A EXPERIMENTS

A.1 EXPERIMENTAL SETUP

For the MNIST dataset: We use the convolution neural network with the structure: $n \times 32C$-(1024)-d. The output dimension d is 1 or 10, with respect to the loss function MSE loss or cross-entropy. The parameters of the convolution layer is initialized by the Gaussian $(0, \sigma_1^2)$, and the parameters of the linear layer is Gaussian $(0, \sigma_2^2)$. σ_1 is given by $(\frac{(c_{in}+c_{out})*m^2}{2})^{-\gamma}$ where c_{in} and c_{out} are the number of in channels and out channels respectively, σ_2 is given empirically by 0.0001. The data size is n which is randomly chosen from the MNIST dataset. The training method is GD or Adam with full batch.

A.2 CIFAR10 EXAMPLES

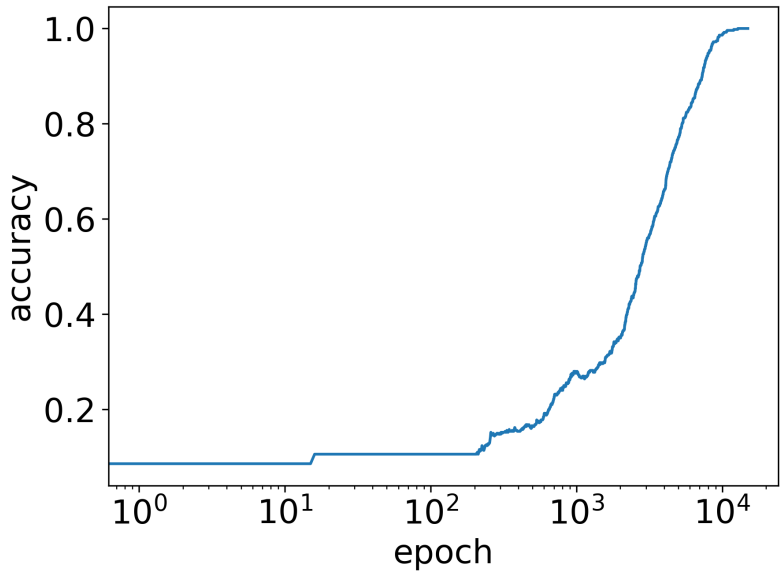


Figure 5: The training accuracy of the experiment in Fig: 3

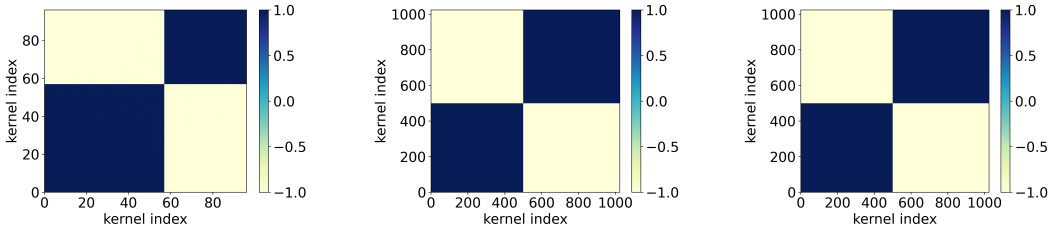
A.3 MNIST EXAMPLES

When we are training the MNIST dataset, we use MSE loss with one dimension output as the criterion.

We display the initial condensation of the convolution neural network with one convolution layer on the MNIST dataset whose data size is $n = 500$. The colors in Fig.8 show the cosine similarity $D(u, v)$ of different kernel in one convolution layer. Yellow $D(u, v) \sim 1$(purple $D(u, v) \sim -1$), means the kernel weight vectors are at the same(opposite) directions, The activation function of the convolution layer are $\mathrm{Sigmoid(x)}$, $\mathrm{ReLU}(x)$ and $\mathrm{tanh(x)}$. The Fig.8c implies that during the initial stage of the training, the convolutional kernel in each layer will condense at two opposite directions when the activation function is $\mathrm{tanh(x)}$. This phenomenon will also happen when the activation functions are $\mathrm{ReLU}(x)$ and $\mathrm{Sigmoid(x)}$ which is shown in Fig.8a and Fig.8b.

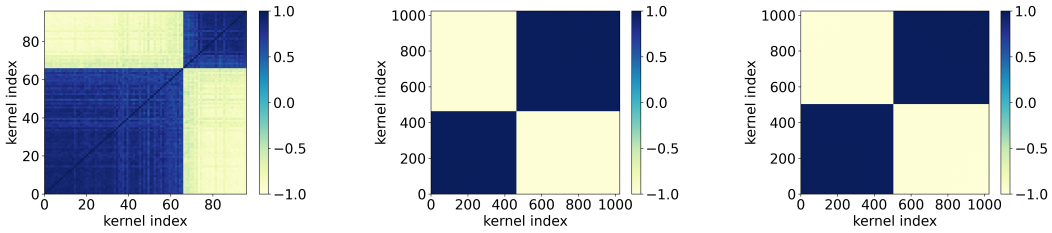
Condensation of multi-layer convolution neural network on MNIST dataset is shown in Fig.9. The activation function between each convolution layer is $\mathrm{tanh(x)}$. We have the kernel weight of different layer are all condensation on two opposite directions, i.e. one line.

In the MNIST example, it has only one input channel, thus we directly project the first eigenvector to $\mathbb{1}$ and find the inner product is close to 1, i.e. 0.99974 ± 0.00003, computed from 50 independent trials, where each trail randomly selected 500 images.



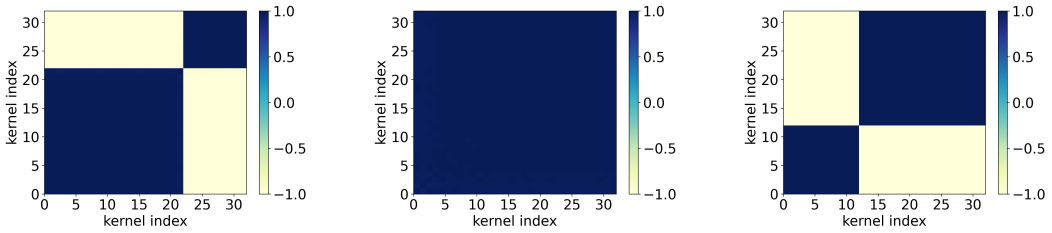
(a) layer 1 (b) layer 2 (c) layer 3

Figure 6: Condensation of convolution network with three convolutional kernel training on CIFAR10 dataset with data size= 500. The kernel size $m = 3$. The colors in the figure show the cosine similarity of normalized weight vectors of each convolutional kernel. The activation for all convolution layers are $\tanh(\mathrm{x})$. The number of the steps are at epoch $=$ 200, epoch $=$ 200 and epoch $=$ 200. The convolutional kernels are initialized by $\gamma = 2$. The learning rate is 5×10^{-6}. We use one dimension output instead of label and use MSE loss. The optimizer is full batch Adam.



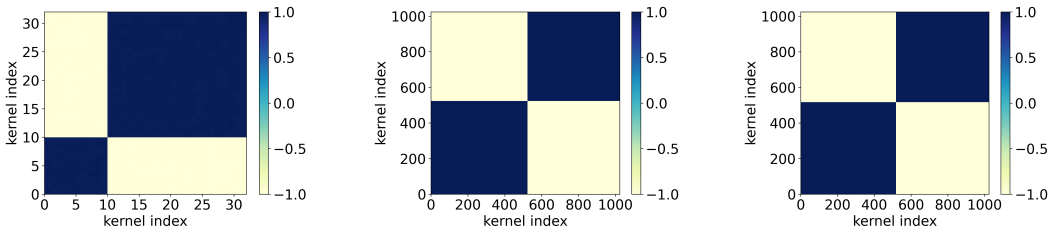
(a) layer 1 (b) layer 2 (c) layer 3

Figure 7: Condensation of three-convolution-layer CNNs. The activation functions are $\tanh(\mathrm{x})$. The numbers of steps selected in the sub-pictures are epoch$=$ 200, epoch$=$ 300 and epoch$=$ 300, respectively, while the NN is only trained once. The color indicates $D(\boldsymbol{u}, \boldsymbol{v})$ of two different kernels, whose indexes are indicated by the abscissa and the ordinate, respectively. The training data is CIFAR10 dataset. ReLU is used for linear layer, softmax for output layer, MSE for loss function and full batch Adam for optimizer. $m = 5$, $lr = 1 \times 10^{-6}$, and $\gamma = 1.2$.



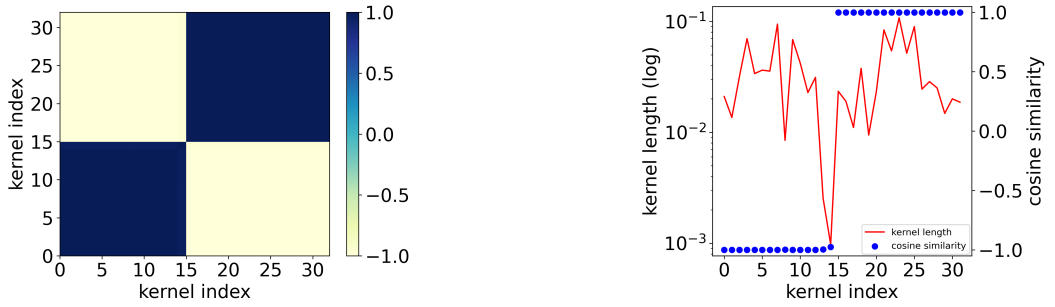
(a) $\mathrm{ReLU}(x)$ (b) $\mathrm{Sigmoid}(\mathrm{x})$ (c) $\tanh(\mathrm{x})$

Figure 8: Condensation of CNNs with different activations (indicated by sub-captions) for convolution layers. The network has 32 kernels in the convolution layer, followed by 1-dimensional output. The kernel size is $m = 5$. The learning rate is 5×10^{-7}, 5×10^{-7} and 5×10^{-6} separately. The number of the selected steps are at epoch $=$ 252, epoch $=$ 302, and epoch $=$ 200. The convolution layer is initialized by $\gamma = 2$. We use full batch Adam optimizer with MSE loss on MNIST dataset.



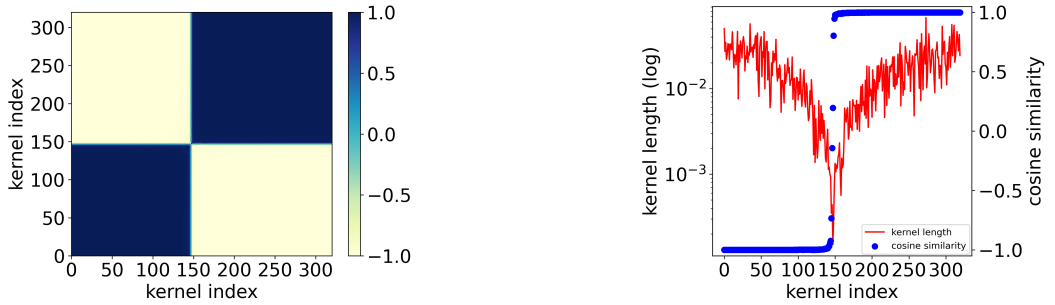
(a) layer 1 (b) layer 2 (c) layer 3

Figure 9: Condensation of $\tanh$ CNN with three convolution layer by MSE loss. The kernel size is $m = 5$. The color indicates the cosine similarity between kernels. The number of the steps are all at epoch $= 100$. The learning rate is 5×10^{-7}. The convolution layer is initialized by $\gamma = 2$. We use 1-dimension output. The optimizer is full batch Adam on MNIST dataset.



(a) condensation (b) directions

Figure 10: Condensation of two-layer CNN by GD and MSE training on MNIST dataset with data size $n = 500$. The network has 32 kernels. (a) consine similarity. (b) left ordinate (red): the amplitude of each kernel; (b) right ordinate: cosine similarity between each kernel and $\mathbb{1}$. The activation function of the convolution part is $\tanh(x)$. The kernel size is $m = 3$. The learning rate is 5×10^{-6}. The number of the selected steps is epoch $= 3600$. The convolution layer is initialized by $\gamma = 2$.



(a) condensation (b) directions

Figure 11: Condensation of two-layer CNN by GD and MSE training on MNIST dataset with data size $n = 500$.The network has 320 kernels. The activation function of the convolution part is $\tanh(x)$. The kernel size is $m = 3$. The learning rate is 5×10^{-6}. The number of the selected steps is epoch $= 7000$. The convolution layer is initialized by $\gamma = 2$.

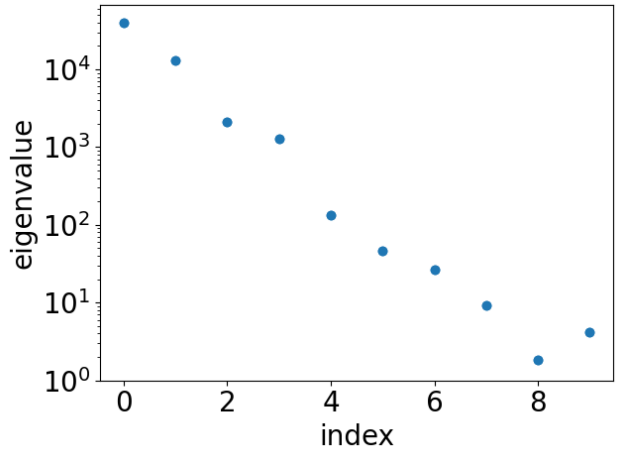


Figure 12: The largest 10 eigenvalues of $\boldsymbol{Z}^\intercal \boldsymbol{Z}$ in tanh(x) CNN on MNIST dataset. A clear spectral gap ($\Delta\lambda := \lambda_1 - \lambda_2$) could be observed, which satisfies Assumption 4.

The loss at the initial stage is shown as follows:

B PRELIMINARIES

B.1 SOME NOTATIONS

For a matrix $\mathbf{A}$, we use $\mathbf{A}_{i,j}$ to denote its (i,j)-th entry. We also use $\mathbf{A}_{i,:}$ to denote the i-th row vector of $\mathbf{A}$ and define $\mathbf{A}_{i,j:k} := (\mathbf{A}_{i,j}, \mathbf{A}_{i,j+1}, \cdots, \mathbf{A}_{i,k})$ as part of the vector. Similarly $\mathbf{A}_{:,k}$ is the k-th column vector and $\mathbf{A}_{i:j,k} := (\mathbf{A}_{i,k}, \mathbf{A}_{i+1,k}, \cdots, \mathbf{A}_{j,k})^\intercal$ is part of the k-th column vector.

