Linearized Networks

Yilan Chen

August 21, 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 $l = 1, ..., L$ and readout layer $n_{L+1} = k$.

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

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

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

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

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

Empirical Tangent Kernel:

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

Continuous time gradient descent:

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

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

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

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

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

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

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

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

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 varing with time, one has to use either a numerical method, or an approximation method such as Magnus expansion.

2 Linearized Model

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

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

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

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

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,

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

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

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

Take this into equation (12)

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

Integrate over time,

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

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

Take this into equation (10), we get

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

3 Infinite width limit yields a Gaussian process

Let

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

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.

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

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

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

That is

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

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

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

This is also a gaussian distribution given by

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

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

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 Θ. For any t, (19) is Gaussian distributed because it describe an affine transform of the Gaussian $[f_0(x), f_0(\mathcal{X})]$.

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

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

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

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

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

5 Gradient flow dynamics for training only the readout-layer

This is appendix D in paper[1].

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

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

where θ_i^{L+1} is the i_{th} column of θ^{L+1}. Note $n_{L+1} = k$.

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

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

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

In the case of MSE loss,

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

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

Since different dimension of outputs are independent, we consider only one output,

$$
\begin{aligned}
\dot{\theta}^{L+1} &= -\eta \nabla_{\theta^{L+1}} f(\mathcal{X})^T \nabla_{f_t(\mathcal{X})} \mathcal{L} = -\eta \nabla_{\theta^{L+1}} f(\mathcal{X})^T (\overline{x}(\mathcal{X}) \theta^{L+1} - \mathcal{Y}) \\
&= -\eta \overline{x}(\mathcal{X})^T (\overline{x}(\mathcal{X}) \theta^{L+1} - \mathcal{Y})
\end{aligned} \tag{38}
$$

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

$$
\overline{x}(\mathcal{X}) \theta_t^{L+1} = e^{-\eta \overline{x}(\mathcal{X})^T \overline{x}(\mathcal{X}) t} (\overline{x}(\mathcal{X}) \theta_0^{L+1} - \mathcal{Y}) + \mathcal{Y} \tag{39}
$$

$$
f_t(\mathcal{X}) = e^{-\eta \overline{x}(\mathcal{X})^T \overline{x}(\mathcal{X}) t} (f_0(\mathcal{X}) - \mathcal{Y}) + \mathcal{Y} \tag{40}
$$

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

$$
\begin{aligned}
f_t(\mathcal{X}) &= e^{-\eta \overline{x}(\mathcal{X})^T \overline{x}(\mathcal{X}) t} (f_0(\mathcal{X}) - \mathcal{Y}) + \mathcal{Y} \\
&= e^{-\eta \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) t} \left(f_0(\mathcal{X}) - \mathcal{Y}\right) + \mathcal{Y}
\end{aligned} \tag{41}
$$

We can see

$$
\overline{x}(\mathcal{X})^T \overline{x}(\mathcal{X}) = \hat{\Theta}_0(\mathcal{X}, \mathcal{X}) = \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X}) \tag{42}
$$

Take this into (19)

$$
f_t(x) = f_0(x) - \hat{\mathcal{K}}(x, \mathcal{X}) \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X})^{-1} \left(I - e^{-\eta \hat{\mathcal{K}}(\mathcal{X}, \mathcal{X}) t}\right) \left(f_0(\mathcal{X}) - \mathcal{Y}\right) \tag{43}
$$

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

$$
\mathbb{E}[f_t(x)] = \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}\right) \mathcal{Y} \tag{44}
$$

$$
\begin{aligned}
\Sigma(x, x) &= \mathcal{K}(x, x) - \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}\right) \mathcal{K}(\mathcal{X}, x) - \mathcal{K}(x, \mathcal{X}) \left(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}\right) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \mathcal{K}(\mathcal{X}, x) \\
&\quad + \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}\right) \mathcal{K}(\mathcal{X}, \mathcal{X}) \left(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}\right) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \mathcal{K}(\mathcal{X}, x) \\
&= \mathcal{K}(x, x) - 2\mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}\right) \mathcal{K}(\mathcal{X}, x) \\
&\quad + \mathcal{K}(x, \mathcal{X}) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \left(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}\right) \mathcal{K}(\mathcal{X}, \mathcal{X}) \left(I - e^{-\eta \mathcal{K}(\mathcal{X}, \mathcal{X}) t}\right) \mathcal{K}(\mathcal{X}, \mathcal{X})^{-1} \mathcal{K}(\mathcal{X}, x)
\end{aligned} \tag{45}
$$

