x_1) \\ \vdots \\ \nabla_{\theta} f_1(x_{|D|}) \\ \vdots \\ \nabla_{\theta} f_k(x_{|D|}) \end{bmatrix} = \begin{bmatrix} \nabla_{\theta_1} f_1(x_1) & \cdots & \nabla_{\theta_{|\theta|}} f_1(x_1) \\ \vdots & \ddots & \vdots \\ \nabla_{\theta_1} f_k(x_1) & \cdots & \nabla_{\theta_{|\theta|}} f_k(x_1) \\ \vdots & \ddots & \vdots \\ \nabla_{\theta_1} f_1(x_{|D|}) & \cdots & \nabla_{\theta_{|\theta|}} f_1(x_{|D|}) \\ \vdots & \ddots & \vdots \\ \nabla_{\theta_1} f_k(x_{|D|}) & \cdots & \nabla_{\theta_{|\theta|}} f_k(x_{|D|}) \end{bmatrix} \in \mathbb{R}^{k|D| \times |\theta|} \quad (3)$$

where  $f_t(\mathcal{X}) = \text{vec}([f_t(x)]_{x \in \mathcal{X}}) \in \mathbb{R}^{k|D| \times 1}$ .

Empirical Tangent Kernel:

$$\begin{aligned} \hat{\Theta}_t &\equiv \hat{\Theta}_t(\mathcal{X}, \mathcal{X}) = \nabla_{\theta} f_t(\mathcal{X}) f_t(\mathcal{X})^T \\ &= \begin{bmatrix} \nabla_{\theta} f_1(x_1) \nabla_{\theta} f_1(x_1)^T & \cdots & \nabla_{\theta} f_1(x_1) \nabla_{\theta} f_k(x_1)^T & \cdots & \nabla_{\theta} f_1(x_1) \nabla_{\theta} f_1(x_{|D|})^T & \cdots & \nabla_{\theta} f_1(x_1) \nabla_{\theta} f_k(x_{|D|})^T \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \nabla_{\theta} f_k(x_1) \nabla_{\theta} f_1(x_1)^T & \cdots & \nabla_{\theta} f_k(x_1) \nabla_{\theta} f_k(x_1)^T & \cdots & \nabla_{\theta} f_k(x_1) \nabla_{\theta} f_1(x_{|D|})^T & \cdots & \nabla_{\theta} f_k(x_1) \nabla_{\theta} f_k(x_{|D|})^T \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \nabla_{\theta} f_1(x_{|D|}) \nabla_{\theta} f_1(x_1)^T & \cdots & \nabla_{\theta} f_1(x_{|D|}) \nabla_{\theta} f_k(x_1)^T & \cdots & \nabla_{\theta} f_1(x_{|D|}) \nabla_{\theta} f_1(x_{|D|})^T & \cdots & \nabla_{\theta} f_1(x_{|D|}) \nabla_{\theta} f_k(x_{|D|})^T \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \nabla_{\theta} f_k(x_{|D|}) \nabla_{\theta} f_1(x_1)^T & \cdots & \nabla_{\theta} f_k(x_{|D|}) \nabla_{\theta} f_k(x_1)^T & \cdots & \nabla_{\theta} f_k(x_{|D|}) \nabla_{\theta} f_1(x_{|D|})^T & \cdots & \nabla_{\theta} f_k(x_{|D|}) \nabla_{\theta} f_k(x_{|D|})^T \end{bmatrix} \\ &= \begin{bmatrix} \hat{\Theta}_t(x_1, \mathcal{X}) \\ \vdots \\ \hat{\Theta}_t(x_{|D|}, \mathcal{X}) \end{bmatrix} \in \mathbb{R}^{k|D| \times k|D|} \end{aligned} \quad (4)$$

Continuous time gradient descent:

$$\dot{\theta}_t = -\eta \nabla_{\theta} f_t(\mathcal{X})^T \nabla_{f_t(\mathcal{X})} \mathcal{L} \in \mathbb{R}^{|\theta| \times 1} \quad (5)$$

$$\dot{f}_t(\mathcal{X}) = \nabla_{\theta} f_t(\mathcal{X}) \dot{\theta}_t = -\eta \nabla_{\theta} f_t(\mathcal{X}) \nabla_{\theta} f_t(\mathcal{X})^T \nabla_{f_t(\mathcal{X})} \mathcal{L} = -\eta \hat{\Theta}_t(\mathcal{X}, \mathcal{X}) \nabla_{f_t(\mathcal{X})} \mathcal{L} \quad (6)$$

In the case of an MSE loss, i.e.,  $\ell(\hat{y}, y) = \frac{1}{2} \|\hat{y} - y\|_2^2$ ,

$$\nabla_{f_t(\mathcal{X})} \mathcal{L} = f_t(\mathcal{X}) - \mathcal{Y} \in \mathbb{R}^{k|D| \times 1} \quad (7)$$

$$\dot{f}_t(\mathcal{X}) = -\eta \hat{\Theta}_t(\mathcal{X}, \mathcal{X}) (f_t(\mathcal{X}) - \mathcal{Y}) \quad (8)$$

which a first-order system of linear differential equations or a matrix differential equation.

When  $\hat{\Theta}_t(\mathcal{X}, \mathcal{X}) \rightarrow \Theta$  is a constant matrix, the solution is given by

$$\begin{aligned} f_t(\mathcal{X}) &= e^{-\eta \Theta t} (f_0(\mathcal{X}) - \mathcal{Y}) + \mathcal{Y} \\ &= (I - e^{-\eta \Theta t}) \mathcal{Y} + e^{-\eta \Theta t} f_0(\mathcal{X}) \end{aligned} \quad (9)$$

where  $e^X = \sum_{k=0}^{\infty} \frac{1}{k!} X^k$ . There is no closed-form solution for the differential equations when  $\hat{\Theta}_t(\mathcal{X}, \mathcal{X})$  is varying with time, one has to use either a numerical method, or an approximation method such as Magnus expansion.