We let $[n] = \{1, 2, \ldots, n\}$. We set $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ as the normal distribution with mean $\boldsymbol{\mu}$ and covariance Σ. For a vector $\mathbf{v}$, we use $\|\mathbf{v}\|_2$ to denote its Euclidean norm, and we use $\langle \cdot, \cdot \rangle$ to denote the standard inner product between two vectors. For a matrix $\mathbf{A}$, we use $\|\mathbf{A}\|_{\mathrm{F}}$ to denote its Frobenius norm and $\|\mathbf{A}\|_{2\to 2}$ to denote its operator norm. Finally, we use $\mathcal{O}(\cdot)$ and $\Omega(\cdot)$ for the standard Big-O and Big-Omega notations.

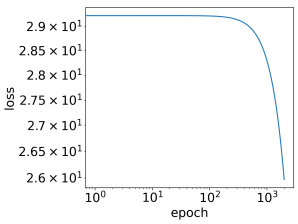
B.2 PROBLEM SETUP

We focus on the empirical risk minimization problem given by the quadratic loss:

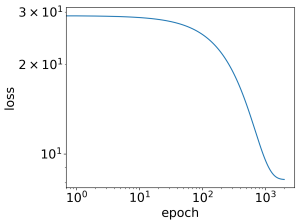
$$\min_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}) = \frac{1}{2n} \sum_{i=1}^{n} \left(f(\boldsymbol{x}_i, \boldsymbol{\theta}) - y_i \right)^2 . \tag{16}$$

In the above, n is the total number of training samples, $\{\boldsymbol{x}_i\}_{i=1}^n$ are the training inputs, $\{y_i\}_{i=1}^n$ are the labels, $f(\boldsymbol{x}_i, \boldsymbol{\theta})$ is the prediction function, and $\boldsymbol{\theta}$ are the parameters to be optimized, and their dependence is modeled by a $(L+1)$-layer convolution neural network (CNN) with filter size $m \times m$. We denote $\boldsymbol{x}^{[l]}(i)$ as the output of the l-th layer with respect to the i-th sample for $l \geq 1$, and $\boldsymbol{x}^{[0]}(i) := \boldsymbol{x}_i$ is the i-th input. For any $l \in [0:L]$, we denote the size of width, height, channel of $\boldsymbol{x}^{[l]}$ as W_l, H_l, and C_l respectively, i.e., $\{\boldsymbol{x}^{[l]}(i)\}_{i=1}^n \subset \mathbb{R}^{W_l \times H_l \times C_l}$. We introduce a filter operator $\chi(\cdot, \cdot)$, which maps the width and height indices of the output of all layers to a binary variable, i.e., for a filter of size $m \times m$, the filter operator reads

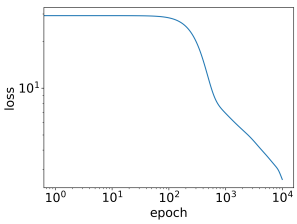
$$\chi(p,q) = \begin{cases} 1, & \text{for } 0 \leqslant p, q \leqslant m-1 \\ 0, & \text{otherwise,} \end{cases} \tag{17}$$



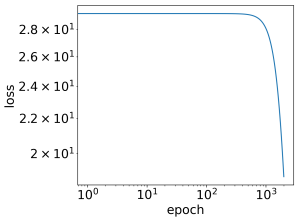
(a) Fig.8a



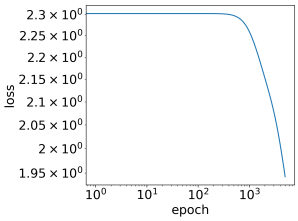
(b) Fig.8b



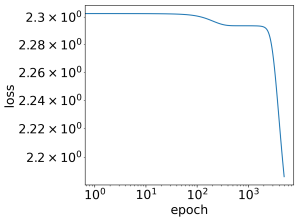
(c) Fig.8c



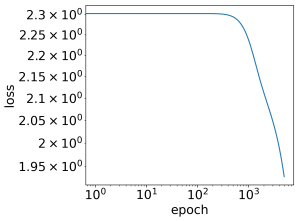
(d) Fig.9



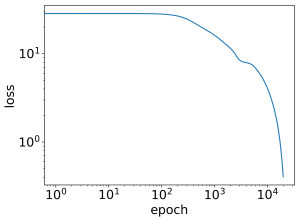
(e) Fig.4a



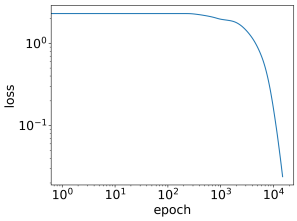
(f) Fig.4b



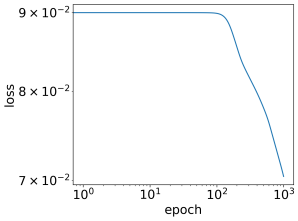
(g) Fig.4c



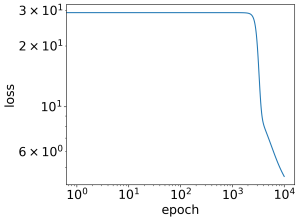
(h) Fig.6



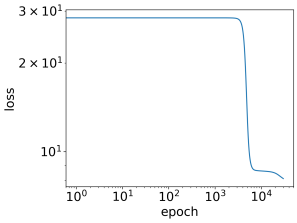
(i) Fig.3(d)-(f)



(j) Fig.7



(k) Fig.10a



(l) Fig.2b

Figure 13: Losses of the experiments on MNIST and CIFAR10 dataset. The original figures corresponding to each sub-picture are written in the sub-captions.

and the $(L+1)$-layer CNN with filter size $m \times m$ is recursively defined for $l \in [2:L]$,

$$\boldsymbol{x}_{u,v,\beta}^{[1]} := \left[\sum_{\alpha=1}^{C_0}\left(\sum_{p=-\infty}^{+\infty}\sum_{q=-\infty}^{+\infty}\boldsymbol{x}_{u+p,v+q,\alpha}^{[0]}\cdot\boldsymbol{W}_{p,q,\alpha,\beta}^{[1]}\cdot\chi(p,q)\right)\right]+\boldsymbol{b}_{\beta}^{[1]},$$

$$\boldsymbol{x}_{u,v,\beta}^{[l]} := \left[\sum_{\alpha=1}^{C_{l-1}}\left(\sum_{p=-\infty}^{+\infty}\sum_{q=-\infty}^{\infty}\sigma\left(\boldsymbol{x}_{u+p,v+q,\alpha}^{[l-1]}\right)\cdot\boldsymbol{W}_{p,q,\alpha,\beta}^{[l]}\cdot\chi(p,q)\right)\right]+\boldsymbol{b}_{\beta}^{[l]},$$

$$f(\boldsymbol{x},\boldsymbol{\theta}) := f_{\mathrm{CNN}}(\boldsymbol{x},\boldsymbol{\theta}) := \left\langle \boldsymbol{a}, \sigma\left(\boldsymbol{x}^{[L]}\right)\right\rangle = \sum_{\beta=1}^{C_L}\sum_{u=1}^{W_L}\sum_{v=1}^{H_L}\boldsymbol{a}_{u,v,\beta}\cdot\sigma\left(\boldsymbol{x}_{u,v,\beta}^{[L]}\right),$$

where $\sigma(\cdot)$ is the activation function applied coordinate-wisely to its input, and for each layer $l \in [L]$, all parameters belonging to this layer are initialized by: For $p, q \in [m-1]$, $\alpha \in [C_{l-1}]$ and $\beta \in [C_l]$,

$$\boldsymbol{W}_{p,q,\alpha,\beta}^{[l]} \sim \mathcal{N}(0,\beta_1^2), \quad \boldsymbol{b}_{\beta}^{[l]} \sim \mathcal{N}(0,\beta_1^2). \tag{18}$$

Moreover, for $u \in [W_L]$ and $v \in [H_L]$,

$$\boldsymbol{a}_{u,v,\beta} \sim \mathcal{N}(0,\beta_2^2), \tag{19}$$

and for convenience we set $\beta_1 = \beta_2 = \varepsilon$, where $\varepsilon > 0$ is the scaling parameter. Finally, for all $i \in [n]$, we denote hereafter that

$$e_i := e_i(\boldsymbol{\theta}) := f(\boldsymbol{x}_i,\boldsymbol{\theta}) - y_i,$$

and

$$\boldsymbol{e} := \boldsymbol{e}(\boldsymbol{\theta}) := [e_1(\boldsymbol{\theta}), e_2(\boldsymbol{\theta}), \ldots, e_n(\boldsymbol{\theta})]^{\intercal} \in \mathbb{R}^n.$$

B.3 GD Dynamics

In this paper, we train all layers of the neural network with continuous time gradient descent (GD): For any time $t \geq 0$,

$$\frac{\mathrm{d}\boldsymbol{\theta}}{\mathrm{d}t} = -\nabla_{\boldsymbol{\theta}} R_S(\boldsymbol{\theta}). \tag{20}$$

We remark that details of the dynamics (20) are hard to write out in matrix form, so we turn to the alternative approach to derive the GD dynamics of each individual parameter. In order for that, it is natural for us to define a series of auxiliary variables: For each $i \in [n]$ and $l \in [L]$, we define $\boldsymbol{z}_{u,v,\beta}^{[l]}(i)$ as the partial derivative of $f(\boldsymbol{x}_i,\boldsymbol{\theta})$ with respect to $\boldsymbol{x}_{u,v,\beta}^{[l]}(i)$, i.e.,

$$\boldsymbol{z}_{u,v,\beta}^{[l]}(i) := \frac{\partial f(\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{x}_{u,v,\beta}^{[l]}(i)}, \tag{21}$$

we obtain that for $l \in [L-1]$,

$$\boldsymbol{z}_{u,v,\beta}^{[l]}(i) = \sum_{\alpha=1}^{C_{l+1}}\sum_{s=1}^{W_{l+1}}\sum_{t=1}^{H_{l+1}}\boldsymbol{z}_{s,t,\alpha}^{[l+1]}(i)\cdot\boldsymbol{W}_{u-s,v-t,\beta,\alpha}^{[l+1]}\cdot\sigma^{(1)}\left(\boldsymbol{x}_{u,v,\beta}^{[l]}(i)\right)\cdot\chi(u-s,v-t),$$

$$\boldsymbol{z}_{u,v,\beta}^{[L]}(i) = \boldsymbol{a}_{u,v,\beta}\cdot\sigma^{(1)}\left(\boldsymbol{x}_{u,v,\beta}^{[L]}(i)\right),$$

hence we obtain that for $l \in [2:L]$,

$$\begin{aligned}
\frac{\partial f(\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{W}_{p,q,\alpha,\beta}^{[1]}} &= \sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{z}_{u,v,\beta}^{[1]}(i)\cdot\boldsymbol{x}_{u+p,v+q,\alpha}^{[0]}(i),\\
\frac{\partial f(\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{W}_{p,q,\alpha,\beta}^{[l]}} &= \sum_{u=1}^{W_l}\sum_{v=1}^{H_l}\boldsymbol{z}_{u,v,\beta}^{[l]}(i)\cdot\sigma\left(\boldsymbol{x}_{u+p,v+q,\alpha}^{[l-1]}(i)\right),\\
\frac{\partial f(\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{b}_{\beta}^{[1]}} &= \sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{z}_{u,v,\beta}^{[1]}(i),\\
\frac{\partial f(\boldsymbol{x}_i,\boldsymbol{\theta})}{\partial \boldsymbol{b}_{\beta}^{[l]}} &= \sum_{u=1}^{W_l}\sum_{v=1}^{H_l}\boldsymbol{z}_{u,v,\beta}^{[l]}(i).
\end{aligned} \tag{22}$$

With the above notations, for $l \in [2:L]$, the dynamics (20) reads

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}^{[1]}_{p,q,\alpha,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{z}^{[1]}_{u,v,\beta}(i) \cdot \boldsymbol{x}^{[0]}_{u+p,v+q,\alpha}(i) \right), \\
\frac{\mathrm{d}\boldsymbol{W}^{[l]}_{p,q,\alpha,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_l}\sum_{v=1}^{H_l} \boldsymbol{z}^{[l]}_{u,v,\beta}(i) \cdot \sigma\left(\boldsymbol{x}^{[l-1]}_{u+p,v+q,\alpha}(i) \right) \right), \\
\frac{\mathrm{d}\boldsymbol{b}^{[1]}_{\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{z}^{[1]}_{u,v,\beta}(i) \right), \\
\frac{\mathrm{d}\boldsymbol{b}^{[l]}_{\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_l}\sum_{v=1}^{H_l} \boldsymbol{z}^{[l]}_{u,v,\beta}(i) \right), \\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \sigma\left(\boldsymbol{x}^{[L]}_{u,v,\beta}(i) \right).
\end{aligned} \tag{23}$$

C Two-Layer CNNs with Single Channel

C.1 Activation Function

In this part, we shall impose some technical conditions on the activation function and input samples. We start with a technical condition (Zhou et al., 2022, Definition 1) on the activation function $\sigma(\cdot)$

Definition 2 (Multiplicity r). *$\sigma(\cdot) : \mathbb{R} \to \mathbb{R}$ has multiplicity r if there exists an integer $r \geq 1$, such that for all $0 \leq s \leq r-1$, the s-th order derivative satisfies $\sigma^{(s)}(0) = 0$, and $\sigma^{(r)}(0) \neq 0$.*

We list out some examples of activation functions with different multiplicity.

Remark 2.