where

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

Take this into the equation,

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

6 Infinite width networks are linearized networks

Theorem 2.1 (informal). *Let $n_1 = \cdots = n_L = n$ and assume $\lambda_{min}(\Theta) > 0$. Applying gradient descent with learning rate $\eta < \eta_{critical}$ (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,*

$$\sup_{t\geq 0}\left\|f_t(x) - f_t^{lin}(x)\right\|_2, \ \sup_{t\geq 0}\frac{\|\theta_t-\theta_0\|_2}{\sqrt{n}}, \ \sup_{t\geq 0}\left\|\hat{\Theta}_t - \hat{\Theta}_0\right\|_F = \mathcal{O}(n^{-\frac{1}{2}}), \ \textit{as} \quad n\to\infty \tag{48}$$

Some short-hand notations:

$$f(\theta_t) = f(\mathcal{X},\theta_t) \in \mathbb{R}^{|\mathcal{X}|k} \tag{49}$$

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

$$J(\theta_t) = \nabla f(\theta_t) \in \mathbb{R}^{(|\mathcal{X}|k)\times|\theta|} \tag{51}$$

$$\begin{cases} \hat{\Theta}_t := \hat{\Theta}_t(\mathcal{X},\mathcal{X}) = \frac{1}{n}J(\theta_t)J(\theta_t)^T \\ \Theta := \lim_{n\to\infty}\hat{\Theta}_0 \quad \text{in probability.}\end{cases} \tag{52}$$

The gradient descent update with learning rate η:

$$\theta_{t+1} = \theta_t - \eta J(\theta_t)^T g(\theta_t) \tag{53}$$

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*

$$\begin{cases} \frac{1}{\sqrt{n}}\left\|J(\theta) - J(\tilde{\theta})\right\|_F & \leq K\left\|\theta - \tilde{\theta}\right\|_2 \\ \frac{1}{\sqrt{n}}\|J(\theta)\|_F & \leq K \end{cases}, \quad \forall \theta,\tilde{\theta} \in B(\theta_0, Cn^{-\frac{1}{2}}) \tag{54}$$

where

$$B(\theta_0, R) := \{\theta : \|\theta - \theta_0\|_2 < R\} \tag{55}$$

Theorem G.1 (Gradient descent). *Assume* **Assumptions [1-4]**. *For $\delta_0 > 0$ and $\eta_0 < \eta_{critical}$, 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 $\eta = \frac{\eta_0}{n}$,*

$$\begin{cases} \|g(\theta_t)\|_2 \leq (1 - \frac{\eta_0\lambda_{min}}{3})^t R_0 \\ \sum_{j=1}^t \|\theta_j - \theta_{j-1}\|_2 \leq \frac{\eta_0 K R_0}{\sqrt{n}}\sum_{j=1}^t(1-\frac{\eta_0\lambda_{min}}{3})^{j-1} \leq \frac{3KR_0}{\lambda_{min}}n^{-\frac{1}{2}} \end{cases} \tag{56}$$

and

$$\sup_t \left\|\hat{\Theta}_0 - \hat{\Theta}_t\right\|_F \leq \frac{6K^3R_0}{\lambda_{min}}n^{-\frac{1}{2}} \tag{57}$$

The first inequation of (56) indicates when $t\to\infty$, $g(\theta_t)\to 0$, i.e. the convergence of training. The second inequation of (56) bounds the change of θ with n. The larger n is, the less θ changes. The inequation of (57) bounds the change of $\hat{\Theta}$ with n during training.

6.1 Proof of Theorem G.1

We first prove (56) 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,

$$\|g(\theta_0)\|_2 < R_0 \tag{58}$$

This is the case of $t = 0$ for (56). Assume (56) holds for $t = t$. Then for $t + 1$

$$\|\theta_{t+1} - \theta_t\|_2 = \left\|-\eta J(\theta_t)^T g(\theta_t)\right\|_2 \leq \eta \|J(\theta_t)\|_{op} \|g(\theta_t)\|_2 \leq \frac{\eta_0}{n} \|J(\theta_t)\|_{op} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^t R_0 \tag{59}$$

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

$$\|J(\theta_t)\|_2 = \sigma_{max}(J(\theta_t)) \leq \|J(\theta_t)\|_F \leq K\sqrt{n} \tag{60}$$