## 2 Linearized Model

$$f_t^{lin}(x) \equiv f_0(x) + \nabla_{\theta} f_0(x)|_{\theta=\theta_0} \omega_t, \quad \omega_t \equiv \theta_t - \theta_0 \quad (10)$$

$$\nabla_{\theta} f_t^{lin}(x) = 0 + \nabla_{\theta} f_0(x)|_{\theta=\theta_0} = \nabla_{\theta} f_0(x) \quad (11)$$

$$\begin{aligned} \dot{\omega}_t &= \dot{\theta}_t = -\eta \nabla_{\theta} f_t^{lin}(\mathcal{X})^T \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L} \\ &= -\eta \nabla_{\theta} f_0(\mathcal{X})^T \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L} \end{aligned} \quad (12)$$

$$\begin{aligned} \dot{f}_t^{lin}(\mathcal{X}) &= \nabla_{\theta} f_t^{lin}(\mathcal{X}) \dot{\theta}_t = -\eta \nabla_{\theta} f_t^{lin}(\mathcal{X}) \nabla_{f_t^{lin}(\mathcal{X})}^T \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L} \\ &= -\eta \nabla_{\theta} f_0(x) \nabla_{\theta} f_0(x)^T \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L} = -\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) \nabla_{f_t^{lin}(\mathcal{X})} \mathcal{L} \end{aligned} \quad (13)$$

Note that (13) is identical to (6) if tangent kernel is deterministic during training, i.e.  $\hat{\Theta}_t(\mathcal{X}, \mathcal{X}) = \hat{\Theta}_0(\mathcal{X}, \mathcal{X})$ .

In the case of an MSE loss,

$$\dot{f}_t^{lin}(\mathcal{X}) = -\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) (f_t^{lin}(\mathcal{X}) - \mathcal{Y}) \quad (14)$$

Since  $\hat{\Theta}_0(\mathcal{X}, \mathcal{X})$  is a constant matrix, (14) has a solution like (9)

$$f_t^{lin}(\mathcal{X}) = e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t} (f_0(\mathcal{X}) - \mathcal{Y}) + \mathcal{Y} \quad (15)$$

Take this into equation (12)

$$\begin{aligned} \dot{\omega}_t &= -\eta \nabla_{\theta} f_0(\mathcal{X})^T (f_t^{lin}(\mathcal{X}) - \mathcal{Y}) \\ &= -\eta \nabla_{\theta} f_0(\mathcal{X})^T e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t} (f_0(\mathcal{X}) - \mathcal{Y}) \end{aligned} \quad (16)$$

Integrate over time,

$$\begin{aligned} \int_0^t e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t} dt &= \int_0^t \sum_{k=0}^{\infty} \frac{1}{k!} (-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t)^k dt \\ &= \sum_{k=0}^{\infty} \frac{1}{k!} (-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t)^k \frac{t}{k+1} \\ &= -\frac{1}{\eta} \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \sum_{k=0}^{\infty} \frac{1}{(k+1)!} (-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t)^{k+1} \\ &= -\frac{1}{\eta} \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \sum_{k=1}^{\infty} \frac{1}{k!} (-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t)^k \\ &= -\frac{1}{\eta} \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} (e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t} - I) \\ &= \frac{1}{\eta} \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} (I - e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t}) \end{aligned} \quad (17)$$

$$\omega_t = -\nabla_{\theta} f_0(\mathcal{X})^T \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X})t} \right) (f_0(\mathcal{X}) - \mathcal{Y}) \quad (18)$$

Take this into equation (10), we get

$$\begin{aligned} f_t^{lin}(x) &= f_0(x) + \nabla_{\theta} f_0(x)|_{\theta=\theta_0} \omega_t \\ &= f_0(x) - \nabla_{\theta} f_0(x) \nabla_{\theta} f_0(\mathcal{X})^T \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X})t} \right) (f_0(\mathcal{X}) - \mathcal{Y}) \\ &= f_0(x) - \hat{\Theta}_0(x, \mathcal{X}) \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X})t} \right) (f_0(\mathcal{X}) - \mathcal{Y}) \\ &= \hat{\Theta}_0(x, \mathcal{X}) \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X})t} \right) \mathcal{Y} \\ &\quad + f_0(x) - \hat{\Theta}_0(x, \mathcal{X}) \hat{\Theta}_0(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X})t} \right) f_0(\mathcal{X}) \end{aligned} \quad (19)$$

### 3 Infinite width limit yields a Gaussian process

Let

$$\mathcal{K}^{i,j}(x, x') = \lim_{\min(n_1, \dots, n_L) \rightarrow \infty} \mathbb{E}[f_0^i(x) f_0^j(x')] \quad (20)$$

denotes the covariance between the  $i$ -th output of  $x$  and  $j$ -th output of  $x'$  at initialization. Then  $f_0(\mathcal{X}) \sim \mathcal{N}(0, \mathcal{K}(\mathcal{X}, \mathcal{X}))$ , i.e.

$$p(f(\mathcal{X})|\mathcal{X}) = \mathcal{N}(0, \mathcal{K}(\mathcal{X}, \mathcal{X})) \quad (21)$$

For a test input  $x \in \mathcal{X}_T$ , the joint output distribution  $f([x, \mathcal{X}])$  is also multivariate Gaussian, whose covariance matrix is

$$\mathbf{K} = \begin{bmatrix} \mathcal{K}(\mathcal{X}, \mathcal{X}) & \mathcal{K}(x, \mathcal{X})^T \\ \mathcal{K}(x, \mathcal{X}) & \mathcal{K}(x, x) \end{bmatrix} \quad (22)$$

That is

$$p(f(x), f(\mathcal{X})|x, \mathcal{X}) = \mathcal{N}(0, \mathbf{K}) \quad (23)$$

We have to predict the probability distribution of  $f(x)$ , conditioning on the training samples,  $f(\mathcal{X}) = \mathcal{Y}$ , i.e.  $p(f(\mathcal{X}) = \mathcal{Y}|\mathcal{X}, \mathcal{Y}) = 1$ .

$$\begin{aligned} p(f(x)|x, \mathcal{X}, \mathcal{Y}) &= p(f(x)|x, \mathcal{X}, f(\mathcal{X})) \\ &= \frac{p(f(x), x, \mathcal{X}, f(\mathcal{X}))}{p(x, \mathcal{X}, f(\mathcal{X}))} \\ &= \frac{p(f(x), f(\mathcal{X})|x, \mathcal{X}) p(x, \mathcal{X})}{p(x, \mathcal{X}, f(\mathcal{X}))} \\ &= \frac{p(f(x), f(\mathcal{X})|x, \mathcal{X})}{p(f(\mathcal{X})|x, \mathcal{X})} \\ &= \frac{p(f(x), f(\mathcal{X})|x, \mathcal{X})}{p(f(\mathcal{X})|\mathcal{X})} \\ &= \frac{\mathcal{N}(0, \mathbf{K})}{\mathcal{N}(0, \mathcal{K}(\mathcal{X}, \mathcal{X}))} \end{aligned} \quad (24)$$

This is also a gaussian distribution given by

$$\mu(x) = \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \mathcal{Y} \quad (25)$$

$$\Sigma(x) = \mathcal{K}(x, x) - \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \mathcal{K}(\mathcal{X}, x)^T \quad (26)$$

This corresponds to only the readout layer  $n_{L+1} = k$  is being trained (see [1] appendix D).