- $\tanh(\mathrm{x}) := \frac{\exp(\mathrm{x})-\exp(-\mathrm{x})}{\exp(\mathrm{x})+\exp(-\mathrm{x})}$ *is with multiplicity $r = 1$;*
- $\mathrm{SiLU}(\mathrm{x}) := \frac{\mathrm{x}}{1+\exp(-\mathrm{x})}$ *is with multiplicity $r = 1$;*
- $\mathrm{xtanh}(\mathrm{x}) := \frac{\mathrm{x}\exp(\mathrm{x})-\mathrm{x}\exp(-\mathrm{x})}{\exp(\mathrm{x})+\exp(-\mathrm{x})}$ *is with multiplicity $r = 2$.*

Assumption 5 (Multiplicity 1). *The activation function $\sigma \in \mathcal{C}^2(\mathbb{R})$, and there exists a universal constant $C_L > 0$, such that its first and second derivatives satisfy*

$$\left\| \sigma^{(1)}(\cdot) \right\|_{\infty} \leq C_L, \quad \left\| \sigma^{(2)}(\cdot) \right\|_{\infty} \leq C_L. \tag{24}$$

Moreover,

$$\sigma(0) = 0, \quad \sigma^{(1)}(0) = 1. \tag{25}$$

Remark 3. *We remark that σ has multiplicity 1. $\sigma^{(1)}(0) = 1$ can be replaced by $\sigma^{(1)}(0) \neq 0$, and we set $\sigma^{(1)}(0) = 1$ for simplicity, and it can be easily satisfied by replacing the original activation $\sigma(\cdot)$ with $\frac{\sigma(\cdot)}{\sigma^{(1)}(0)}$.*

We note that Assumption 5 can be satisfied by using the $\tanh$ *activation:*

$$\sigma(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)},$$

and the scaled SiLU activation

$$\sigma(x) = \frac{2x}{1 + \exp(-x)}.$$

In the case of two-layer CNNs, as $L = 2$, then we shall set $\boldsymbol{W}_{p,q,\alpha,\beta} := \boldsymbol{W}^{[1]}_{p,q,\alpha,\beta}$, $\boldsymbol{b}_\beta := \boldsymbol{b}^{[1]}_\beta$, and $\boldsymbol{x}_{u+p,v+q,\alpha}(i) := \boldsymbol{x}^{[0]}_{u+p,v+q,\alpha}(i)$ for simplicity, then (23) reads

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}_{p,q,\alpha,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right) \cdot \boldsymbol{x}_{u+p,v+q,\alpha}(i) \right), \\
\frac{\mathrm{d}\boldsymbol{b}_\beta}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right) \right), \\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \sigma\left(\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right).
\end{aligned}$$

Moreover, since the MNIST Deng (2012) images are black and white, therefore we do not need three different color-channels to represent the final color, and only one channel is enough, i.e., $C_0 = 1$, hence the above dynamics can be further simplified into

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}_{p,q,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right) \cdot \boldsymbol{x}_{u+p,v+q}(i) \right), \\
\frac{\mathrm{d}\boldsymbol{b}_\beta}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right) \right), \\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \sigma\left(\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right).
\end{aligned} \tag{26}$$

We identify the vectorized parameters $\boldsymbol{\theta}$ as variables of order 1 by setting $\boldsymbol{\theta} = \varepsilon\bar{\boldsymbol{\theta}}$, and

$$\begin{aligned}
\boldsymbol{x}^{[1]}_{u,v,\beta} &= \left(\sum_{p=-\infty}^{+\infty}\sum_{q=-\infty}^{+\infty} \boldsymbol{x}_{u+p,v+q} \cdot \varepsilon\overline{\boldsymbol{W}}_{p,q,\beta} \cdot \chi(p,q) \right) + \varepsilon\bar{\boldsymbol{b}}^{[1]}_\beta \\
&= \varepsilon\bar{\boldsymbol{x}}^{[1]}_{u,v,\beta},
\end{aligned} \tag{27}$$

where

$$\bar{\boldsymbol{x}}^{[1]}_{u,v,\beta} := \left(\sum_{p=-\infty}^{+\infty}\sum_{q=-\infty}^{+\infty} \boldsymbol{x}_{u+p,v+q} \cdot \overline{\boldsymbol{W}}_{p,q,\beta} \cdot \chi(p,q) \right) + \bar{\boldsymbol{b}}^{[1]}_\beta,$$

is also of order 1, and the rescaled dynamics can be written into

$$\begin{aligned}
\frac{\mathrm{d}\overline{\boldsymbol{W}}_{p,q,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \bar{\boldsymbol{a}}_{u,v,\beta} \cdot \sigma^{(1)}\left(\varepsilon\bar{\boldsymbol{x}}^{[1]}_{u,v,\beta}(i)\right) \cdot \boldsymbol{x}_{u+p,v+q}(i) \right), \\
\frac{\mathrm{d}\bar{\boldsymbol{b}}_\beta}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \bar{\boldsymbol{a}}_{u,v,\beta} \cdot \sigma^{(1)}\left(\varepsilon\bar{\boldsymbol{x}}^{[1]}_{u,v,\beta}(i)\right) \right), \\
\frac{\mathrm{d}\bar{\boldsymbol{a}}_{u,v,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \frac{\sigma\left(\varepsilon\bar{\boldsymbol{x}}^{[1]}_{u,v,\beta}(i)\right)}{\varepsilon},
\end{aligned}$$

with the following initialization

$$\overline{\boldsymbol{W}}^0_{p,q,\beta} \sim \mathcal{N}(0,1), \quad \bar{\boldsymbol{b}}^0_\beta \sim \mathcal{N}(0,1), \quad \bar{\boldsymbol{a}}^0_{u,v,\beta} \sim \mathcal{N}(0,1). \tag{28}$$

In the following discussion throughout this paper, we always refer to the above rescaled dynamics and drop all the "bar"s of $\bar{\boldsymbol{\theta}}$, $\overline{\boldsymbol{W}}_{p,q,\beta}$, $\bar{\boldsymbol{b}}_\beta$, $\bar{\boldsymbol{x}}^{[1]}_{u,v,\beta}$ and $\bar{\boldsymbol{a}}_{u,v,\beta}$ for notational simplicity. Moreover, we remark that $p \in [0 : m-1]$, $q \in [0 : m-1]$ and $\beta \in [M]$, where m is the filter size, and $M := C_1$, the number of channels in $\boldsymbol{x}^{[1]}(i)$, which can be heuristically understood as the 'width' of the hidden layer in the case of two-layer neural networks (NNs). Before we end this section, we assume hereafter that

Assumption 6. *The training inputs $\{\boldsymbol{x}_i\}_{i=1}^n$ and labels $\{y_i\}_{i=1}^n$ satisfy that there exists a universal constant $c > 0$, such that given any $i \in [n]$, then for each $u \in [W_0]$, $v \in [H_0]$ and $\alpha \in [C_0]$, the following holds*

$$\frac{1}{c} \leq \left|\boldsymbol{x}_{u,v,\alpha}(i)\right|, \quad |y_i| \leq c.$$

We assume further that

Assumption 7. *The following limit exists*

$$\gamma := \lim_{M\to\infty} -\frac{\log \varepsilon^2}{\log M}. \tag{29}$$

C.2 Effective Linear Dynamics

As the normalized flow reads

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}_{p,q,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^n e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right) \cdot \boldsymbol{x}_{u+p,v+q}(i)\right), \\
\frac{\mathrm{d}\boldsymbol{b}_{\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^n e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right)\right), \\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^n e_i \cdot \frac{\sigma\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right)}{\varepsilon},
\end{aligned}$$

since $e_i \approx -y_i$, and by means of perturbation expansion with respect to ε and keep the order 1 term, we obtain that

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}_{p,q,\beta}}{\mathrm{d}t} &\approx \frac{1}{n}\sum_{i=1}^n y_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \boldsymbol{x}_{u+p,v+q}(i)\right), \\
\frac{\mathrm{d}\boldsymbol{b}_{\beta}}{\mathrm{d}t} &\approx \frac{1}{n}\sum_{i=1}^n y_i \cdot \sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta}, \\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &\approx \frac{1}{n}\sum_{i=1}^n y_i \cdot \boldsymbol{x}_{u,v,\beta}^{[1]}(i) \\
&= \frac{1}{n}\sum_{i=1}^n y_i \cdot \left[\left(\sum_{p=0}^{m-1}\sum_{q=0}^{m-1} \boldsymbol{x}_{u+p,v+q}(i) \cdot \boldsymbol{W}_{p,q,\beta}\right) + \boldsymbol{b}_{\beta}\right].
\end{aligned} \tag{30}$$

Given any $u \in [W_1]$ and $v \in [H_1]$, then for all $p, q \in [0 : m-1]$, we set

$$\begin{aligned}
\boldsymbol{z}_{u+p,v+q} &:= \frac{1}{n}\sum_{i=1}^n y_i \boldsymbol{x}_{u+p,v+q}(i), \\
z &:= \frac{1}{n}\sum_{i=1}^n y_i,
\end{aligned} \tag{31}$$

then the dynamics (30) can be further simplified into: For any $\beta \in [M]$,

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}_{p,q,\beta}}{\mathrm{d}t} &\approx \sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \boldsymbol{z}_{u+p,v+q}, \\
\frac{\mathrm{d}\boldsymbol{b}_{\beta}}{\mathrm{d}t} &\approx \sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot z, \\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &\approx \left(\sum_{p=0}^{m-1}\sum_{q=0}^{m-1} \boldsymbol{z}_{u+p,v+q} \cdot \boldsymbol{W}_{p,q,\beta}\right) + \boldsymbol{b}_{\beta} \cdot z.
\end{aligned} \tag{32}$$

We observe that (32) reveals that the training dynamics of two-layer CNNs at initial stage has a close relationship to power iteration of a matrix $\boldsymbol{A}$ that only depends on the input samples, i.e., the dynamics (32) takes the form

$$\frac{\mathrm{d}\boldsymbol{\theta}_\beta}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{\theta}_\beta, \tag{33}$$

where

$$\begin{aligned}\boldsymbol{\theta}_\beta := \Big(&\boldsymbol{W}_{0,0,\beta}, \boldsymbol{W}_{0,1,\beta}, \cdots, \boldsymbol{W}_{0,m-1,\beta}; \boldsymbol{W}_{1,0,\beta}, \cdots, \boldsymbol{W}_{1,m-1,\beta}; \cdots\cdots \boldsymbol{W}_{m-1,m-1,\beta}; \boldsymbol{b}_\beta; \\ &\boldsymbol{a}_{1,1,\beta}, \boldsymbol{a}_{1,2,\beta}, \cdots, \boldsymbol{a}_{1,H_1,\beta}; \boldsymbol{a}_{2,1,\beta}, \cdots, \boldsymbol{a}_{2,H_1,\beta}; \cdots\cdots \boldsymbol{a}_{W_1,H_1,\beta}\Big)^\intercal,\end{aligned}$$

and if we would like to simplify our notations

$$\begin{aligned}\boldsymbol{\theta}_\beta = \Big(&\boldsymbol{W}_{0,0:(m-1),\beta}; \boldsymbol{W}_{1,0:(m-1),\beta}; \cdots\cdots \boldsymbol{W}_{m-1,0:(m-1),\beta}; \boldsymbol{b}_\beta; \\ &\boldsymbol{a}_{1,1:H_1,\beta}; \boldsymbol{a}_{2,1:H_1,\beta}; \cdots\cdots \boldsymbol{a}_{W_1,1:H_1,\beta}\Big)^\intercal,\end{aligned}$$

$$\boldsymbol{A} := \begin{bmatrix} \boldsymbol{0}_{(m^2+1)\times(m^2+1)} & \boldsymbol{Z}^\intercal \\ \boldsymbol{Z} & \boldsymbol{0}_{W_1H_1\times W_1H_1} \end{bmatrix}, \tag{34}$$

where $\boldsymbol{Z} \in \mathbb{R}^{W_1H_1\times(m^2+1)}$ and $\boldsymbol{Z}$ depends sorely on the input samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, whose entries read

$$\boldsymbol{Z} := \begin{bmatrix}
\boldsymbol{z}_{1,1} & \boldsymbol{z}_{1,2} & \cdots & \boldsymbol{z}_{1,m}; & \boldsymbol{z}_{2,1} & \cdots & \boldsymbol{z}_{2,m}; & \cdots\cdots & \boldsymbol{z}_{m,m}; & z \\
\boldsymbol{z}_{1,2} & \boldsymbol{z}_{1,3} & \cdots & \boldsymbol{z}_{1,m+1}; & \boldsymbol{z}_{2,2} & \cdots & \boldsymbol{z}_{2,m+1}; & \cdots\cdots & \boldsymbol{z}_{m,m+1}; & z \\
\vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{1,H_1} & \boldsymbol{z}_{1,H_1+1} & \cdots & \boldsymbol{z}_{1,H_0}; & \boldsymbol{z}_{2,H_1} & \cdots & \boldsymbol{z}_{2,H_0}; & \cdots\cdots & \boldsymbol{z}_{m,H_0}; & z \\
\boldsymbol{z}_{2,1} & \boldsymbol{z}_{2,2} & \cdots & \boldsymbol{z}_{2,m}; & \boldsymbol{z}_{3,1} & \cdots & \boldsymbol{z}_{3,m}; & \cdots\cdots & \boldsymbol{z}_{m+1,m}; & z \\
\vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{2,H_1} & \boldsymbol{z}_{2,H_1+1} & \cdots & \boldsymbol{z}_{2,H_0}; & \boldsymbol{z}_{3,H_1} & \cdots & \boldsymbol{z}_{3,H_0}; & \cdots\cdots & \boldsymbol{z}_{m+1,H_0}; & z \\
\vdots & \vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{W_1,H_1} & \boldsymbol{z}_{W_1,H_1+1} & \cdots & \boldsymbol{z}_{W_1,H_0}; & \boldsymbol{z}_{W_1+1,H_1} & \cdots & \boldsymbol{z}_{W_1+1,H_0}; & \cdots\cdots & \boldsymbol{z}_{W_0,H_0}; & z
\end{bmatrix}. \tag{35}$$

If we would like to simplify our notations,

$$\boldsymbol{Z} := \begin{bmatrix}
\boldsymbol{z}_{1,1:m}; & \boldsymbol{z}_{2,1:m}; & \cdots\cdots & \boldsymbol{z}_{m,1:m}; & z \\
\boldsymbol{z}_{1,2:(m+1)}; & \boldsymbol{z}_{2,2:(m+1)}; & \cdots\cdots & \boldsymbol{z}_{m,2:(m+1)}; & z \\
\vdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{1,H_1:H_0}; & \boldsymbol{z}_{2,H_1:H_0}; & \cdots\cdots & \boldsymbol{z}_{m,H_1:H_0}; & z \\
\boldsymbol{z}_{2,1:m}; & \boldsymbol{z}_{3,1:m}; & \cdots\cdots & \boldsymbol{z}_{m+1,1:m}; & z \\
\vdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{2,H_1:H_0}; & \boldsymbol{z}_{3,H_1:H_0}; & \cdots\cdots & \boldsymbol{z}_{m+1,H_1:H_0}; & z \\
\vdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{W_1,H_1:H_0}; & \boldsymbol{z}_{W_1+1,H_1:H_0}; & \cdots\cdots & \boldsymbol{z}_{W_0,H_1:H_0}; & z
\end{bmatrix}.$$