So we get

$$\|\theta_{t+1} - \theta_t\|_2 \leq \frac{K\eta_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^t R_0 \tag{61}$$

Then

$$\begin{aligned}
\|\theta_{t+1} - \theta_0\|_2 &= \|\theta_{t+1} - \theta_t + \theta_t - \theta_{t-1} + \cdots - \theta_0\|_2 \\
&\leq \|\theta_{t+1} - \theta_t\|_2 + \cdots + \|\theta_1 - \theta_0\|_2 \\
&\leq \sum_{j=1}^{t+1} \frac{K\eta_0 R_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{j-1}
\end{aligned} \tag{62}$$

which is the sum of a geometric progression,

$$\sum_{j=1}^{t+1} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{j-1} = \frac{1 - \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{t+1}}{1 - \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)} = 3\frac{1 - \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{t+1}}{\eta_0 \lambda_{min}} \leq \frac{3}{\eta_0 \lambda_{min}} \tag{63}$$

So we get the second inequation of (56).

$$\|\theta_{t+1} - \theta_0\|_2 \leq \sum_{j=1}^{t+1} \|\theta_j - \theta_{j-1}\|_2 \leq \sum_{j=1}^{t+1} \frac{K\eta_0 R_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{j-1} \leq \frac{K\eta_0 R_0}{\sqrt{n}} \frac{3}{\eta_0 \lambda_{min}} = \frac{3KR_0}{\lambda_{min}} n^{-\frac{1}{2}} \tag{64}$$

For the first inequation of (56), we apply the mean value theorem.

$$\begin{aligned}
\|g(\theta_{t+1})\|_2 &= \|g(\theta_{t+1}) - g(\theta_t) + g(\theta_t)\|_2 \\
&= \left\|J(\tilde{\theta}_t)(\theta_{t+1} - \theta_t) + g(\theta_t)\right\|_2 \\
&= \left\|-\eta J(\tilde{\theta}_t) J(\theta_t)^T g(\theta_t) + g(\theta_t)\right\|_2 \\
&= \left\|\left(I - \eta J(\tilde{\theta}_t) J(\theta_t)^T\right) g(\theta_t)\right\|_2 \\
&\leq \left\|I - \eta J(\tilde{\theta}_t) J(\theta_t)^T\right\|_{op} \|g(\theta_t)\|_2 \\
&\leq \left\|I - \eta J(\tilde{\theta}_t) J(\theta_t)^T\right\|_{op} (1 - \frac{\eta_0 \lambda_{min}}{3})^t R_0
\end{aligned} \tag{65}$$

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

$$\begin{aligned}
\left\|I - \eta J(\tilde{\theta}_t) J(\theta_t)^T\right\|_{op} &= \left\|I - \eta J(\theta_0) J(\theta_0)^T + \eta J(\theta_0) J(\theta_0)^T - \eta J(\tilde{\theta}_t) J(\theta_t)^T\right\|_{op} \\
&= \left\|I - \eta_0 \Theta_0 + \eta J(\theta_0) J(\theta_0)^T - \eta J(\tilde{\theta}_t) J(\theta_t)^T\right\|_{op} \\
&= \left\|I - \eta_0 \Theta + \eta_0 \Theta - \eta_0 \Theta_0 + \eta J(\theta_0) J(\theta_0)^T - \eta J(\tilde{\theta}_t) J(\theta_t)^T\right\|_{op} \\
&\leq \|I - \eta_0 \Theta\|_{op} + \eta_0 \|\Theta - \Theta_0\|_{op} + \eta \left\|J(\theta_0) J(\theta_0)^T - J(\tilde{\theta}_t) J(\theta_t)^T\right\|_{op}
\end{aligned} \tag{66}$$

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

$$\|I - \eta_0 \Theta\|_{op} = \sigma_{max}(I - \eta_0 \Theta) = 1 - \eta_0 \sigma_{min}(\Theta) = 1 - \eta_0 \lambda_{min} \tag{67}$$

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

$$\eta_0 \left\|\Theta - \hat{\Theta}_0\right\|_{op} \le \eta_0 \left\|\Theta - \hat{\Theta}_0\right\|_F \le \frac{\eta_0 \lambda_{min}}{3} \tag{68}$$