### 4 Corollary 1

In the infinite width setting,  $[f_0(x), f_0(\mathcal{X})]$  is Gaussian distributed and  $\hat{\Theta}_0$  converges in probability to a deterministic kernel  $\Theta$ . For any  $t$ , (19) is Gaussian distributed because it describe an affine transform of the Gaussian  $[f_0(x), f_0(\mathcal{X})]$ .

$$\begin{aligned} \mathbb{E}[f_t^{lin}(x)] &= \mathbb{E}[\Theta(x, \mathcal{X}) \Theta(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta \Theta(\mathcal{X}, \mathcal{X})t} \right) \mathcal{Y}] \\ &\quad + \mathbb{E}[f_0(x)] - \Theta(x, \mathcal{X}) \Theta(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta \Theta(\mathcal{X}, \mathcal{X})t} \right) \mathbb{E}[f_0(\mathcal{X})] \\ &= \Theta(x, \mathcal{X}) \Theta(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta \Theta(\mathcal{X}, \mathcal{X})t} \right) \mathcal{Y} \end{aligned} \quad (27)$$

$$\begin{aligned}
\Sigma(x, x) &= \mathbb{E}[f_t^{lin}(x) f_t^{lin}(x)^T] \\
&= \mathbb{E}[[f_0(x) - \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})(f_0(\mathcal{X}) - \mathcal{Y})] \\
&\quad [f_0(x) - \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})(f_0(\mathcal{X}) - \mathcal{Y})]^T] \\
&= \mathbb{E}[[f_0(x) - \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})(f_0(\mathcal{X}) - \mathcal{Y})] \\
&\quad [f_0(x)^T - (f_0(\mathcal{X}) - \mathcal{Y})^T(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\Theta(\mathcal{X}, \mathcal{X})^{-1}\Theta(\mathcal{X}, x)]] \\
&= \mathbb{E}[f_0(x)f_0(x)^T - \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})(f_0(\mathcal{X}) - \mathcal{Y})f_0(x)^T \\
&\quad - f_0(x)(f_0(\mathcal{X}) - \mathcal{Y})^T(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\Theta(\mathcal{X}, \mathcal{X})^{-1}\Theta(\mathcal{X}, x) \\
&\quad + \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})(f_0(\mathcal{X}) - \mathcal{Y})(f_0(\mathcal{X}) - \mathcal{Y})^T(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\Theta(\mathcal{X}, \mathcal{X})^{-1}\Theta(\mathcal{X}, x)] \\
&= \mathcal{K}(x, x) - \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\mathcal{K}(\mathcal{X}, x) - \mathcal{K}(x, \mathcal{X})(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\Theta(\mathcal{X}, \mathcal{X})^{-1}\Theta(\mathcal{X}, x) \\
&\quad + \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\mathcal{K}(\mathcal{X}, \mathcal{X})(I - e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t})\Theta(\mathcal{X}, \mathcal{X})^{-1}\Theta(\mathcal{X}, x)
\end{aligned} \tag{28}$$

When  $t \rightarrow \infty$ ,  $e^{-\eta\Theta(\mathcal{X}, \mathcal{X})t} \rightarrow \mathbf{0}$ ,

$$\mathbb{E}[f_t^{lin}(x)] = \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}\mathcal{Y} \tag{29}$$

$$\begin{aligned}
\Sigma(x, x) &= \mathcal{K}(x, x) - \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}\mathcal{K}(\mathcal{X}, x) - \mathcal{K}(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}\Theta(\mathcal{X}, x) \\
&\quad + \Theta(x, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}\mathcal{K}(\mathcal{X}, \mathcal{X})\Theta(\mathcal{X}, \mathcal{X})^{-1}\Theta(\mathcal{X}, x)
\end{aligned} \tag{30}$$

## 5 Gradient flow dynamics for training only the readout-layer

This is appendix D in paper[1].

$$f(x) = \bar{x}(x)\theta^{L+1} = [\bar{x}(x) \cdot \theta_1^{L+1}, \dots, \bar{x}(x) \cdot \theta_i^{L+1}, \dots, \bar{x}(x) \cdot \theta_k^{L+1}] \tag{31}$$

$$\bar{x}(x) = [\frac{\sigma_\omega x^L(x)}{\sqrt{n_l}}, \sigma_b] \in \mathbb{R}^{n_L+1}, \quad \theta^{L+1} = \begin{bmatrix} W^{L+1} \\ b^{L+1} \end{bmatrix} \in \mathbb{R}^{(n_L+1) \times k} \tag{32}$$

where  $\theta_i^{L+1}$  is the  $i_{th}$  column of  $\theta^{L+1}$ . Note  $n_{L+1} = k$ .

$$\begin{aligned}
\hat{\mathcal{K}}(x, x') &= \mathbb{E}[f_0^i(x) \cdot f_0^i(x')] = \mathbb{E}[\bar{x}(x) \cdot \theta_i^{L+1} \cdot \bar{x}(x') \cdot \theta_i^{L+1}] \\
&= \bar{x}(x) \cdot \bar{x}(x') = \frac{\sigma_\omega^2}{n_l} x^L(x) \cdot x^L(x') + \sigma_b^2 \rightarrow \mathcal{K}(x, x')
\end{aligned} \tag{33}$$

$$\nabla_{\theta^{L+1}} f(x) = \begin{bmatrix} \nabla_{\theta^{L+1}} f_1(x) \\ \vdots \\ \nabla_{\theta^{L+1}} f_k(x) \end{bmatrix} = \begin{bmatrix} \bar{x}(x), & \mathbf{0}, & \dots, & \mathbf{0}, & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}, & \mathbf{0}, & \dots, & \mathbf{0}, & \bar{x}(x) \end{bmatrix} \in \mathbb{R}^{k \times (n_L+1)k} \tag{34}$$

$$\begin{aligned}
\nabla_{\theta^{L+1}} f(\mathcal{X}) &= \begin{bmatrix} \nabla_{\theta^{L+1}} f_1(x_1) \\ \vdots \\ \nabla_{\theta^{L+1}} f_k(x_1) \\ \vdots \\ \nabla_{\theta^{L+1}} f_1(x_{|D|}) \\ \vdots \\ \nabla_{\theta^{L+1}} f_k(x_{|D|}) \end{bmatrix} = \begin{bmatrix} \bar{x}(x_1), & \mathbf{0}, & \dots, & \mathbf{0}, & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}, & \mathbf{0}, & \dots, & \mathbf{0}, & \bar{x}(x_1) \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \bar{x}(x_{|D|}), & \mathbf{0}, & \dots, & \mathbf{0}, & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0}, & \mathbf{0}, & \dots, & \mathbf{0}, & \bar{x}(x_{|D|}) \end{bmatrix} \in \mathbb{R}^{k|D| \times (n_L+1)k}
\end{aligned} \tag{35}$$