In order to solve out the simplified dynamics (32), we need to perform Singular value decomposition (SVD) on $\boldsymbol{Z}$, i.e.,

$$\boldsymbol{Z} = \boldsymbol{U}\Lambda\boldsymbol{V}^\intercal, \tag{36}$$

where

$$\boldsymbol{U} = [\boldsymbol{u}_1, \boldsymbol{u}_2, \cdots, \boldsymbol{u}_{W_1H_1}], \quad \boldsymbol{V} = [\boldsymbol{v}_1, \boldsymbol{v}_2, \cdots, \boldsymbol{v}_{m^2+1}],$$

and as we denote $r := \mathrm{rank}(\boldsymbol{Z})$, naturally, $r \le \min\{W_1H_1, m^2+1\}$ we have r singular values,

$$\lambda_1 \ge \lambda_2 \ge \cdots \ge \lambda_r > 0,$$

and WLOG, we assume that

Assumption 8 (Spectral Gap of $\boldsymbol{Z}$). *The singular values $\{\lambda_k\}_{k=1}^r$ of $\boldsymbol{Z}$ satisfy that*

$$\lambda_1 > \lambda_2 \geq \cdots \geq \lambda_r > 0, \tag{37}$$

and we denote the spectral gap between λ_1 and λ_2 by

$$\Delta\lambda := \lambda_1 - \lambda_2.$$

Moreover, as we denote further that

$$\begin{aligned}\boldsymbol{\theta}_{\boldsymbol{W},\beta} &:= (\boldsymbol{W}_{0,0,\beta}, \boldsymbol{W}_{0,1,\beta}, \cdots, \boldsymbol{W}_{0,m-1,\beta}; \boldsymbol{W}_{1,0,\beta}, \cdots, \boldsymbol{W}_{1,m-1,\beta}; \cdots\cdots \boldsymbol{W}_{m-1,m-1,\beta}; \boldsymbol{b}_\beta)^\intercal, \\ \boldsymbol{\theta}_{\boldsymbol{a},\beta} &:= (\boldsymbol{a}_{1,1,\beta}, \boldsymbol{a}_{1,2,\beta}, \cdots, \boldsymbol{a}_{1,H_1,\beta}; \boldsymbol{a}_{2,1,\beta}, \cdots, \boldsymbol{a}_{2,H_1,\beta}; \cdots\cdots \boldsymbol{a}_{W_1,H_1,\beta})^\intercal,\end{aligned}$$

hence

$$\boldsymbol{\theta}_\beta = \left(\boldsymbol{\theta}_{\boldsymbol{W},\beta}^\intercal, \boldsymbol{\theta}_{\boldsymbol{a},\beta}^\intercal\right)^\intercal,$$

and the linear dynamics (32) read

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{W},\beta}}{\mathrm{d}t} &= \boldsymbol{Z}^\intercal \boldsymbol{\theta}_{\boldsymbol{a},\beta}, \quad \boldsymbol{\theta}_{\boldsymbol{W},\beta}(0) = \boldsymbol{\theta}_{\boldsymbol{W},\beta}^0, \\ \frac{\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{a},\beta}}{\mathrm{d}t} &= \boldsymbol{Z}\boldsymbol{\theta}_{\boldsymbol{W},\beta}, \quad \boldsymbol{\theta}_{\boldsymbol{a},\beta}(0) = \boldsymbol{\theta}_{\boldsymbol{a},\beta}^0,\end{aligned} \tag{38}$$

hence (38) can be simplified further into two separate second order differential equations,

$$\frac{\mathrm{d}^2\boldsymbol{\theta}_{\boldsymbol{W},\beta}}{\mathrm{d}t^2} = \boldsymbol{Z}^\intercal\boldsymbol{Z}\boldsymbol{\theta}_{\boldsymbol{W},\beta}, \;\; \boldsymbol{\theta}_{\boldsymbol{W},\beta}(0) = \boldsymbol{\theta}_{\boldsymbol{W},\beta}^0, \;\; \frac{\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{W},\beta}}{\mathrm{d}t}(0) = \boldsymbol{Z}^\intercal\boldsymbol{\theta}_{\boldsymbol{a},\beta}^0, \tag{39}$$

and

$$\frac{\mathrm{d}^2\boldsymbol{\theta}_{\boldsymbol{a},\beta}}{\mathrm{d}t^2} = \boldsymbol{Z}\boldsymbol{Z}^\intercal\boldsymbol{\theta}_{\boldsymbol{a},\beta}, \;\; \boldsymbol{\theta}_{\boldsymbol{a},\beta}(0) = \boldsymbol{\theta}_{\boldsymbol{a},\beta}^0, \;\; \frac{\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{a},\beta}}{\mathrm{d}t}(0) = \boldsymbol{Z}\boldsymbol{\theta}_{\boldsymbol{W},\beta}^0. \tag{40}$$

We observe that

$$\begin{aligned}\boldsymbol{Z}^\intercal\boldsymbol{Z} &= \sum_{k=1}^r \lambda_k^2 \boldsymbol{v}_k\boldsymbol{v}_k^\intercal, \\ \boldsymbol{Z}\boldsymbol{Z}^\intercal &= \sum_{k=1}^r \lambda_k^2 \boldsymbol{u}_k\boldsymbol{u}_k^\intercal,\end{aligned}$$

hence the solutions to (39) and (40) respectively reads

$$\begin{aligned}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(t) &= \left(\sum_{k=1}^r \left[c_{\lambda_k,\boldsymbol{W},\beta}\exp(\lambda_k t) + d_{\lambda_k,\boldsymbol{W},\beta}\exp(-\lambda_k t)\right]\boldsymbol{v}_k\right) \\ &\quad + \mathcal{P}_{(r+1):(m^2+1)}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(0), \\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}(t) &= \left(\sum_{k=1}^r \left[c_{\lambda_k,\boldsymbol{a},\beta}\exp(\lambda_k t) + d_{\lambda_k,\boldsymbol{a},\beta}\exp(-\lambda_k t)\right]\boldsymbol{u}_k\right) \\ &\quad + \mathcal{P}_{(r+1):(W_1H_1)}\boldsymbol{\theta}_{\boldsymbol{a},\beta}(0),\end{aligned} \tag{41}$$

where for each $k \in [r]$ and $\beta \in [M]$, the constants $c_{\lambda_k,\boldsymbol{W},\beta}$, $d_{\lambda_k,\boldsymbol{W},\beta}$, $c_{\lambda_k,\boldsymbol{a},\beta}$ and $d_{\lambda_k,\boldsymbol{a},\beta}$ depend on $\left\langle\boldsymbol{\theta}_{\boldsymbol{W},\beta}^0, \boldsymbol{v}_k\right\rangle$ and $\left\langle\boldsymbol{\theta}_{\boldsymbol{a},\beta}^0, \boldsymbol{u}_k\right\rangle$, i.e., the constants $c_{\lambda_k,\boldsymbol{W},\beta}$ and $d_{\lambda_k,\boldsymbol{W},\beta}$ are determined by,

$$\begin{aligned}c_{\lambda_k,\boldsymbol{W},\beta} + d_{\lambda_k,\boldsymbol{W},\beta} &= \left\langle\boldsymbol{\theta}_{\boldsymbol{W},\beta}^0, \boldsymbol{v}_k\right\rangle, \\ \lambda_k c_{\lambda_k,\boldsymbol{W},\beta} - \lambda_k d_{\lambda_k,\boldsymbol{W},\beta} &= \left\langle\boldsymbol{Z}^\intercal\boldsymbol{\theta}_{\boldsymbol{a},\beta}^0, \boldsymbol{v}_k\right\rangle = \lambda_k\left\langle\boldsymbol{\theta}_{\boldsymbol{a},\beta}^0, \boldsymbol{u}_k\right\rangle,\end{aligned} \tag{42}$$

thus

$$\begin{aligned}c_{\lambda_k,\boldsymbol{W},\beta} &= \frac{1}{2}\left(\left\langle\boldsymbol{\theta}_{\boldsymbol{W},\beta}^0, \boldsymbol{v}_k\right\rangle + \left\langle\boldsymbol{\theta}_{\boldsymbol{a},\beta}^0, \boldsymbol{u}_k\right\rangle\right), \\ d_{\lambda_k,\boldsymbol{W},\beta} &= \frac{1}{2}\left(\left\langle\boldsymbol{\theta}_{\boldsymbol{W},\beta}^0, \boldsymbol{v}_k\right\rangle - \left\langle\boldsymbol{\theta}_{\boldsymbol{a},\beta}^0, \boldsymbol{u}_k\right\rangle\right),\end{aligned}$$

which matches the same constants as the ones in two-layer NNs Chen et al. (2023), a special case where $r = 1$. Then with slight misuse of notations, $\mathcal{P}_{1:r}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(0)$ refers to the projection of $\boldsymbol{\theta}_{\boldsymbol{W},\beta}^0$ towards $\operatorname{span}\{\boldsymbol{v}_k\}_{k=1}^r$, $\mathcal{P}_{(r+1):(m^2+1)}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(0)$ refers to the projection of $\boldsymbol{\theta}_{\boldsymbol{W},\beta}^0$ towards $\operatorname{span}\{\boldsymbol{v}_k\}_{k=r+1}^{m^2+1}$, and $\mathcal{P}_{(r+1):(W_1H_1)}\boldsymbol{\theta}_{\boldsymbol{a},\beta}(0)$ refers to the projection of $\boldsymbol{\theta}_{\boldsymbol{a},\beta}^0$ towards $\operatorname{span}\{\boldsymbol{u}_k\}_{k=r+1}^{W_1H_1}$.

Proposition 1. *The solution to the linear differential equation*

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{W},\beta}}{\mathrm{d}t} &= \boldsymbol{Z}^{\intercal}\boldsymbol{\theta}_{\boldsymbol{a},\beta}, \quad \boldsymbol{\theta}_{\boldsymbol{W},\beta}(0) = \boldsymbol{\theta}^0_{\boldsymbol{W},\beta},\\ \frac{\mathrm{d}\boldsymbol{\theta}_{\boldsymbol{a},\beta}}{\mathrm{d}t} &= \boldsymbol{Z}\boldsymbol{\theta}_{\boldsymbol{W},\beta}, \quad \boldsymbol{\theta}_{\boldsymbol{a},\beta}(0) = \boldsymbol{\theta}^0_{\boldsymbol{a},\beta},\end{aligned} \tag{43}$$

reads

$$\begin{aligned}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(t) &= \left(\sum_{k=1}^{r}\left[c_{\lambda_k,\boldsymbol{W},\beta}\exp(\lambda_k t) + d_{\lambda_k,\boldsymbol{W},\beta}\exp(-\lambda_k t)\right]\boldsymbol{v}_k\right)\\ &\quad + \mathcal{P}_{(r+1):(m^2+1)}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(0),\\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}(t) &= \left(\sum_{k=1}^{r}\left[c_{\lambda_k,\boldsymbol{a},\beta}\exp(\lambda_k t) + d_{\lambda_k,\boldsymbol{a},\beta}\exp(-\lambda_k t)\right]\boldsymbol{u}_k\right)\\ &\quad + \mathcal{P}_{(r+1):(W_1H_1)}\boldsymbol{\theta}_{\boldsymbol{a},\beta}(0).\end{aligned} \tag{44}$$

Remark 4. *It is noteworthy that $\boldsymbol{\theta}_{\boldsymbol{W},\beta}$ shall be understood as two components, one is the projection of $\boldsymbol{\theta}_{\boldsymbol{W},\beta}$ into* $\operatorname{span}\{\boldsymbol{v}_k\}_{k=1}^{r}$,

$$\mathcal{P}_{1:r}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(t) := \left(\sum_{k=1}^{r}\left[c_{\lambda_k,\boldsymbol{W},\beta}\exp(\lambda_k t) + d_{\lambda_k,\boldsymbol{W},\beta}\exp(-\lambda_k t)\right]\boldsymbol{v}_k\right),$$

which evolves with respect to time t, and the other is the projection of $\boldsymbol{\theta}_{\boldsymbol{W},\beta}$ into $(\operatorname{span}\{\boldsymbol{v}_k\}_{k=1}^{r})^{\perp} = \operatorname{span}\{\boldsymbol{v}_k\}_{k=r+1}^{m^2+1}$,

$$\mathcal{P}_{(r+1):(m^2+1)}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(t) = \mathcal{P}_{(r+1):(m^2+1)}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(0),$$

which remains frozen as t evolves.