For the third part

$$\begin{aligned}
\left\|J(\theta_0)J(\theta_0)^T - J(\tilde{\theta}_t)J(\theta_t)^T\right\|_{op} &= \left\|J(\theta_0)J(\theta_0)^T - J(\theta_0)J(\theta_t)^T + J(\theta_0)J(\theta_t)^T - J(\tilde{\theta}_t)J(\theta_t)^T\right\|_{op} \\
&= \left\|J(\theta_0)[J(\theta_0)^T - J(\theta_t)^T] + [J(\theta_0) - J(\tilde{\theta}_t)]J(\theta_t)^T\right\|_{op} \\
&\le \|J(\theta_0)\|_{op} \left\|J(\theta_0)^T - J(\theta_t)^T\right\|_{op} + \left\|J(\theta_0) - J(\tilde{\theta}_t)\right\|_{op} \left\|J(\theta_t)^T\right\|_{op} \\
&\le \|J(\theta_0)\|_F \left\|J(\theta_0)^T - J(\theta_t)^T\right\|_F + \left\|J(\theta_0) - J(\tilde{\theta}_t)\right\|_F \left\|J(\theta_t)^T\right\|_F \\
&\le K\sqrt{n}K\sqrt{n}\left\|\theta_0 - \theta_t\right\|_2 + K\sqrt{n}K\sqrt{n}\left\|\theta_0 - \tilde{\theta}_t\right\|_2 \\
&= K^2 n \left\|\theta_t - \theta_0\right\|_2 + K^2 n \left\|\tilde{\theta}_t - \theta_0\right\|_2 \\
&\le 2K^2 n \frac{3KR_0}{\lambda_{min}} n^{-\frac{1}{2}}
\end{aligned} \tag{69}$$

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

$$\begin{aligned}
\left\|1 - \eta J(\tilde{\theta}_t)J(\theta_t)^T\right\|_{op} &\le 1 - \eta_0\lambda_{min} + \frac{\eta_0\lambda_{min}}{3} + 2\eta_0 K^2 \frac{3KR_0}{\lambda_{min}} n^{-\frac{1}{2}} \\
&\le 1 - \frac{\eta_0\lambda_{min}}{3}
\end{aligned} \tag{70}$$

Take this into (65),

$$\|g(\theta_{t+1})\|_2 \le (1 - \frac{\eta_0\lambda_{min}}{3})^{t+1} R_0 \tag{71}$$

For (57),

$$\begin{aligned}
\left\|\hat{\Theta}_0 - \hat{\Theta}_t\right\|_F &= \frac{1}{n}\left\|J(\theta_0)J(\theta_0)^T - J(\theta_t)J(\theta_t)^T\right\|_F \\
&= \frac{1}{n}\left\|J(\theta_0)J(\theta_0)^T - J(\theta_0)J(\theta_t)^T + J(\theta_0)J(\theta_t)^T - J(\theta_t)J(\theta_t)^T\right\|_F \\
&= \frac{1}{n}\left\|J(\theta_0)[J(\theta_0)^T - J(\theta_t)^T] + [J(\theta_0) - J(\theta_t)]J(\theta_t)^T\right\|_F \\
&\le \frac{1}{n}\|J(\theta_0)\|_F \left\|J(\theta_0)^T - J(\theta_t)^T\right\|_F + \frac{1}{n}\|J(\theta_0) - J(\theta_t)\|_F \left\|J(\theta_t)^T\right\|_F \\
&\le K^2 \|\theta_t - \theta_0\|_2 + K^2 \|\theta_t - \theta_0\|_2 \\
&\le \frac{6K^3R_0}{\lambda_{min}} n^{-\frac{1}{2}}
\end{aligned} \tag{72}$$

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

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 $\|x\|_2 \le 1$, for $\delta_0 > 0$ arbitrarily small, there exist $R_0 > 0$ and $N \in \mathbb{N}$ such that for every $n \ge N$, with probability at least $(1 - \delta_0)$ over random initialization,*

$$\sup_t \left\|g^{lin}(t) - g(t)\right\|_2, \quad \sup_t \left\|g^{lin}(t,x) - g(t,x)\right\|_2 \lesssim n^{-\frac{1}{2}} R_0^2 \tag{73}$$

Proof.