In the case of MSE loss,

$$\mathcal{L} = \frac{1}{2} \|f(\mathcal{X}) - \mathcal{Y}\|_2^2 = \frac{1}{2} \|\bar{x}(\mathcal{X})\theta^{L+1} - \mathcal{Y}\|_2^2 \tag{36}$$

$$\begin{aligned}
\dot{\theta}^{L+1} &= -\eta \nabla_{\theta^{L+1}} f(\mathcal{X})^T \nabla_{f_t(\mathcal{X})} \mathcal{L} = -\eta \nabla_{\theta^{L+1}} f(\mathcal{X})^T (\bar{x}(\mathcal{X}) \theta^{L+1} - \mathcal{Y}) \\
&= -\eta \begin{bmatrix} \bar{x}(x_1)^T, & \cdots & \mathbf{0}, & \cdots, & \bar{x}(x_{|D|})^T, & \cdots & \mathbf{0} \\ \vdots & \ddots & \vdots & \ddots & \vdots & \ddots & \vdots \\ \mathbf{0}, & \cdots & \bar{x}(x_1)^T, & \cdots, & \mathbf{0}, & \cdots & \bar{x}(x_{|D|})^T \end{bmatrix} \begin{bmatrix} \bar{x}(x_1) \theta_1^{L+1} - y_{1,1} \\ \vdots \\ \bar{x}(x_1) \theta_k^{L+1} - y_{1,k} \\ \vdots \\ \bar{x}(x_{|D|}) \theta_1^{L+1} - y_{|D|,1} \\ \vdots \\ \bar{x}(x_{|D|}) \theta_k^{L+1} - y_{|D|,k} \end{bmatrix} \quad (37)
\end{aligned}$$

Since different dimensions of an output are independent, we consider only one output, i.e.  $k = 1$ , where we have

$$\nabla_{\theta^{L+1}} f(\mathcal{X}) = \bar{x}(\mathcal{X}) \in \mathbb{R}^{n_L+1} \quad (38)$$

$$\begin{aligned}
\dot{\theta}^{L+1} &= -\eta \nabla_{\theta^{L+1}} f(\mathcal{X})^T \nabla_{f_t(\mathcal{X})} \mathcal{L} \\
&= -\eta \bar{x}(\mathcal{X})^T (\bar{x}(\mathcal{X}) \theta^{L+1} - \mathcal{Y}) \quad (39)
\end{aligned}$$

Since  $\bar{x}(\mathcal{X})^T \bar{x}(\mathcal{X})$  is a constant matrix, we can get the result as

$$\bar{x}(\mathcal{X}) \theta_t^{L+1} = e^{-\eta \bar{x}(\mathcal{X})^T \bar{x}(\mathcal{X}) t} (\bar{x}(\mathcal{X}) \theta_0^{L+1} - \mathcal{Y}) + \mathcal{Y} \quad (40)$$

$$f_t(\mathcal{X}) = e^{-\eta \bar{x}(\mathcal{X})^T \bar{x}(\mathcal{X}) t} (f_0(\mathcal{X}) - \mathcal{Y}) + \mathcal{Y} \quad (41)$$

When we only train the readout-layer, the original network and its linearization are identical. Compare with (15),

$$\begin{aligned}
f_t(\mathcal{X}) &= e^{-\eta \bar{x}(\mathcal{X})^T \bar{x}(\mathcal{X}) t} (f_0(\mathcal{X}) - \mathcal{Y}) + \mathcal{Y} \\
&= e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t} (f_0(\mathcal{X}) - \mathcal{Y}) + \mathcal{Y} \quad (42)
\end{aligned}$$

We can see

$$\bar{x}(\mathcal{X})^T \bar{x}(\mathcal{X}) = \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) = \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X}) \quad (43)$$

Take this into (19)

$$f_t(x) = f_0(x) - \hat{\mathcal{K}}(x, \mathcal{X}) \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1} (I - e^{-\eta \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X}) t}) (f_0(\mathcal{X}) - \mathcal{Y}) \quad (44)$$

In the infinite width setting,  $\hat{\mathcal{K}} \rightarrow \mathcal{K}$ , and  $\Theta = \mathcal{K}$ . Take this into the results of Corollary 1,

$$\mathbb{E}[f_t(x)] = \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{Y} \quad (45)$$

$$\begin{aligned}
\Sigma(x, x) &= \mathcal{K}(x, x) - \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, x) - \mathcal{K}(x, \mathcal{X}) (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \mathcal{K}(\mathcal{X}, x) \\
&\quad + \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, \mathcal{X}) (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \mathcal{K}(\mathcal{X}, x) \\
&= \mathcal{K}(x, x) - 2\mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, x) \\
&\quad + \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, \mathcal{X}) (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \mathcal{K}(\mathcal{X}, x) \quad (46)
\end{aligned}$$

where

$$\begin{aligned}
&(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, \mathcal{X}) (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \\
&= (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \mathcal{K}(\mathcal{X}, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \\
&= (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) (I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}) \\
&= I - 2e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t} + e^{-2\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t} \quad (47)
\end{aligned}$$