C.3 Difference between Real and Linear Dynamics

For any $\beta \in [M]$, the real dynamics (26) can be written into

$$\begin{aligned}\frac{\mathrm{d}\boldsymbol{W}_{p,q,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{a}_{u,v,\beta}\cdot\sigma^{(1)}\left(\varepsilon\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right)\cdot\boldsymbol{x}_{u+p,v+q}(i)\right)\\ &\quad - \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{a}_{u,v,\beta}\cdot\boldsymbol{x}_{u+p,v+q}(i)\right)\\ &\quad + \sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{a}_{u,v,\beta}\cdot\boldsymbol{z}_{u+p,v+q},\\ \frac{\mathrm{d}\boldsymbol{b}_{\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{a}_{u,v,\beta}\cdot\sigma^{(1)}\left(\varepsilon\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right)\right)\\ &\quad - \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{a}_{u,v,\beta}\right)\\ &\quad + \sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{a}_{u,v,\beta}\cdot z,\\ \frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &= -\frac{1}{n}\sum_{i=1}^{n} e_i \cdot \frac{\sigma\left(\varepsilon\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\right)}{\varepsilon} - \frac{1}{n}\sum_{i=1}^{n} y_i\cdot\boldsymbol{x}^{[1]}_{u,v,\beta}(i)\\ &\quad + \left(\sum_{p=0}^{m-1}\sum_{q=0}^{m-1}\boldsymbol{z}_{u+p,v+q}\cdot\boldsymbol{W}_{p,q,\beta}\right) + \boldsymbol{b}_{\beta}\cdot z.\end{aligned}$$

Hence the difference between the real and linear dynamics is characterized by $\{f_{p,q,\beta}, f_\beta, g_{u,v,\beta}\}_{p,q\in[0:m-1],u\in[W_1],v\in[H_1],\beta\in[M]}$, where

$$\begin{aligned}
f_{p,q,\beta} &:= \frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right) \cdot \boldsymbol{x}_{u+p,v+q}(i)\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \boldsymbol{x}_{u+p,v+q}(i)\right), \\
f_\beta &:= \frac{1}{n}\sum_{i=1}^{n} e_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right)\right) \\
&\quad + \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta}\right), \\
g_{u,v,\beta} &:= \frac{1}{n}\sum_{i=1}^{n} e_i \cdot \frac{\sigma\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right)}{\varepsilon} + \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \boldsymbol{x}_{u,v,\beta}^{[1]}(i),
\end{aligned}$$

and for each $\beta \in [M]$, we set

$$\begin{aligned}
\boldsymbol{f}_\beta &:= (f_{0,0,\beta}, f_{0,1,\beta}, \cdots, f_{0,m-1,\beta}; f_{1,0,\beta}, \cdots, f_{1,m-1,\beta}; \cdots\cdots f_{m-1,m-1,\beta}; f_\beta)^{\mathsf{T}}, \\
\boldsymbol{g}_\beta &:= (g_{1,1,\beta}, g_{1,2,\beta}, \cdots, g_{1,H_1,\beta}; g_{2,1,\beta}, \cdots, g_{2,H_1,\beta}; \cdots\cdots g_{W_1,H_1,\beta})^{\mathsf{T}},
\end{aligned}$$

and we observe further that for any $\beta \in [M]$, the real dynamics read

$$\frac{\mathrm{d}\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta} \\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix}}{\mathrm{d}t} = \begin{pmatrix}\boldsymbol{f}_\beta \\ \boldsymbol{g}_\beta\end{pmatrix} + \boldsymbol{A}\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta} \\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix}. \tag{45}$$

Definition 3 (Neuron energy)**.** *In real dynamics, we define the energy at time t for each $\beta \in [M]$,*

$$E_\beta(t) := \left(\|\boldsymbol{\theta}_{\boldsymbol{W},\beta}(t)\|_2^2 + \|\boldsymbol{\theta}_{\boldsymbol{a},\beta}(t)\|_2^2\right)^{\frac{1}{2}}, \tag{46}$$

and we denote

$$E_{\max}(t) := \max_{\beta\in[M]} E_\beta(t). \tag{47}$$

For simplicity, we hereafter drop the (t)s for all $E_\beta(t)$ and $E_{\max}(t)$. Then the estimates on $\{\boldsymbol{f}_\beta, \boldsymbol{g}_\beta\}_{\beta=1}^M$ read

Proposition 2. *For any $\varepsilon > 0$ and any time $t > 0$,*

$$\begin{aligned}
\|\boldsymbol{f}_\beta\|_2 &\le \left(M\varepsilon^2 E_{\max}^2 + \varepsilon E_{\max}\right)\|\boldsymbol{\theta}_{\boldsymbol{a},\beta}\|_2, \\
\|\boldsymbol{g}_\beta\|_2 &\le \left(M\varepsilon^2 E_{\max}^2 + \varepsilon E_{\max}\right)\|\boldsymbol{\theta}_{\boldsymbol{W},\beta}\|_2.
\end{aligned} \tag{48}$$

Moreover, we obtain that

$$\left(\|\boldsymbol{f}_\beta\|_2 + \|\boldsymbol{g}_\beta\|_2^2\right)^{\frac{1}{2}} \le \left(M\varepsilon^2 E_{\max}^2 + \varepsilon E_{\max}\right) E_\beta \le \left(M\varepsilon^2 E_{\max}^2 + \varepsilon E_{\max}\right) E_{\max}.$$

Proof. We obtain that for each $i \in [n]$,

$$
\begin{aligned}
|e_i + y_i| &= \left| \sum_{\beta=1}^{M} \sum_{u=1}^{W_1} \sum_{v=1}^{H_1} \varepsilon \boldsymbol{a}_{u,v,\beta} \cdot \sigma\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right) \right| \leq \varepsilon \sum_{\beta=1}^{M} \left| \sum_{u=1}^{W_1} \sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right) \right| \\
&\leq \varepsilon^2 \sum_{\beta=1}^{M} \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_1 \max_{u \in [W_1], v \in [H_1]} \left| \boldsymbol{x}_{u,v,\beta}^{[1]}(i) \right| \\
&\leq \varepsilon^2 \sum_{\beta=1}^{M} \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_1 \max_{u \in [W_1], v \in [H_1]} \left| \left(\sum_{p=0}^{m-1} \sum_{q=0}^{m-1} \boldsymbol{x}_{u+p,v+q}(i) \cdot \boldsymbol{W}_{p,q,\beta} \right) + \boldsymbol{b}_\beta \right| \\
&\leq \varepsilon^2 c \sum_{\beta=1}^{M} \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_1 \left\| \boldsymbol{\theta}_{\boldsymbol{W},\beta} \right\|_1 \\
&\leq \varepsilon^2 c \sqrt{W_1 H_1} \sqrt{m^2+1} \sum_{\beta=1}^{M} \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_2 \left\| \boldsymbol{\theta}_{\boldsymbol{W},\beta} \right\|_2,
\end{aligned}
$$

for simplicity we omit the constant $\sqrt{W_1 H_1}\sqrt{m^2+1}$ since it is a universal constant. Hence, we obtain further that

$$
\begin{aligned}
\left\| \boldsymbol{f}_\beta \right\|_2 &\leq c\sqrt{m^2+1} \left| \frac{1}{n} \sum_{i=1}^{n} (e_i + y_i) \cdot \left(\sum_{u=1}^{W_1} \sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \sigma^{(1)}\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right) \right) \right| \\
&\quad + c\sqrt{m^2+1} \left| \frac{1}{n} \sum_{i=1}^{n} y_i \cdot \left(\sum_{u=1}^{W_1} \sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta} \cdot \left(\sigma^{(1)}\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right) - 1 \right) \right) \right| \\
&\leq \frac{c\sqrt{m^2+1}\sqrt{W_1 H_1}}{n} \sum_{i=1}^{n} \left(\left(\sum_{\beta=1}^{M} \varepsilon^2 \left\| \boldsymbol{\theta}_{\boldsymbol{W},\beta} \right\|_2 \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_2 \right) \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_2 \right) \\
&\quad + c^2 \sqrt{m^2+1}\sqrt{W_1 H_1} \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_2 \left\| \varepsilon \boldsymbol{\theta}_{\boldsymbol{W},\beta} \right\|_2 \\
&\leq \left(M\varepsilon^2 E_{\max}^2 + \varepsilon E_{\max} \right) \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_2,
\end{aligned}
$$

where we also omit the constant $\sqrt{W_1 H_1}\sqrt{m^2+1}$, and similarly

$$
\begin{aligned}
\left\| \boldsymbol{g}_\beta \right\|_2 &\leq c\sqrt{W_1 H_1} \max_{u \in [W_1], v \in [H_1]} \left| \frac{1}{n} \sum_{i=1}^{n} (e_i + y_i) \cdot \frac{\sigma\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right)}{\varepsilon} \right| \\
&\quad + c\sqrt{W_1 H_1} \max_{u \in [W_1], v \in [H_1]} \left| \frac{1}{n} \sum_{i=1}^{n} y_i \cdot \left(\frac{\sigma\left(\varepsilon \boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right)}{\varepsilon} - \boldsymbol{x}_{u,v,\beta}^{[1]}(i) \right) \right| \\
&\leq \frac{c\sqrt{m^2+1}\sqrt{W_1 H_1}}{n} \sum_{i=1}^{n} \left(\left(\sum_{\beta=1}^{M} \varepsilon^2 \left\| \boldsymbol{\theta}_{\boldsymbol{W},\beta} \right\|_2 \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta} \right\|_2 \right) \left\| \boldsymbol{\theta}_{\boldsymbol{W},\beta} \right\|_2 \right) \\
&\quad + c^2 \sqrt{m^2+1}\sqrt{W_1 H_1} \varepsilon \left\| \boldsymbol{\theta}_{\boldsymbol{W},\beta} \right\|_2^2 \\
&\leq \left(M\varepsilon^2 E_{\max}^2 + \varepsilon E_{\max} \right) \left\| \boldsymbol{\theta}_{\boldsymbol{W},\beta} \right\|_2.
\end{aligned}
$$

□

C.4 Several Estimate on the Initial Parameters

We begin this part by an estimate on standard Gaussian vectors

Lemma 1 (Bounds on initial parameters)**.** *Given any $\delta \in (0,1)$, we have with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,*

$$
\max_{\beta \in [M]} \left\{ \left\| \boldsymbol{\theta}_{\boldsymbol{W},\beta}^0 \right\|_\infty, \left\| \boldsymbol{\theta}_{\boldsymbol{a},\beta}^0 \right\|_\infty \right\} \leq \sqrt{2 \log \frac{2M(m^2+1+W_1 H_1)}{\delta}}. \tag{49}
$$

Proof. If $\mathrm{X} \sim \mathcal{N}(0,1)$, then for any $\eta > 0$,

$$\mathbb{P}(|\mathrm{X}| > \eta) \leq 2\exp\left(-\frac{1}{2}\eta^2\right).$$

Since given any $\beta \in [M]$, for each $u \in [W_1]$, $v \in [H_1]$, $p \in [0:m-1]$ and $q \in [0:m-1]$,

$$\boldsymbol{W}^0_{p,q,\beta} \sim \mathcal{N}(0,1), \quad \boldsymbol{b}^0_\beta \sim \mathcal{N}(0,1), \quad \boldsymbol{a}^0_{u,v,\beta} \sim \mathcal{N}(0,1),$$

and they are all independent with each other. As we set

$$\eta = \sqrt{2\log\frac{2M(m^2+1+W_1H_1)}{\delta}},$$

we obtain that

$$\begin{aligned}
&\mathbb{P}\left(\max_{\beta\in[M]}\left\{\left\|\boldsymbol{\theta}^0_{\boldsymbol{W},\beta}\right\|_\infty, \left\|\boldsymbol{\theta}^0_{\boldsymbol{a},\beta}\right\|_\infty\right\} > \eta\right) \\
&= \mathbb{P}\left(\max_{\beta\in[M],u\in[W_1],v\in[H_1],p\in[0:m-1],q\in[0:m-1]}\left\{\left|\boldsymbol{W}^0_{p,q,\beta}\right|, \left|\boldsymbol{b}^0_\beta\right|, \left|\boldsymbol{a}^0_{u,v,\beta}\right|\right\} > \eta\right) \\
&\leq 2M(m^2+1)\exp\left(-\frac{1}{2}\eta^2\right) + 2MW_1H_1\exp\left(-\frac{1}{2}\eta^2\right) \\
&= 2M(m^2+1+W_1H_1)\exp\left(-\frac{1}{2}\eta^2\right) = \delta.
\end{aligned}$$

□

Next we would like to introduce the sub-exponential norm Vershynin (2010) of a random variable and Bernstein's Inequality.

Definition 4 (Sub-exponential norm)**.** *The sub-exponential norm of a random variable* X *is defined as*

$$\|\mathrm{X}\|_{\psi_1} := \inf\left\{s > 0 \mid \mathbb{E}_{\mathrm{X}}\left[\exp\left(\frac{|\mathrm{X}|}{s}\right)\right] \leq 2\right\}. \tag{50}$$

In particular, we denote $\mathrm{X} := \chi^2(d)$ *as a chi-square distribution with* d *degrees of freedom Laurent and Massart (2000), and its sub-exponential norm by*

$$C_{\psi,d} := \|\mathrm{X}\|_{\psi_1}.$$

Remark 5. *As the probability density function of* $\mathrm{X} = \chi^2(d)$ *reads*

$$f_{\mathrm{X}}(z) := \frac{1}{2^{\frac{d}{2}}\Gamma(\frac{d}{2})} z^{\frac{d}{2}-1}\exp\left(-\frac{z}{2}\right), \quad z \geq 0,$$

we note that

$$\mathbb{E}_{\mathrm{X}\sim\chi^2(1)}\exp\left(\frac{|\mathrm{X}|}{s}\right) = \int_0^{+\infty}\frac{1}{2^{\frac{1}{2}}\Gamma(\frac{1}{2})}z^{-\frac{1}{2}}\exp\left(-\left(\frac{1}{2}-\frac{1}{s}\right)z\right)\mathrm{d}z = \frac{1}{\sqrt{1-\frac{2}{s}}},$$

Then we obtain that

$$\frac{8}{3} \leq C_{\psi,1} < 3.$$

Moreover, we notice that

$$\mathbb{E}_{\mathrm{X}\sim\chi^2(d)}\exp\left(\frac{|\mathrm{X}|}{s}\right) = \left(\mathbb{E}_{\mathrm{Y}\sim\chi^2(1)}\exp\left(\frac{|\mathrm{Y}|}{s}\right)\right)^d,$$

as we set

$$\frac{1}{\sqrt{1-\frac{2}{s}}} = 2^{\frac{1}{d}},$$

then

$$s = \frac{2}{1 - 2^{-\frac{2}{d}}},$$

hence

$$\frac{2}{1 - 2^{-\frac{2}{d}}} \leq C_{\psi,d} < 3,$$

and

$$C_{\psi,d} \geq C_{\psi,1},$$

for $d \geq 1$.

Theorem 2 (Bernstein's inequality)**.** *Let* $\{\mathrm{X}_k\}_{k=1}^m$ *be i.i.d. sub-exponential random variables satisfying*

$$\mathbb{E}\mathrm{X}_1 = \mu,$$

then for any $\eta \geq 0$*, we have*

$$\mathbb{P}\left(\left|\frac{1}{m}\sum_{k=1}^{m}\mathrm{X}_k - \mu\right| \geq \eta\right) \leq 2\exp\left(-C_0 m \min\left(\frac{\eta^2}{\|\mathrm{X}_1\|_{\psi_1}^2}, \frac{\eta}{\|\mathrm{X}_1\|_{\psi_1}}\right)\right),$$

for some absolute constant C_0.