Recall

$$\dot{f}_t(\mathcal{X}) = -\eta\hat{\Theta}_t(\mathcal{X},\mathcal{X})\big(f_t(\mathcal{X}) - \mathcal{Y}\big) \tag{74}$$

$$\dot{f}_t^{lin}(\mathcal{X}) = -\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})\big(f_t^{lin}(\mathcal{X}) - \mathcal{Y}\big) \tag{75}$$

$$f_t^{lin}(\mathcal{X}) = e^{-\eta\hat{\Theta}_0(\mathcal{X},\mathcal{X})t}\big(f_0(\mathcal{X}) - \mathcal{Y}\big) + \mathcal{Y} \tag{76}$$

That is

$$\dot{g}(t) = -\eta\hat{\Theta}_t g(t) \tag{77}$$
$$\dot{g}^{lin}(t) = -\eta\hat{\Theta}_0 g^{lin}(t) \tag{78}$$
$$g^{lin}(t) = e^{-\eta\hat{\Theta}_0 t} g^{lin}(0) \tag{79}$$

$$\begin{aligned}
&\frac{d}{dt}\left(e^{\eta\hat{\Theta}_0 t}\big(g^{lin}(t) - g(t)\big)\right) \\
&= \frac{d}{dt}\left(g^{lin}(0) - e^{\eta\hat{\Theta}_0 t} g(t)\right) \\
&= 0 - \eta\hat{\Theta}_0 e^{\eta\hat{\Theta}_0 t} g(t) + e^{\eta\hat{\Theta}_0 t}\eta\hat{\Theta}_t g(t) \\
&= -\eta e^{\eta\hat{\Theta}_0 t}\hat{\Theta}_0 g(t) + \eta e^{\eta\hat{\Theta}_0 t}\hat{\Theta}_t g(t) \\
&= \eta e^{\eta\hat{\Theta}_0 t}\big(\hat{\Theta}_t - \hat{\Theta}_0\big) g(t)
\end{aligned} \tag{80}$$

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

Integrating both sides

$$e^{\eta\hat{\Theta}_0 t}\big(g^{lin}(t) - g(t)\big) = \int_0^t \eta e^{\eta\hat{\Theta}_0 s}\big(\hat{\Theta}_s - \hat{\Theta}_0\big) g(s)\,\mathrm{d}s \tag{81}$$

$$g^{lin}(t) - g(t) = \int_0^t \eta e^{\eta\hat{\Theta}_0 (s-t)}\big(\hat{\Theta}_s - \hat{\Theta}_0\big) g(s)\,\mathrm{d}s \tag{82}$$

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

$$\begin{aligned}
g^{lin}(t) - g(t) = &-\int_0^t \eta e^{\eta\hat{\Theta}_0 (s-t)}\big(\hat{\Theta}_s - \hat{\Theta}_0\big)\big(g^{lin}(s) - g(s)\big)\,\mathrm{d}s \\
&+ \int_0^t \eta e^{\eta\hat{\Theta}_0 (s-t)}\big(\hat{\Theta}_s - \hat{\Theta}_0\big) g^{lin}(s)\,\mathrm{d}s
\end{aligned} \tag{83}$$

$$\begin{aligned}
\left\|g^{lin}(t) - g(t)\right\|_2 \leq &\eta\int_0^t \left\|e^{\eta\hat{\Theta}_0 (s-t)}\right\|_{op}\left\|\big(\hat{\Theta}_s - \hat{\Theta}_0\big)\right\|_{op}\left\|g^{lin}(s) - g(s)\right\|_2\,\mathrm{d}s \\
&+ \eta\int_0^t \left\|e^{\eta\hat{\Theta}_0 (s-t)}\right\|_{op}\left\|\big(\hat{\Theta}_s - \hat{\Theta}_0\big)\right\|_{op}\left\|g^{lin}(s)\right\|_2\,\mathrm{d}s
\end{aligned} \tag{84}$$

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

If λ is an eigenvalue of $\hat{\Theta}_0$, then e^{λ} 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 $\hat{\Theta}_0\eta(s-t)$.

$$\left\|e^{\eta\hat{\Theta}_0 (s-t)}\right\|_{op} = \sigma_{max}(e^{\eta\hat{\Theta}_0 (s-t)}) = e^{\lambda_0\eta(s-t)} \tag{85}$$