Take this into the equation,

$$\begin{aligned}
\Sigma(x, x) &= \mathcal{K}(x, x) - 2\mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta\mathcal{K}(\mathcal{X}, \mathcal{X})t} \right) \mathcal{K}(\mathcal{X}, x) \\
&\quad + \mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left( I - 2e^{-\eta\mathcal{K}(\mathcal{X}, \mathcal{X})t} + e^{-2\eta\mathcal{K}(\mathcal{X}, \mathcal{X})t} \right) \mathcal{K}(\mathcal{X}, x) \\
&= \mathcal{K}(x, x) + \mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left( e^{-\eta\mathcal{K}(\mathcal{X}, \mathcal{X})t} - I \right) \mathcal{K}(\mathcal{X}, x) \\
&= \mathcal{K}(x, x) - \mathcal{K}(x, \mathcal{X})\mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left( I - e^{-\eta\mathcal{K}(\mathcal{X}, \mathcal{X})t} \right) \mathcal{K}(x, \mathcal{X})^T
\end{aligned} \tag{48}$$

## 6 Infinite width networks are linearized networks

**Theorem 2.1** (informal). *Let  $n_1 = \dots = n_L = n$  and assume  $\lambda_{\min}(\Theta) > 0$ . Applying gradient descent with learning rate  $\eta < \eta_{\text{critical}}$  (or gradient flow), for every  $x \in \mathbb{R}^{n_0}$  with  $\|x\|_2 \leq 1$ , with probability arbitrarily close to 1 over random initialization,*

$$\sup_{t \geq 0} \|f_t(x) - f_t^{\text{lin}}(x)\|_2, \sup_{t \geq 0} \frac{\|\theta_t - \theta_0\|_2}{\sqrt{n}}, \sup_{t \geq 0} \|\hat{\Theta}_t - \hat{\Theta}_0\|_F = \mathcal{O}(n^{-\frac{1}{2}}), \text{ as } n \rightarrow \infty \tag{49}$$

Some short-hand notations:

$$f(\theta_t) = f(\mathcal{X}, \theta_t) \in \mathbb{R}^{|\mathcal{X}|^k} \tag{50}$$

$$g(\theta_t) = g(\mathcal{X}, \theta_t) - \mathcal{Y} \in \mathbb{R}^{|\mathcal{X}|^k} \tag{51}$$

$$J(\theta_t) = \nabla f(\theta_t) \in \mathbb{R}^{(|\mathcal{X}|^k) \times |\theta|} \tag{52}$$

$$\begin{cases} \hat{\Theta}_t := \hat{\Theta}_t(\mathcal{X}, \mathcal{X}) = \frac{1}{n} J(\theta_t) J(\theta_t)^T \\ \Theta := \lim_{n \rightarrow \infty} \hat{\Theta}_0 \text{ in probability.} \end{cases} \tag{53}$$

The gradient descent update with learning rate  $\eta$ :

$$\theta_{t+1} = \theta_t - \eta J(\theta_t)^T g(\theta_t) \tag{54}$$

**Lemma 1 (Local Lipschitzness of the Jacobian).** *There is a  $K > 0$  such that for every  $C > 0$ , with high probability over random initialization (w.h.p.o.r.i.) the following holds*

$$\begin{cases} \frac{1}{\sqrt{n}} \|J(\theta) - J(\tilde{\theta})\|_F \leq K \|\theta - \tilde{\theta}\|_2 \\ \frac{1}{\sqrt{n}} \|J(\theta)\|_F \leq K \end{cases}, \quad \forall \theta, \tilde{\theta} \in B(\theta_0, Cn^{-\frac{1}{2}}) \tag{55}$$

where

$$B(\theta_0, R) := \{\theta : \|\theta - \theta_0\|_2 < R\} \tag{56}$$

**Theorem G.1 (Gradient descent).** *Assume Assumptions [1-4]. For  $\delta_0 > 0$  and  $\eta_0 < \eta_{\text{critical}}$ , there exist  $R_0 > 0$ ,  $N \in \mathbb{N}$  and  $K > 1$ , such that for every  $n \geq N$ , the following holds with probability at least  $(1 - \delta_0)$  over random initialization when applying gradient descent with learning rate  $\eta = \frac{\eta_0}{n}$ ,*

$$\begin{cases} \|g(\theta_t)\|_2 \leq (1 - \frac{\eta_0 \lambda_{\min}}{3})^t R_0 \\ \sum_{j=1}^t \|\theta_j - \theta_{j-1}\|_2 \leq \frac{\eta_0 K R_0}{\sqrt{n}} \sum_{j=1}^t (1 - \frac{\eta_0 \lambda_{\min}}{3})^{j-1} \leq \frac{3K R_0}{\lambda_{\min}} n^{-\frac{1}{2}} \end{cases} \tag{57}$$

and

$$\sup_t \|\hat{\Theta}_0 - \hat{\Theta}_t\|_F \leq \frac{6K^3 R_0}{\lambda_{\min}} n^{-\frac{1}{2}} \tag{58}$$

The first inequation of (57) indicates when  $t \rightarrow \infty$ ,  $g(\theta_t) \rightarrow 0$ , i.e. the convergence of training. The second inequation of (57) bounds the change of  $\theta$  with  $n$ . The larger  $n$  is, the less  $\theta$  changes. The inequation of (58) bounds the change of  $\hat{\Theta}$  with  $n$  during training.

## 6.1 Proof of Theorem G.1

We first prove (57) by induction.

Since  $f(x_0)$  and  $g(x_0)$  are gaussian distributed, for arbitrarily small  $\delta_0 > 0$ , there exist  $R_0$  and  $n_0$  (both may depend on  $\delta_0, |\mathcal{X}|$  and  $\mathcal{K}$ ) such that for every  $n \geq n_0$ , with probability at least  $(1 - \delta_0)$  over random initialization,

$$\|g(\theta_0)\|_2 < R_0 \quad (59)$$

This is the case of  $t = 0$  for (57). Assume (57) holds for  $t = t$ . Then for  $t + 1$

$$\|\theta_{t+1} - \theta_t\|_2 = \|-\eta J(\theta_t)^T g(\theta_t)\|_2 \leq \eta \|J(\theta_t)\|_{op} \|g(\theta_t)\|_2 \leq \frac{\eta_0}{n} \|J(\theta_t)\|_{op} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^t R_0 \quad (60)$$

Here the  $\|\cdot\|_{op}$  is the induced 2-norm for a matrix. From Lemma 1 and the property of matrix norm,

$$\|J(\theta_t)\|_2 = \sigma_{max}(J(\theta_t)) \leq \|J(\theta_t)\|_F \leq K\sqrt{n} \quad (61)$$

So we get

$$\|\theta_{t+1} - \theta_t\|_2 \leq \frac{K\eta_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^t R_0 \quad (62)$$