In order to study the phenomenon of condensation, we need to concatenate the vectors $\{\boldsymbol{\theta}_{\boldsymbol{W},\beta}\}_{\beta=1}^M$ into

$$\boldsymbol{\theta}_{\boldsymbol{W}} := \mathrm{vec}\left(\{\boldsymbol{\theta}_{\boldsymbol{W},\beta}\}_{\beta=1}^M\right),$$

and we obtain that

Proposition 3 (Upper and lower bounds of initial parameters)**.** *Given any* $\delta \in (0,1)$*, if*

$$M = \Omega\left(\log\frac{2}{\delta}\right),$$

then with probability at least $1-\delta$ *over the choice of* $\boldsymbol{\theta}^0$*,*

$$\sqrt{\frac{M(m^2+1)}{2}} \leq \left\|\boldsymbol{\theta}_{\boldsymbol{W}}^0\right\|_2 \leq \sqrt{\frac{3M(m^2+1)}{2}}. \tag{51}$$

Proof. Since given any $\beta \in [M]$, for each $p \in [0:m-1]$ and $q \in [0:m-1]$,

$$\left(\boldsymbol{W}_{p,q,\beta}^0\right)^2, \quad \left(\boldsymbol{b}_\beta^0\right)^2 \sim \chi^2(1)$$

are sub-exponential random variables with

$$\mathbb{E}\left(\boldsymbol{W}_{p,q,\beta}^0\right)^2 = 1, \quad \mathbb{E}\left(\boldsymbol{b}_\beta^0\right)^2 = 1.$$

Since $C_{\psi,1} \geq \frac{8}{3} > 2$, then for any $0 \leq \eta \leq 2$, it is obvious that

$$\min\left(\frac{\eta^2}{C_{\psi,1}^2}, \frac{\eta}{C_{\psi,1}}\right) = \frac{\eta^2}{C_{\psi,1}^2}.$$

Hence, by application of Theorem 2,

$$\begin{aligned}
&\mathbb{P}\left(\left|\frac{1}{M(m^2+1)}\sum_{\beta=1}^{M}\left[\left(\sum_{p=0}^{m-1}\sum_{q=0}^{m-1}\left(\boldsymbol{W}_{p,q,\beta}^0\right)^2\right) + \left(\boldsymbol{b}_\beta\right)^2\right] - 1\right| \geq \eta\right) \\
&\leq 2\exp\left(-\frac{C_0 M(m^2+1)\eta^2}{C_{\psi,1}^2}\right),
\end{aligned}$$

as we set

$$2\exp\left(-\frac{C_0 M(m^2+1)\eta^2}{C_{\psi,1}^2}\right) = \delta,$$

and consequently,

$$\eta = \sqrt{\frac{C_{\psi,1}^2}{C_0 M(m^2+1)} \log \frac{2}{\delta}},$$

then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,

$$\left\|\boldsymbol{\theta}_{\boldsymbol{W}}^0\right\|_2^2 \geq M(m^2+1)\left(1 - \sqrt{\frac{C_{\psi,1}^2}{C_0 M(m^2+1)} \log \frac{2}{\delta}}\right),$$

$$\left\|\boldsymbol{\theta}_{\boldsymbol{W}}^0\right\|_2^2 \leq M(m^2+1)\left(1 + \sqrt{\frac{C_{\psi,1}^2}{C_0 M(m^2+1)} \log \frac{2}{\delta}}\right),$$

and by choosing

$$M \geq \frac{4C_{\psi,1}^2}{C_0(m^2+1)} \log \frac{2}{\delta},$$

we obtain that

$$\sqrt{\frac{M(m^2+1)}{2}} \leq \left\|\boldsymbol{\theta}_{\boldsymbol{W}}^0\right\|_2 \leq \sqrt{\frac{3M(m^2+1)}{2}}.$$

□

C.5 Lower Bound on Effective Time

We denote a useful quantity

$$\phi(t) := \sup_{0 \leq s \leq t} E_{\max}(s), \tag{52}$$

then directly from Lemma 1, we have with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$,

$$\max_{\beta \in [M]} \left\{\left\|\boldsymbol{\theta}_{\boldsymbol{W},\beta}^0\right\|_\infty, \left\|\boldsymbol{\theta}_{\boldsymbol{a},\beta}^0\right\|_\infty\right\} \leq \sqrt{2 \log \frac{2M(m^2+1+W_1H_1)}{\delta}},$$

hence

$$\phi(0) \leq \sqrt{2(m^2+1+W_1H_1) \log \frac{2M(m^2+1+W_1H_1)}{\delta}}. \tag{53}$$

We define

$$T_{\text{eff}} := \inf\left\{t > 0 \mid M\varepsilon^2\phi^3(t) > M^{-\tau}, \quad \tau = \frac{\gamma-1}{4}\right\}, \tag{54}$$

then for M large enough,

$$M\varepsilon^2\phi^3(0) \leq M\varepsilon^2\left(2(m^2+1+W_1H_1)\log\frac{2M(m^2+1+W_1H_1)}{\delta}\right)^{\frac{3}{2}} \leq M^{-\frac{\gamma-1}{2}},$$

hence $T_{\text{eff}} \geq 0$. We observe further that as the real dynamics read

$$\frac{\mathrm{d}\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta} \\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix}}{\mathrm{d}t} = \begin{pmatrix}\boldsymbol{f}_\beta \\ \boldsymbol{g}_\beta\end{pmatrix} + \boldsymbol{A}\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta} \\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix}. \tag{55}$$

then by taking the 2-norm on both sides

$$\begin{aligned} E_\beta(t) \leq{}& \exp\left(t\left\|\boldsymbol{A}\right\|_{2\to 2}\right) E_\beta(0) \\ &+ \int_0^t \exp\left((t-s)\left\|\boldsymbol{A}\right\|_{2\to 2}\right)\left(M\varepsilon^2 E_{\max}^2(s) + \varepsilon E_{\max}(s)\right) E_\beta(s)\mathrm{d}s, \end{aligned}$$

by taking supreme over the index β and time $0 \leq t \leq T_{\text{eff}}$ on both sides, for M large enough,

$$\begin{aligned} \phi(t) &\leq \phi(0)\exp(\lambda_1 t) + 2M^{-\min\{1,\tau\}} \int_0^t \exp(\lambda_1(t-s))\mathrm{d}s \\ &\leq \phi(0)\exp(\lambda_1 t) + 2M^{-\min\{1,\tau\}} \frac{\exp(\lambda_1 t)-1}{\lambda_1} \\ &\leq \phi(0)\exp(\lambda_1 t) + 2M^{-\min\{1,\tau\}} \frac{\exp(\lambda_1 t)}{\lambda_1}, \end{aligned} \tag{56}$$

then based on Lemma 1, with probability $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, for sufficiently large M,

$$\begin{aligned}\phi(t) &\leq \phi(0)\exp(\lambda_1 t) + \frac{2}{\lambda_1} M^{-\min\{1,\tau\}}\exp(\lambda_1 t)\\ &\leq 2\phi(0)\exp(\lambda_1 t)\\ &\leq 2\sqrt{2(m^2+1+W_1H_1)\log\frac{2M(m^2+1+W_1H_1)}{\delta}}\exp(\lambda_1 t),\end{aligned} \tag{57}$$

we set t_0 as the time satisfying

$$2\sqrt{2(m^2+1+W_1H_1)\log\frac{2M(m^2+1+W_1H_1)}{\delta}}\exp(\lambda_1 t) = \frac{1}{2}M^{\frac{\gamma-1}{4}}, \tag{58}$$

then we obtain that, for any $\eta_0 > \frac{\gamma-1}{100} > 0$,

$$T_{\text{eff}} \geq t_0 > \frac{1}{\lambda_1}\left[\log\left(\frac{1}{4}\right) + \left(\frac{\gamma-1}{4} - \eta_0\right)\log(M)\right]. \tag{59}$$

Recall that

$$\boldsymbol{\theta}_{\boldsymbol{W}} = \operatorname{vec}\left(\{\boldsymbol{\theta}_{\boldsymbol{W},\beta}\}_{\beta=1}^{M}\right),$$

and we denote further that

$$\boldsymbol{\theta}_{\boldsymbol{W},\boldsymbol{v}_1} := \mathcal{P}_1\boldsymbol{\theta}_{\boldsymbol{W}} := \left(\langle\boldsymbol{\theta}_{\boldsymbol{W},1},\boldsymbol{v}_1\rangle, \langle\boldsymbol{\theta}_{\boldsymbol{W},2},\boldsymbol{v}_1\rangle, \cdots \langle\boldsymbol{\theta}_{\boldsymbol{W},M},\boldsymbol{v}_1\rangle\right)^{\intercal},$$

where $\boldsymbol{v}_1$ is the eigenvector of the largest eigenvalue of $\boldsymbol{Z}^{\intercal}\boldsymbol{Z}$, or the first column vector of $\boldsymbol{V}$ in (36).

Theorem 3. *Given any $\delta \in (0,1)$, under Assumption 5, Assumption 6, Assumption 7 and Assumption 8, if $\gamma > 1$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lim_{M\to+\infty}\sup_{t\in[0,T_{\text{eff}}]}\frac{\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)-\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} = +\infty, \tag{60}$$

and

$$\lim_{M\to+\infty}\sup_{t\in[0,T_{\text{eff}}]}\frac{\|\boldsymbol{\theta}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)\|_2} = 1. \tag{61}$$

Proof. Since for each $\beta \in [M]$, the real dynamics read

$$\frac{\mathrm{d}\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta}\\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix}}{\mathrm{d}t} = \boldsymbol{A}\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta}\\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix} + \begin{pmatrix}\boldsymbol{f}_\beta\\ \boldsymbol{g}_\beta\end{pmatrix}, \quad \begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta}(0)\\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}(0)\end{pmatrix} = \begin{pmatrix}\boldsymbol{\theta}^0_{\boldsymbol{W},\beta}\\ \boldsymbol{\theta}^0_{\boldsymbol{a},\beta}\end{pmatrix},$$

and

$$\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta}\\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix} = \exp(t\boldsymbol{A})\begin{pmatrix}\boldsymbol{\theta}^0_{\boldsymbol{W},\beta}\\ \boldsymbol{\theta}^0_{\boldsymbol{a},\beta}\end{pmatrix} + \int_0^t \exp((t-s)\boldsymbol{A})\begin{pmatrix}\boldsymbol{f}_\beta\\ \boldsymbol{g}_\beta\end{pmatrix}\mathrm{d}s.$$

As we notice that for any $\beta \in [M]$, $\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta}\\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix}$ can be written into two parts, the first one is the linear part, the second one is the residual part. For simplicity of proof, we need to introduce some further notations, i.e., as we denote

$$\begin{aligned}\begin{pmatrix}\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\beta}\\ \bar{\boldsymbol{\theta}}_{\boldsymbol{a},\beta}\end{pmatrix} &:= \exp(t\boldsymbol{A})\begin{pmatrix}\boldsymbol{\theta}^0_{\boldsymbol{W},\beta}\\ \boldsymbol{\theta}^0_{\boldsymbol{a},\beta}\end{pmatrix},\\ \begin{pmatrix}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\beta}\\ \widetilde{\boldsymbol{\theta}}_{\boldsymbol{a},\beta}\end{pmatrix} &:= \int_0^t \exp((t-s)\boldsymbol{A})\begin{pmatrix}\boldsymbol{f}_\beta\\ \boldsymbol{g}_\beta\end{pmatrix}\mathrm{d}s,\end{aligned}$$

then

$$\begin{pmatrix}\boldsymbol{\theta}_{\boldsymbol{W},\beta}\\ \boldsymbol{\theta}_{\boldsymbol{a},\beta}\end{pmatrix} = \begin{pmatrix}\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\beta}\\ \bar{\boldsymbol{\theta}}_{\boldsymbol{a},\beta}\end{pmatrix} + \begin{pmatrix}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\beta}\\ \widetilde{\boldsymbol{\theta}}_{\boldsymbol{a},\beta}\end{pmatrix},$$

directly from Proposition 1, we obtain that

$$\begin{aligned}\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\beta}(t) &= \left(\sum_{k=1}^{r}\left[c_{\lambda_k,\boldsymbol{W},\beta}\exp(\lambda_k t)+d_{\lambda_k,\boldsymbol{W},\beta}\exp(-\lambda_k t)\right]\boldsymbol{v}_k\right)\\ &\quad+\mathcal{P}_{(r+1):(m^2+1)}\boldsymbol{\theta}^0_{\boldsymbol{W},\beta},\\ \bar{\boldsymbol{\theta}}_{\boldsymbol{a},\beta}(t) &= \left(\sum_{k=1}^{r}\left[c_{\lambda_k,\boldsymbol{a},\beta}\exp(\lambda_k t)+d_{\lambda_k,\boldsymbol{a},\beta}\exp(-\lambda_k t)\right]\boldsymbol{u}_k\right)\\ &\quad+\mathcal{P}_{(r+1):(W_1H_1)}\boldsymbol{\theta}^0_{\boldsymbol{a},\beta}.\end{aligned}\tag{62}$$

We are hereby to prove (60). Firstly, we observe that

$$\boldsymbol{\theta}_{\boldsymbol{W}}(0)=\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0),$$

hence

$$\begin{aligned}&\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)-\boldsymbol{\theta}_{\boldsymbol{W}}(0)\right\|_2^2\\ =&\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right\|_2^2\\ =&\left\|\mathcal{P}_{1:r}\left(\boldsymbol{\theta}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)+\mathcal{P}_{(r+1):(m^2+1)}\left(\boldsymbol{\theta}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2^2\\ =&\left\|\mathcal{P}_{1:r}\left(\boldsymbol{\theta}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2^2+\left\|\mathcal{P}_{(r+1):(m^2+1)}\left(\boldsymbol{\theta}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2^2\\ =&\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)+\mathcal{P}_{1:r}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2^2+\left\|\mathcal{P}_{(r+1):(m^2+1)}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2^2,\end{aligned}$$

by choosing $\eta_0=\frac{\gamma-1}{8}$, then for time $0\le t\le \bar{t}_0:=\frac{1}{\lambda_1}\left[\left(\frac{\gamma-1}{8}\right)\log(M)-\log(2)\right]$ and any $\beta\in[M]$,

$$\begin{aligned}\left\|\begin{pmatrix}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\beta}\\ \widetilde{\boldsymbol{\theta}}_{\boldsymbol{a},\beta}\end{pmatrix}\right\|_2&=\left\|\int_0^t\exp\left((t-s)\boldsymbol{A}\right)\begin{pmatrix}\boldsymbol{f}_k\\ \boldsymbol{g}_k\end{pmatrix}\mathrm{d}s\right\|_2\\ &\le\left(M\varepsilon^2\phi^3(t)+\varepsilon\phi^2(t)\right)\int_0^t\exp(\lambda_1(t-s))\mathrm{d}s\\ &\le 2M^{-\min\left\{\tau,\frac{1}{2}\right\}}\int_0^t\exp(\lambda_1(t-s))\mathrm{d}s\\ &\le 2M^{-\frac{\gamma-1}{4}}\frac{\exp(\lambda_1 t)}{\lambda_1}\le 2M^{-\frac{\gamma-1}{4}}\exp(\lambda_1\bar{t}_0)=M^{-\frac{\gamma-1}{8}}.\end{aligned}$$