Take this into the equation,

$$\begin{aligned}
\left\|g^{lin}(t) - g(t)\right\|_2 \leq &\eta\int_0^t e^{\lambda_0\eta(s-t)}\left\|\big(\hat{\Theta}_s - \hat{\Theta}_0\big)\right\|_{op}\left\|g^{lin}(s) - g(s)\right\|_2\,\mathrm{d}s \\
&+ \eta\int_0^t e^{\lambda_0\eta(s-t)}\left\|\big(\hat{\Theta}_s - \hat{\Theta}_0\big)\right\|_{op}\left\|g^{lin}(s)\right\|_2\,\mathrm{d}s
\end{aligned} \tag{86}$$

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

Let

$$u(t) \equiv e^{\lambda_0\eta t}\left\|g^{lin}(t) - g(t)\right\|_2 \tag{88}$$
$$\alpha(t) \equiv \eta\int_0^t e^{\lambda_0\eta s}\left\|\big(\hat{\Theta}_s - \hat{\Theta}_0\big)\right\|_{op}\left\|g^{lin}(s)\right\|_2\,\mathrm{d}s \tag{89}$$
$$\beta(t) \equiv \eta\left\|\big(\hat{\Theta}_t - \hat{\Theta}_0\big)\right\|_{op} \tag{90}$$

The above can be written as

$$u(t) \leq \alpha(t) + \int_0^t \beta(s) u(s) \, \mathrm{d}s \tag{91}$$

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

$$u(t) \leq \alpha(t) e^{\int_0^t \beta(s) \, \mathrm{d}s} \tag{92}$$

Recall

$$\left\| g^{lin}(t) \right\|_2 = \left\| e^{-\eta \hat{\Theta}_0 t} g^{lin}(0) \right\|_2 \leq \left\| e^{-\eta \hat{\Theta}_0 t} \right\|_{op} \left\| g^{lin}(0) \right\|_2 = e^{-\lambda_0 \eta t} \left\| g^{lin}(0) \right\|_2 \tag{93}$$

Then

$$\begin{aligned} \alpha(t) &= \eta \int_0^t e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \left\| g^{lin}(s) \right\|_2 \, \mathrm{d}s \\ &\leq \eta \int_0^t e^{\lambda_0 \eta s} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} e^{-\eta \lambda_0 t} \left\| g^{lin}(0) \right\|_2 \, \mathrm{d}s \\ &= \eta \left\| g^{lin}(0) \right\|_2 \int_0^t e^{\lambda_0 \eta (s-t)} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \, \mathrm{d}s \\ &\leq \eta \left\| g^{lin}(0) \right\|_2 \int_0^t \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \, \mathrm{d}s \end{aligned} \tag{94}$$

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

Take this into (92),

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

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

$$\begin{aligned} e^{\lambda_0 \eta t} \left\| g^{lin}(t) - g(t) \right\|_2 &\leq \eta \left\| g^{lin}(0) \right\|_2 \int_0^t \sigma_t \, \mathrm{d}s \; e^{\int_0^t \eta \sigma_t \, \mathrm{d}s} \\ &= \eta \left\| g^{lin}(0) \right\|_2 \sigma_t t \; e^{\sigma_t \eta t} \\ &= \sigma_t \eta t e^{\sigma_t \eta t} \left\| g^{lin}(0) \right\|_2 \end{aligned} \tag{96}$$

$$\left\| g^{lin}(t) - g(t) \right\|_2 \leq \sigma_t \eta t e^{\sigma_t \eta t - \lambda_0 \eta t} \left\| g^{lin}(0) \right\|_2 \tag{97}$$

As it is proved in Theorem G.1,

$$\sigma_t = \sup_{0 \leq s \leq t} \left\| (\hat{\Theta}_s - \hat{\Theta}_0) \right\|_{op} \leq \sup_t \left\| (\hat{\Theta}_t - \hat{\Theta}_0) \right\|_F \lesssim n^{-\frac{1}{2}} R_0 \to 0 \tag{98}$$

when $n_1 = \cdots = n_L = n \to \infty$. Thus for large n,

$$\eta t e^{\eta \sigma_t t - \lambda_0 \eta t} = \frac{\eta t}{e^{\eta t (\lambda_0 - \sigma_t)}} = \mathcal{O}(1) \tag{99}$$

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

$$\left\| g^{lin}(0) \right\|_2 = \left\| g(0) \right\|_2 < R_0 \tag{100}$$

Therefore

$$\left\| g^{lin}(t) - g(t) \right\|_2 \lesssim \sigma_t R_0 \lesssim n^{-\frac{1}{2}} R_0^2 \to 0 \tag{101}$$

as $n \to \infty$.

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.