Then

$$\begin{aligned} \|\theta_{t+1} - \theta_0\|_2 &= \|\theta_{t+1} - \theta_t + \theta_t - \theta_{t-1} + \dots - \theta_0\|_2 \\ &\leq \|\theta_{t+1} - \theta_t\|_2 + \dots + \|\theta_1 - \theta_0\|_2 \\ &\leq \sum_{j=1}^{t+1} \frac{K\eta_0 R_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{j-1} \end{aligned} \quad (63)$$

which is the sum of a geometric progression,

$$\sum_{j=1}^{t+1} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{j-1} = \frac{1 - \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{t+1}}{1 - \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)} = 3 \frac{1 - \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{t+1}}{\eta_0 \lambda_{min}} \leq \frac{3}{\eta_0 \lambda_{min}} \quad (64)$$

So we get the second inequation of (57).

$$\|\theta_{t+1} - \theta_0\|_2 \leq \sum_{j=1}^{t+1} \|\theta_j - \theta_{j-1}\|_2 \leq \sum_{j=1}^{t+1} \frac{K\eta_0 R_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{j-1} \leq \frac{K\eta_0 R_0}{\sqrt{n}} \frac{3}{\eta_0 \lambda_{min}} = \frac{3K R_0}{\lambda_{min}} n^{-\frac{1}{2}} \quad (65)$$

For the first inequation of (57), we apply the mean value theorem.

$$\begin{aligned} \|g(\theta_{t+1})\|_2 &= \|g(\theta_{t+1}) - g(\theta_t) + g(\theta_t)\|_2 \\ &= \|J(\tilde{\theta}_t)(\theta_{t+1} - \theta_t) + g(\theta_t)\|_2 \\ &= \|- \eta J(\tilde{\theta}_t) J(\theta_t)^T g(\theta_t) + g(\theta_t)\|_2 \\ &= \|(I - \eta J(\tilde{\theta}_t) J(\theta_t)^T) g(\theta_t)\|_2 \\ &\leq \|I - \eta J(\tilde{\theta}_t) J(\theta_t)^T\|_{op} \|g(\theta_t)\|_2 \\ &\leq \|I - \eta J(\tilde{\theta}_t) J(\theta_t)^T\|_{op} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^t R_0 \end{aligned} \quad (66)$$

where  $\tilde{\theta}_t$  is some linear interpolation between  $\theta_t$  and  $\theta_{t+1}$ .

$$\begin{aligned} \|I - \eta J(\tilde{\theta}_t) J(\theta_t)^T\|_{op} &= \|I - \eta J(\theta_0) J(\theta_0)^T + \eta J(\theta_0) J(\theta_0)^T - \eta J(\tilde{\theta}_t) J(\theta_t)^T\|_{op} \\ &= \|I - \eta_0 \Theta_0 + \eta J(\theta_0) J(\theta_0)^T - \eta J(\tilde{\theta}_t) J(\theta_t)^T\|_{op} \\ &= \|I - \eta_0 \Theta + \eta_0 \Theta - \eta_0 \Theta_0 + \eta J(\theta_0) J(\theta_0)^T - \eta J(\tilde{\theta}_t) J(\theta_t)^T\|_{op} \\ &\leq \|I - \eta_0 \Theta\|_{op} + \eta_0 \|\Theta - \Theta_0\|_{op} + \eta \|J(\theta_0) J(\theta_0)^T - J(\tilde{\theta}_t) J(\theta_t)^T\|_{op} \end{aligned} \quad (67)$$

The assumption  $\Theta$  is full-rank and  $\lambda_{min} > 0$  implies

$$\|I - \eta_0 \Theta\|_{op} = \sigma_{max}(I - \eta_0 \Theta) = 1 - \eta_0 \sigma_{min}(\Theta) = 1 - \eta_0 \lambda_{min} \quad (68)$$

Because  $\hat{\Theta}_0 \rightarrow \Theta$  in probability, one can find  $n_2$  such that

$$\eta_0 \left\| \Theta - \hat{\Theta}_0 \right\|_{op} \leq \eta_0 \left\| \Theta - \hat{\Theta}_0 \right\|_F \leq \frac{\eta_0 \lambda_{min}}{3} \quad (69)$$

For the third part

$$\begin{aligned} \left\| J(\theta_0)J(\theta_0)^T - J(\tilde{\theta}_t)J(\theta_t)^T \right\|_{op} &= \left\| J(\theta_0)J(\theta_0)^T - J(\theta_0)J(\theta_t)^T + J(\theta_0)J(\theta_t)^T - J(\tilde{\theta}_t)J(\theta_t)^T \right\|_{op} \\ &= \left\| J(\theta_0)[J(\theta_0)^T - J(\theta_t)^T] + [J(\theta_0) - J(\tilde{\theta}_t)]J(\theta_t)^T \right\|_{op} \\ &\leq \|J(\theta_0)\|_{op} \|J(\theta_0)^T - J(\theta_t)^T\|_{op} + \|J(\theta_0) - J(\tilde{\theta}_t)\|_{op} \|J(\theta_t)^T\|_{op} \\ &\leq \|J(\theta_0)\|_F \|J(\theta_0)^T - J(\theta_t)^T\|_F + \|J(\theta_0) - J(\tilde{\theta}_t)\|_F \|J(\theta_t)^T\|_F \quad (70) \\ &\leq K\sqrt{n}K\sqrt{n} \|\theta_0 - \theta_t\|_2 + K\sqrt{n}K\sqrt{n} \|\theta_0 - \tilde{\theta}_t\|_2 \\ &= K^2 n \|\theta_t - \theta_0\|_2 + K^2 n \|\tilde{\theta}_t - \theta_0\|_2 \\ &\leq 2K^2 n \frac{3KR_0}{\lambda_{min}} n^{-\frac{1}{2}} \end{aligned}$$

Take the three parts together and when  $n \geq \left(\frac{18K^3 R_0}{\lambda_{min}}\right)^2$ ,

$$\begin{aligned} \left\| 1 - \eta J(\tilde{\theta}_t)J(\theta_t)^T \right\|_{op} &\leq 1 - \eta_0 \lambda_{min} + \frac{\eta_0 \lambda_{min}}{3} + 2\eta_0 K^2 \frac{3KR_0}{\lambda_{min}} n^{-\frac{1}{2}} \\ &\leq 1 - \frac{\eta_0 \lambda_{min}}{3} \end{aligned} \quad (71)$$