We conclude that for $t\le\bar{t}_0$, the following holds

$$\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2\le\sqrt{M}\left\|\begin{pmatrix}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\beta}\\ \widetilde{\boldsymbol{\theta}}_{\boldsymbol{a},\beta}\end{pmatrix}\right\|_2\le\sqrt{M}M^{-\frac{\gamma-1}{8}},$$

thus the ratio reads

$$\begin{aligned}&\left(\frac{\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)-\boldsymbol{\theta}_{\boldsymbol{W}}(0)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\right\|_2}\right)^2\\ =&\frac{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)+\mathcal{P}_{1:r}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2^2+\left\|\mathcal{P}_{(r+1):(m^2+1)}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2^2}{\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\right\|_2^2}\\ =&\underbrace{\frac{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)+\mathcal{P}_{1:r}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2^2}{\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\right\|_2^2}}_{\mathrm{I}}+\underbrace{\frac{\left\|\mathcal{P}_{(r+1):(m^2+1)}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2^2}{\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\right\|_2^2}}_{\mathrm{II}}.\end{aligned}$$

For part II, we obtain that

$$\frac{\left\|\mathcal{P}_{(r+1):(m^2+1)}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\right\|_2}\le\frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\left\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\right\|_2},$$

then directly from Proposition 3, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough M, for any $0 \leq t \leq \bar{t}_0 = \frac{1}{\lambda_1}\left[\left(\frac{\gamma-1}{8}\right)\log(M) - \log(2)\right]$, the following holds:

$$\frac{\left\|\mathcal{P}_{(r+1):(m^2+1)}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} \leq \frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} \leq \sqrt{\frac{2}{m^2+1}} M^{-\frac{\gamma-1}{8}},$$

by taking the limit, we obtain that

$$\lim_{M\to\infty} \sup_{t\in[0,\bar{t}_0]} \frac{\left\|\mathcal{P}_{(r+1):(m^2+1)}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} = 0.$$

As for part I, we notice that

$$\begin{aligned}
&\frac{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right) + \mathcal{P}_{1:r}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} \\
&\geq \frac{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2 - \left\|\mathcal{P}_{1:r}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} \\
&\geq \underbrace{\frac{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2}}_{\text{III}} - \underbrace{\frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2}}_{\text{IV}},
\end{aligned}$$

by similar reasoning as shown in part II, for any time $t \in [0, \bar{t}_0]$, part IV tends to zero as $M \to \infty$, i.e.,

$$\lim_{M\to\infty} \sup_{t\in[0,\bar{t}_0]} \frac{\left\|\mathcal{P}_{1:r}\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} \leq \lim_{M\to\infty} \sup_{t\in[0,\bar{t}_0]} \frac{\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} = 0.$$

For part III, we observe that since

$$\begin{aligned}
&\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2^2 \\
&= \sum_{\beta=1}^{M}\sum_{k=1}^{r}\left[c_{\lambda_k,\boldsymbol{W},\beta}\left(\exp(\lambda_k t) - 1\right) + d_{\lambda_k,\boldsymbol{W},\beta}\left(\exp(-\lambda_k t) - 1\right)\right]^2,
\end{aligned}$$

where

$$\begin{aligned}
c_{\lambda_k,\boldsymbol{W},\beta} &= \frac{1}{2}\left(\left\langle \boldsymbol{\theta}^0_{\boldsymbol{W},\beta}, \boldsymbol{v}_k\right\rangle + \left\langle \boldsymbol{\theta}^0_{\boldsymbol{a},\beta}, \boldsymbol{u}_k\right\rangle\right), \\
d_{\lambda_k,\boldsymbol{W},\beta} &= \frac{1}{2}\left(\left\langle \boldsymbol{\theta}^0_{\boldsymbol{W},\beta}, \boldsymbol{v}_k\right\rangle - \left\langle \boldsymbol{\theta}^0_{\boldsymbol{a},\beta}, \boldsymbol{u}_k\right\rangle\right),
\end{aligned}$$

we observe that given $\boldsymbol{u}_k$ and $\boldsymbol{v}_k$, $\mathrm{Y}_{k,\beta} := \left\langle \boldsymbol{\theta}^0_{\boldsymbol{a},\beta}, \boldsymbol{u}_k\right\rangle \sim \mathcal{N}(0,1)$ and $\mathrm{X}_{k,\beta} : \left\langle \boldsymbol{\theta}^0_{\boldsymbol{W},\beta}, \boldsymbol{v}_k\right\rangle \sim \mathcal{N}(0,1)$. Moreover, $\{\mathrm{X}_{k,\beta}\}_{\beta=1}^M \sim \mathcal{N}(0,1)$ and $\{\mathrm{Y}_{k,\beta}\}_{\beta=1}^M \sim \mathcal{N}(0,1)$ are i.i.d. Gaussian variables, and they are independent with each other. We denote further that $r_k(t) := \exp\left(\frac{1}{2}\lambda_k t\right)$, and by application of Theorem 2, with probability $1-\frac{\delta}{4}$ over the choice of $\boldsymbol{\theta}^0$, for M large enough,

$$\frac{(m^2+1)}{2} \leq \frac{1}{M}\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2^2 \leq \frac{3(m^2+1)}{2},$$

and with probability $1-\frac{\delta}{4}$ over the choice of $\boldsymbol{\theta}^0$, for M large enough,

$$\frac{1}{2} \leq \frac{1}{M}\sum_{\beta=1}^{M}\mathrm{X}_{k,\beta}^2 \leq \frac{3}{2},$$

and with probability $1-\frac{\delta}{4}$ over the choice of $\boldsymbol{\theta}^0$, for M large enough,

$$\frac{1}{2} \leq \frac{1}{M}\sum_{\beta=1}^{M}\mathrm{Y}_{k,\beta}^2 \leq \frac{3}{2},$$

and with probability $1-\frac{\delta}{4}$ over the choice of $\boldsymbol{\theta}^0$, for M large enough,

$$-\frac{1}{4} \leq \frac{1}{M} \sum_{\beta=1}^{M} \mathrm{X}_{k,\beta} \mathrm{Y}_{k,\beta} \leq \frac{1}{4},$$

then we obtain that, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$

$$\begin{aligned}
&\frac{1}{M} \left\| \mathcal{P}_{1:r} \left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0) \right) \right\|_2^2 \\
=&\frac{1}{M} \sum_{\beta=1}^{M} \sum_{k=1}^{r} \left[c_{\lambda_k, \boldsymbol{W}, \beta} \left(\exp(\lambda_k t) - 1 \right) + d_{\lambda_k, \boldsymbol{W}, \beta} \left(\exp(-\lambda_k t) - 1 \right) \right]^2 \\
=&\frac{1}{4M} \sum_{\beta=1}^{M} \sum_{k=1}^{r} \left[\mathrm{X}_{k,\beta} \left(r_k^2(t) + r_k^{-2}(t) - 2 \right) + \mathrm{Y}_{k,\beta} \left(r_k^2(t) - r_k^{-2}(t) \right) \right]^2 \\
=&\frac{1}{4M} \sum_{\beta=1}^{M} \sum_{k=1}^{r} \left(r_k(t) - r_k^{-1}(t) \right)^2 \left[\mathrm{X}_{k,\beta} \left(r_k(t) - r_k^{-1}(t) \right) + \mathrm{Y}_{k,\beta} \left(r_k(t) + r_k^{-1}(t) \right) \right]^2 \\
\geq&\frac{1}{8} \sum_{k=1}^{r} \left(r_k(t) - r_k^{-1}(t) \right)^2 \left(r_k^2(t) + 3 r_k^{-2}(t) \right) \geq \frac{1}{8} \sum_{k=1}^{r} \left(r_k(t) - r_k^{-1}(t) \right)^4.
\end{aligned}$$

Hence, with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough M, for any $0 \leq t \leq \bar{t}_0 = \frac{1}{\lambda_1} \left[\left(\frac{\gamma-1}{8} \right) \log(M) - \log(2) \right]$,

$$\begin{aligned}
&\frac{1}{M} \left\| \mathcal{P}_{1:r} \left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0) \right) \right\|_2^2 \\
\geq&\frac{1}{8} \sum_{k=1}^{r} \left(r_k(\bar{t}_0) - r_k^{-1}(\bar{t}_0) \right)^4 \geq \frac{1}{8} \sum_{k=1}^{r} \left(r_k(\bar{t}_0) - 1 \right)^4 \\
\gtrsim&\frac{1}{8} \sum_{k=1}^{r} \exp\left(\frac{4}{2} \frac{\lambda_k}{\lambda_1} \left(\frac{\gamma-1}{8} \right) \log(M) \right) = \frac{1}{8} \sum_{k=1}^{r} M^{\frac{\lambda_k}{\lambda_1} \frac{\gamma-1}{4}},
\end{aligned}$$

then for part III, we obtain that with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough M, for any $0 \leq t \leq \bar{t}_0 = \frac{1}{\lambda_1} \left[\left(\frac{\gamma-1}{8} \right) \log(M) - \log(2) \right]$,

$$\begin{aligned}
\frac{\left\| \mathcal{P}_{1:r} \left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(\bar{t}_0) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0) \right) \right\|_2^2}{\left\| \boldsymbol{\theta}_{\boldsymbol{W}}(0) \right\|_2^2} &= \frac{\frac{1}{M} \left\| \mathcal{P}_{1:r} \left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(\bar{t}_0) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0) \right) \right\|_2^2}{\frac{1}{M} \left\| \boldsymbol{\theta}_{\boldsymbol{W}}(0) \right\|_2^2} \\
&\geq \frac{2}{3(m^2+1)} \frac{1}{M} \left\| \mathcal{P}_{1:r} \left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0) \right) \right\|_2^2 \\
&\gtrsim \frac{2}{3(m^2+1)} \frac{1}{8} \sum_{k=1}^{r} M^{\frac{\lambda_k}{\lambda_1} \frac{\gamma-1}{4}},
\end{aligned}$$

by taking the limit, we obtain that

$$\lim_{M \to \infty} \sup_{t \in [0, \bar{t}_0]} \frac{\left\| \mathcal{P}_{1:r} \left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(\bar{t}_0) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0) \right) \right\|_2}{\left\| \boldsymbol{\theta}_{\boldsymbol{W}}(0) \right\|_2} = \infty.$$

To sum up, since $\bar{t}_0 \leq T_{\text{eff}}$, we have that

$$\lim_{m \to +\infty} \sup_{t \in [0, T_{\text{eff}}]} \frac{\left\| \boldsymbol{\theta}_{\boldsymbol{W}}(t) - \boldsymbol{\theta}_{\boldsymbol{W}}(0) \right\|_2}{\left\| \boldsymbol{\theta}_{\boldsymbol{W}}(0) \right\|_2} = +\infty + 0 = +\infty, \tag{63}$$

which finishes the proof of (60).

In order to prove (61), firstly we have

$$\frac{\left\| \boldsymbol{\theta}_{\boldsymbol{W}, \boldsymbol{v}_1}(t) \right\|_2}{\left\| \boldsymbol{\theta}_{\boldsymbol{W}}(t) \right\|_2} \leq 1,$$

moreover, we observe that

$$\left(\frac{\|\boldsymbol{\theta}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)\|_2}\right)^2 = \frac{\|\boldsymbol{\theta}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\|_2^2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)\|_2^2} = \frac{\|\boldsymbol{\theta}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\|_2^2}{\|\mathcal{P}_{1:r}\boldsymbol{\theta}_{\boldsymbol{W}}(t)\|_2^2 + \left\|\mathcal{P}_{(r+1):(m^2+1)}\boldsymbol{\theta}_{\boldsymbol{W}}(t)\right\|_2^2}$$
$$= \frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\right\|_2^2}{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right)\right\|_2^2 + \left\|\mathcal{P}_{(r+1):(m^2+1)}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right)\right\|_2^2},$$

where

$$\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t) := \mathcal{P}_1\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t),$$
$$\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t) := \mathcal{P}_1\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t).$$

Then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$ and large enough M, for any $0 \le t \le \bar{t}_0 = \frac{1}{\lambda_1}\left[\left(\frac{\gamma-1}{8}\right)\log(M) - \log(2)\right]$, the following holds:

$$\begin{aligned}
\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\right\|_2 &\ge \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\right\|_2 - \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2 \\
&\ge \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2 - \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\right\|_2 - \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2 \\
&\ge \underbrace{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2}_{\text{V}} - \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2 - \left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2 \\
&\ge \sqrt{\frac{M}{8}}\left(r_1(t) - r_1^{-1}(t)\right)^2 - \sqrt{M}M^{-\frac{\gamma-1}{8}} - \sqrt{\frac{3M}{2}},
\end{aligned}$$

hence part V is the term of dominance, and by similar reasoning

$$\begin{aligned}
\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right)\right\|_2 &\ge \underbrace{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2}_{\text{VI}} - \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2 - \left\|\mathcal{P}_{1:r}\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right\|_2 \\
&\ge \sqrt{\frac{M}{8}\sum_{k=1}^{r}\left(r_k(t) - r_k^{-1}(t)\right)^4} - \sqrt{M}M^{-\frac{\gamma-1}{8}} - \sqrt{\frac{3Mr}{2}},
\end{aligned}$$

hence part VI is the term of dominance, and finally

$$\begin{aligned}
&\left\|\mathcal{P}_{(r+1):(m^2+1)}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) + \widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right)\right\|_2 \\
\le &\left\|\mathcal{P}_{(r+1):(m^2+1)}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2 + \left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2 + \left\|\mathcal{P}_{(r+1):(m^2+1)}\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right\|_2 \\
= &\left\|\widetilde{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)\right\|_2 + \left\|\mathcal{P}_{(r+1):(m^2+1)}\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right\|_2 \le \sqrt{M}M^{-\frac{\gamma-1}{8}} + \sqrt{\frac{3M(m^2+1)}{2}},
\end{aligned}$$

which is at most of order $\sqrt{M}$. Then for M large enough, the majority part of the ratio $\frac{\left\|\boldsymbol{\theta}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\right\|_2^2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)\|_2^2}$ is

$$\frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2^2}{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t) - \bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2^2},$$

where

$$\begin{aligned}
&\frac{1}{M}\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2^2\\
=&\frac{1}{M}\sum_{\beta=1}^{M}\left[c_{\lambda_1,\boldsymbol{W},\beta}\left(\exp(\lambda_1 t)-1\right)+d_{\lambda_1,\boldsymbol{W},\beta}\left(\exp(-\lambda_1 t)-1\right)\right]^2\\
=&\frac{\exp(2\lambda_1 t)}{M}\sum_{\beta=1}^{M}\left[c_{\lambda_1,\boldsymbol{W},\beta}\left(1-\exp(-\lambda_1 t)\right)+d_{\lambda_1,\boldsymbol{W},\beta}\left(\exp(-2\lambda_1 t)-\exp(-\lambda_1 t)\right)\right]^2,
\end{aligned}$$

and

$$\begin{aligned}
&\frac{1}{M}\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(t)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2^2\\
=&\frac{1}{M}\sum_{\beta=1}^{M}\sum_{k=1}^{r}\left[c_{\lambda_k,\boldsymbol{W},\beta}\left(\exp(\lambda_k t)-1\right)+d_{\lambda_k,\boldsymbol{W},\beta}\left(\exp(-\lambda_k t)-1\right)\right]^2\\
=&\frac{\exp(2\lambda_1 t)}{M}\sum_{\beta=1}^{M}\sum_{k=1}^{r}\Big[c_{\lambda_k,\boldsymbol{W},\beta}\left(\exp((\lambda_1-\lambda_k)t)-\exp(-\lambda_1 t)\right)\\
&\qquad\qquad+d_{\lambda_1,\boldsymbol{W},\beta}\left(\exp(-(\lambda_1+\lambda_k)t)-\exp(-\lambda_1 t)\right)\Big]^2.
\end{aligned}$$