Take this into (66),

$$\|g(\theta_{t+1})\|_2 \leq \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{t+1} R_0 \quad (72)$$

For (58),

$$\begin{aligned} \left\| \hat{\Theta}_0 - \hat{\Theta}_t \right\|_F &= \frac{1}{n} \left\| J(\theta_0)J(\theta_0)^T - J(\theta_t)J(\theta_t)^T \right\|_F \\ &= \frac{1}{n} \left\| J(\theta_0)J(\theta_0)^T - J(\theta_0)J(\theta_t)^T + J(\theta_0)J(\theta_t)^T - J(\theta_t)J(\theta_t)^T \right\|_F \\ &= \frac{1}{n} \left\| J(\theta_0)[J(\theta_0)^T - J(\theta_t)^T] + [J(\theta_0) - J(\theta_t)]J(\theta_t)^T \right\|_F \quad (73) \\ &\leq \frac{1}{n} \|J(\theta_0)\|_F \|J(\theta_0)^T - J(\theta_t)^T\|_F + \frac{1}{n} \|J(\theta_0) - J(\theta_t)\|_F \|J(\theta_t)^T\|_F \\ &\leq K^2 \|\theta_t - \theta_0\|_2 + K^2 \|\theta_t - \theta_0\|_2 \\ &\leq \frac{6K^3 R_0}{\lambda_{min}} n^{-\frac{1}{2}} \end{aligned}$$

## 6.2 Bounding $\|f_t(x) - f_t^{lin}(x)\|_2$

To simplify the notation, let  $g^{lin}(t) = f_t^{lin}(\mathcal{X}) - \mathcal{Y}$  and  $g(t) = f_t(\mathcal{X}) - \mathcal{Y}$ .

**Theorem H.1.** *Same as in Theorem G.2. For every  $x \in \mathbb{R}^{n_0}$  with  $\|x\|_2 \leq 1$ , for  $\delta_0 > 0$  arbitrarily small, there exist  $R_0 > 0$  and  $N \in \mathbb{N}$  such that for every  $n \geq N$ , with probability at least  $(1 - \delta_0)$  over random initialization,*

$$\sup_t \|g^{lin}(t) - g(t)\|_2, \quad \sup_t \|g^{lin}(t, x) - g(t, x)\|_2 \lesssim n^{-\frac{1}{2}} R_0^2 \quad (74)$$

*Proof.*

Recall

$$\dot{f}_t(\mathcal{X}) = -\eta \hat{\Theta}_t(\mathcal{X}, \mathcal{X})(f_t(\mathcal{X}) - \mathcal{Y}) \quad (75)$$

$$\dot{f}_t^{lin}(\mathcal{X}) = -\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X})(f_t^{lin}(\mathcal{X}) - \mathcal{Y}) \quad (76)$$

$$f_t^{lin}(\mathcal{X}) = e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X})t} (f_0(\mathcal{X}) - \mathcal{Y}) + \mathcal{Y} \quad (77)$$



That is

$$\dot{g}(t) = -\eta \hat{\Theta}_t g(t) \quad (78)$$

$$\dot{g}^{lin}(t) = -\eta \hat{\Theta}_0 g^{lin}(t) \quad (79)$$

$$g^{lin}(t) = e^{-\eta \hat{\Theta}_0 t} g^{lin}(0) \quad (80)$$

$$\begin{aligned} & \frac{d}{dt} \left( e^{\eta \hat{\Theta}_0 t} (g^{lin}(t) - g(t)) \right) \\ &= \frac{d}{dt} \left( g^{lin}(0) - e^{\eta \hat{\Theta}_0 t} g(t) \right) \\ &= 0 - \eta \hat{\Theta}_0 e^{\eta \hat{\Theta}_0 t} g(t) + e^{\eta \hat{\Theta}_0 t} \eta \hat{\Theta}_t g(t) \\ &= -\eta e^{\eta \hat{\Theta}_0 t} \hat{\Theta}_0 g(t) + \eta e^{\eta \hat{\Theta}_0 t} \hat{\Theta}_t g(t) \\ &= \eta e^{\eta \hat{\Theta}_0 t} (\hat{\Theta}_t - \hat{\Theta}_0) g(t) \end{aligned} \quad (81)$$

where  $\hat{\Theta}_0 e^{\eta \hat{\Theta}_0 t} = \hat{\Theta}_0 \sum_{k=0}^{\infty} \frac{1}{k!} (\eta \hat{\Theta}_0 t)^k = \left( \sum_{k=0}^{\infty} \frac{1}{k!} (\eta \hat{\Theta}_0 t)^k \right) \hat{\Theta}_0 = e^{\eta \hat{\Theta}_0 t} \hat{\Theta}_0$

Integrating both sides

$$e^{\eta \hat{\Theta}_0 t} (g^{lin}(t) - g(t)) = \int_0^t \eta e^{\eta \hat{\Theta}_0 s} (\hat{\Theta}_s - \hat{\Theta}_0) g(s) ds \quad (82)$$

$$g^{lin}(t) - g(t) = \int_0^t \eta e^{\eta \hat{\Theta}_0 (s-t)} (\hat{\Theta}_s - \hat{\Theta}_0) g(s) ds \quad (83)$$

This is not easy to bound, so we add a  $g^{lin}(s)$  term.

$$\begin{aligned} g^{lin}(t) - g(t) &= - \int_0^t \eta e^{\eta \hat{\Theta}_0 (s-t)} (\hat{\Theta}_s - \hat{\Theta}_0) (g^{lin}(s) - g(s)) ds \\ &\quad + \int_0^t \eta e^{\eta \hat{\Theta}_0 (s-t)} (\hat{\Theta}_s - \hat{\Theta}_0) g^{lin}(s) ds \end{aligned} \quad (84)$$

$$\begin{aligned} \|g^{lin}(t) - g(t)\|_2 &\leq \eta \int_0^t \left\| e^{\eta \hat{\Theta}_0 (s-t)} \right\|_{op} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s) - g(s)\|_2 ds \\ &\quad + \eta \int_0^t \left\| e^{\eta \hat{\Theta}_0 (s-t)} \right\|_{op} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s)\|_2 ds \end{aligned} \quad (85)$$

where the  $\|\cdot\|_{op}$  is the induced 2-norm for a matrix.

If  $\lambda$  is an eigenvalue of  $\hat{\Theta}_0$ , then  $e^\lambda$  is an eigenvalue of  $e^{\hat{\Theta}_0}$ . Let  $\lambda_0 > 0$  be the smallest eigenvalue of  $\hat{\Theta}_0$ . Since  $s - t < 0$ ,  $\lambda_0 \eta (s - t) < 0$  is the largest eigenvalue of  $\hat{\Theta}_0 \eta (s - t)$ .

$$\left\| e^{\eta \hat{\Theta}_0 (s-t)} \right\|_{op} = \sigma_{max}(e^{\eta \hat{\Theta}_0 (s-t)}) = e^{\lambda_0 \eta (s-t)} \quad (86)$$

Take this into the equation,

$$\begin{aligned} \|g^{lin}(t) - g(t)\|_2 &\leq \eta \int_0^t e^{\lambda_0 \eta (s-t)} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s) - g(s)\|_2 ds \\ &\quad + \eta \int_0^t e^{\lambda_0 \eta (s-t)} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s)\|_2 ds \end{aligned} \quad (87)$$

$$\begin{aligned} e^{\lambda_0 \eta t} \|g^{lin}(t) - g(t)\|_2 &\leq \eta \int_0^t e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s) - g(s)\|_2 ds \\ &\quad + \eta \int_0^t e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s)\|_2 ds \end{aligned} \quad (88)$$

Let

$$u(t) \equiv e^{\lambda_0 \eta t} \|g^{lin}(t) - g(t)\|_2 \quad (89)$$

$$\alpha(t) \equiv \eta \int_0^t e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s)\|_2 ds \quad (90)$$

$$\beta(t) \equiv \eta \left\| (\hat{\Theta}_t - \hat{\Theta}_0) \right\|_{op} \quad (91)$$

The above can be written as

$$u(t) \leq \alpha(t) + \int_0^t \beta(s)u(s) \, ds \quad (92)$$

Since  $e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s)\|_2 \geq 0$ ,  $\alpha(t)$  is non-decreasing. Applying an integral form of the Grönwall's inequality gives

$$u(t) \leq \alpha(t) e^{\int_0^t \beta(s) \, ds} \quad (93)$$

Recall

$$\|g^{lin}(t)\|_2 = \left\| e^{-\eta \hat{\Theta}_0 t} g^{lin}(0) \right\|_2 \leq \left\| e^{-\eta \hat{\Theta}_0 t} \right\|_{op} \|g^{lin}(0)\|_2 = e^{-\lambda_0 \eta t} \|g^{lin}(0)\|_2 \quad (94)$$

Then

$$\begin{aligned} \alpha(t) &= \eta \int_0^t e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s)\|_2 \, ds \\ &\leq \eta \int_0^t e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} e^{-\eta \lambda_0 t} \|g^{lin}(0)\|_2 \, ds \\ &= \eta \|g^{lin}(0)\|_2 \int_0^t e^{\lambda_0 \eta (s-t)} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \, ds \\ &\leq \eta \|g^{lin}(0)\|_2 \int_0^t \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \, ds \end{aligned} \quad (95)$$

since  $e^{\lambda_0 \eta (s-t)} \leq 1$ .

Take this into (93),

$$\begin{aligned} e^{\lambda_0 \eta t} \|g^{lin}(t) - g(t)\|_2 &\leq \eta \int_0^t e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \|g^{lin}(s)\|_2 \, ds e^{\int_0^t \eta \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \, ds} \\ &\leq \eta \|g^{lin}(0)\|_2 \int_0^t \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \, ds e^{\int_0^t \eta \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \, ds} \end{aligned} \quad (96)$$

Let  $\sigma_t = \sup_{0 \leq s \leq t} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op}$ . Then

$$\begin{aligned} e^{\lambda_0 \eta t} \|g^{lin}(t) - g(t)\|_2 &\leq \eta \|g^{lin}(0)\|_2 \int_0^t \sigma_t \, ds e^{\int_0^t \eta \sigma_t \, ds} \\ &= \eta \|g^{lin}(0)\|_2 \sigma_t t e^{\sigma_t \eta t} \\ &= \sigma_t \eta t e^{\sigma_t \eta t} \|g^{lin}(0)\|_2 \end{aligned} \quad (97)$$

$$\|g^{lin}(t) - g(t)\|_2 \leq \sigma_t \eta t e^{\sigma_t \eta t - \lambda_0 \eta t} \|g^{lin}(0)\|_2 \quad (98)$$

As it is proved in Theorem G.1,

$$\sigma_t = \sup_{0 \leq s \leq t} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \leq \sup_t \left\| (\hat{\Theta}_t - \hat{\Theta}_0) \right\|_F \lesssim n^{-\frac{1}{2}} R_0 \rightarrow 0 \quad (99)$$

when  $n_1 = \dots = n_L = n \rightarrow \infty$ . Thus for large  $n$ ,

$$\eta t e^{\eta \sigma_t t - \lambda_0 \eta t} = \frac{\eta t}{e^{\eta t (\lambda_0 - \sigma_t)}} = \mathcal{O}(1) \quad (100)$$

Recall  $g(0) = g^{lin}(0)$  are gaussian distributed and there exist

$$\|g^{lin}(0)\|_2 = \|g(0)\|_2 < R_0 \quad (101)$$

Therefore

$$\|g^{lin}(t) - g(t)\|_2 \lesssim \sigma_t R_0 \lesssim n^{-\frac{1}{2}} R_0^2 \rightarrow 0 \quad (102)$$

as  $n \rightarrow \infty$ .

## References

- [1] Lee J, Xiao L, Schoenholz S S, et al. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent[J]. arXiv preprint arXiv:1902.06720, 2019.
- [2] Jacot A, Gabriel F, Hongler C. Neural tangent kernel: Convergence and generalization in neural networks[C]//Advances in neural information processing systems. 2018: 8571-8580.
- [3] Arora S, Du S S, Hu W, et al. On exact computation with an infinitely wide neural net[C]//Advances in Neural Information Processing Systems. 2019: 8141-8150.
- [4] Lee J, Bahri Y, Novak R, et al. Deep neural networks as gaussian processes[J]. arXiv preprint arXiv:1711.00165, 2017.
- [5] Yang, Greg. "Scaling limits of wide neural networks with weight sharing: Gaussian process behavior, gradient independence, and neural tangent kernel derivation." arXiv preprint arXiv:1902.04760 (2019).
- [6] Huang W, Du W, Da Xu R Y. On the Neural Tangent Kernel of Deep Networks with Orthogonal Initialization[J]. arXiv preprint arXiv:2004.05867, 2020.