By taking $t=\bar{t}_0$, we observe that as the spectral gap $\Delta\lambda>0$, then for any $k\in[2:r]$,

$$\exp((\lambda_1-\lambda_k)\bar{t}_0)\leq\exp(-\Delta\lambda\bar{t}_0)\lesssim M^{-\left(\frac{\gamma-1}{8}\right)\frac{\Delta\lambda}{\lambda_1}}, \tag{64}$$

which tends to zero as $M\to\infty$, hence

$$\begin{aligned}
&\lim_{M\to\infty}\frac{\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(\bar{t}_0)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2^2}{\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(\bar{t}_0)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2^2}\\
=&\lim_{M\to\infty}\frac{\frac{1}{M}\left\|\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(\bar{t}_0)-\bar{\boldsymbol{\theta}}_{\boldsymbol{W},\boldsymbol{v}_1}(0)\right\|_2^2}{\frac{1}{M}\left\|\mathcal{P}_{1:r}\left(\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}\widetilde{\boldsymbol{\theta}}-\bar{\boldsymbol{\theta}}_{\boldsymbol{W}}(0)\right)\right\|_2^2}\\
=&\lim_{M\to\infty}\frac{\frac{\exp(2\lambda_1\bar{t}_0)}{M}\sum_{\beta=1}^{M}c_{\lambda_1,\boldsymbol{W},\beta}^2}{\frac{\exp(2\lambda_1\bar{t}_0)}{M}\sum_{\beta=1}^{M}c_{\lambda_1,\boldsymbol{W},\beta}^2+\underbrace{0+0+\cdots+0}_{r-1\text{ zeros}}}=1,
\end{aligned}$$

and in combination with $\bar{t}_0\leq T_{\text{eff}}$, we finish the proof of (61). □

D TWO-LAYER CNNS WITH MULTI CHANNELS

In the case of two-layer CNNs with multi channels, as $L=2$, then we still set $\boldsymbol{W}_{p,q,\alpha,\beta}:=\boldsymbol{W}_{p,q,\alpha,\beta}^{[1]}$, $\boldsymbol{b}_\beta:=\boldsymbol{b}_\beta^{[1]}$, and $\boldsymbol{x}_{u+p,v+q,\alpha}(i):=\boldsymbol{x}_{u+p,v+q,\alpha}^{[0]}(i)$ for simplicity, then the GD dynamics reads

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}_{p,q,\alpha,\beta}}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^{n}e_i\cdot\left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{a}_{u,v,\beta}\cdot\sigma^{(1)}\left(\boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right)\cdot\boldsymbol{x}_{u+p,v+q,\alpha}(i)\right),\\
\frac{\mathrm{d}\boldsymbol{b}_\beta}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^{n}e_i\cdot\left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1}\boldsymbol{a}_{u,v,\beta}\cdot\sigma^{(1)}\left(\boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right)\right),\\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t}&=-\frac{1}{n}\sum_{i=1}^{n}e_i\cdot\sigma\left(\boldsymbol{x}_{u,v,\beta}^{[1]}(i)\right).
\end{aligned}$$

since $e_i \approx -y_i$, and by means of perturbation expansion with respect to ε and keep the order 1 term, we obtain that

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}_{p,q,\alpha,\beta}}{\mathrm{d}t} &\approx \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \left(\sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta}\cdot \boldsymbol{x}_{u+p,v+q,\alpha}(i)\right),\\
\frac{\mathrm{d}\boldsymbol{b}_{\beta}}{\mathrm{d}t} &\approx \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta},\\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &\approx \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \boldsymbol{x}^{[1]}_{u,v,\beta}(i)\\
&= \frac{1}{n}\sum_{i=1}^{n} y_i \cdot \left[\left(\sum_{p=0}^{m-1}\sum_{q=0}^{m-1} \boldsymbol{x}_{u+p,v+q}(i)\cdot \boldsymbol{W}_{p,q,\alpha,\beta}\right)+\boldsymbol{b}_{\beta}\right].
\end{aligned} \tag{65}$$

Given any $u \in [W_1]$, $v \in [H_1]$ and $\alpha \in [C_0]$, then for all $p, q \in [0 : m-1]$, we set

$$\begin{aligned}
\boldsymbol{z}_{u+p,v+q,\alpha} &:= \frac{1}{n}\sum_{i=1}^{n} y_i \boldsymbol{x}_{u+p,v+q,\alpha}(i),\\
z &:= \frac{1}{n}\sum_{i=1}^{n} y_i,
\end{aligned} \tag{66}$$

then the above dynamics can be further simplified into: For any $\beta \in [M]$,

$$\begin{aligned}
\frac{\mathrm{d}\boldsymbol{W}_{p,q,\alpha,\beta}}{\mathrm{d}t} &\approx \sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta}\cdot \boldsymbol{z}_{u+p,v+q,\alpha},\\
\frac{\mathrm{d}\boldsymbol{b}_{\beta}}{\mathrm{d}t} &\approx \sum_{u=1}^{W_1}\sum_{v=1}^{H_1} \boldsymbol{a}_{u,v,\beta}\cdot z,\\
\frac{\mathrm{d}\boldsymbol{a}_{u,v,\beta}}{\mathrm{d}t} &\approx \sum_{\alpha=1}^{C_0}\left(\sum_{p=0}^{m-1}\sum_{q=0}^{m-1} \boldsymbol{z}_{u+p,v+q,\alpha}\cdot \boldsymbol{W}_{p,q,\alpha,\beta}\right)+\boldsymbol{b}_{\beta}\cdot z.
\end{aligned} \tag{67}$$

We observe that the training dynamics still takes the form

$$\frac{\mathrm{d}\boldsymbol{\theta}_{\beta}}{\mathrm{d}t} = \boldsymbol{A}\boldsymbol{\theta}_{\beta}, \tag{68}$$

except that in this case,

$$\begin{aligned}
\boldsymbol{\theta}_{\beta} := \Big(&\boldsymbol{W}_{0,0,1,\beta}, \boldsymbol{W}_{0,1,1,\beta}, \cdots, \boldsymbol{W}_{0,m-1,1,\beta}; \boldsymbol{W}_{1,0,1,\beta}, \cdots, \boldsymbol{W}_{1,m-1,1,\beta}; \cdots\cdots \boldsymbol{W}_{m-1,m-1,1,\beta};\\
&\boldsymbol{W}_{0,0,2,\beta}, \boldsymbol{W}_{0,1,2,\beta}, \cdots, \boldsymbol{W}_{0,m-1,2,\beta}; \boldsymbol{W}_{1,0,2,\beta}, \cdots, \boldsymbol{W}_{1,m-1,2,\beta}; \cdots\cdots \boldsymbol{W}_{m-1,m-1,2,\beta};\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\boldsymbol{W}_{0,0,C_0,\beta}, \boldsymbol{W}_{0,1,C_0,\beta}, \cdots, \boldsymbol{W}_{0,m-1,C_0,\beta}; \cdots, \boldsymbol{W}_{1,m-1,C_0,\beta}; \cdots\cdots \boldsymbol{W}_{m-1,m-1,C_0,\beta}; \boldsymbol{b}_{\beta};\\
&\boldsymbol{a}_{1,1,\beta}, \boldsymbol{a}_{1,2,\beta}, \cdots, \boldsymbol{a}_{1,H_1,\beta}; \boldsymbol{a}_{2,1,\beta}, \cdots, \boldsymbol{a}_{2,H_1,\beta}; \cdots\cdots \boldsymbol{a}_{W_1,H_1,\beta}\Big)^{\mathsf{T}},
\end{aligned}$$

or in more simplified notations,

$$\begin{aligned}
\boldsymbol{\theta}_{\beta} := \Big(&\boldsymbol{W}_{0,0:(m-1),1,\beta}; \boldsymbol{W}_{1,0:(m-1),1,\beta}; \cdots\cdots \boldsymbol{W}_{m-1,0:(m-1),1,\beta};\\
&\boldsymbol{W}_{0,0:(m-1),2,\beta}; \boldsymbol{W}_{1,0:(m-1),2,\beta}; \cdots\cdots \boldsymbol{W}_{m-1,0:(m-1),2,\beta};\\
&\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\cdots\\
&\boldsymbol{W}_{0,0:(m-1),C_0,\beta}; \boldsymbol{W}_{1,0:(m-1),C_0,\beta}; \cdots\cdots \boldsymbol{W}_{m-1,0:(m-1),C_0,\beta}; \boldsymbol{b}_{\beta};\\
&\boldsymbol{a}_{1,1:H_1,\beta}; \boldsymbol{a}_{2,1:H_1,\beta}; \cdots\cdots \boldsymbol{a}_{W_1,1:H_1,\beta}\Big)^{\mathsf{T}},
\end{aligned}$$

and

$$\boldsymbol{A} := \begin{bmatrix} \mathbf{0}_{(C_0m^2+1)\times(C_0m^2+1)} & \boldsymbol{Z}^\intercal \\ \boldsymbol{Z} & \mathbf{0}_{W_1H_1\times W_1H_1} \end{bmatrix}, \tag{69}$$

where $\boldsymbol{Z} \in \mathbb{R}^{W_1H_1\times(m^2+1)}$ and $\boldsymbol{Z}$ depends sorely on the input samples $\{\boldsymbol{x}_i\}_{i=1}^n$ and $\{y_i\}_{i=1}^n$, whose entries read

$$\begin{bmatrix}
\boldsymbol{z}_{1,1:m,1}; & \cdots & \boldsymbol{z}_{m,1:m,1}; & \boldsymbol{z}_{1,1:m,2}; & \cdots & \boldsymbol{z}_{m,1:m,2}; & \cdots\cdots & \boldsymbol{z}_{m,1:m,C_0}; & z \\
\boldsymbol{z}_{1,2:(m+1),1}; & \cdots & \boldsymbol{z}_{m,2:(m+1),1}; & \boldsymbol{z}_{1,2:(m+1),2}; & \cdots & \boldsymbol{z}_{m,2:(m+1),2}; & \cdots\cdots & \boldsymbol{z}_{m,2:(m+1),C_0}; & z \\
\vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{1,H_1:H_0,1}; & \cdots & \boldsymbol{z}_{m,H_1:H_0,1}; & \boldsymbol{z}_{1,H_1:H_0,2} & \cdots & \boldsymbol{z}_{m,H_1:H_0,2}; & \cdots\cdots & \boldsymbol{z}_{m,H_1:H_0,C_0}; & z \\
\boldsymbol{z}_{2,1:m,1}; & \cdots & \boldsymbol{z}_{m+1,1:m,1}; & \boldsymbol{z}_{2,1:m,2}; & \cdots & \boldsymbol{z}_{m+1,1:m,2}; & \cdots\cdots & \boldsymbol{z}_{m+1,1:m,C_0}; & z \\
\vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{2,H_1:H_0,1}; & \cdots & \boldsymbol{z}_{m+1,H_1:H_0,1}; & \boldsymbol{z}_{2,H_1:H_0,2} & \cdots & \boldsymbol{z}_{m+1,H_1:H_0,2}; & \cdots\cdots & \boldsymbol{z}_{m+1,H_1:H_0,C_0}; & z \\
\vdots & \cdots & \vdots & \vdots & \cdots & \vdots & \cdots\cdots & \vdots & \vdots \\
\boldsymbol{z}_{W_1,H_1:H_0,1}; & \cdots & \boldsymbol{z}_{W_0,H_1:H_0,1}; & \boldsymbol{z}_{W_1,H_1:H_0,2} & \cdots & \boldsymbol{z}_{W_0,H_1:H_0,2}; & \cdots\cdots & \boldsymbol{z}_{W_0,H_1:H_0,C_0}; & z
\end{bmatrix}. \tag{70}$$

We remark that all results in the case of single channel CNNs can be reproduced for multi channel CNNs, and we state a theorem without proof

Theorem 4. *Given any $\delta \in (0,1)$, under Assumption 5, Assumption 6, Assumption 7 and Assumption 8, if $\gamma > 1$, then with probability at least $1-\delta$ over the choice of $\boldsymbol{\theta}^0$, we have*

$$\lim_{M\to+\infty} \sup_{t\in[0,T_{\mathrm{eff}}]} \frac{\|\boldsymbol{\theta}_{\boldsymbol{W}}(t) - \boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(0)\|_2} = +\infty, \tag{71}$$

and

$$\lim_{M\to+\infty} \sup_{t\in[0,T_{\mathrm{eff}}]} \frac{\|\boldsymbol{\theta}_{\boldsymbol{W},\boldsymbol{v}_1}(t)\|_2}{\|\boldsymbol{\theta}_{\boldsymbol{W}}(t)\|_2} = 1. \tag{72}$$

To sum up, the weight vectors condense toward the unit eigenvector equipped with the largest singular value.