A APPENDIX

A.1 PROOF OF LEMMA 3.2

Proof. Let $\gamma_0 \in (0,1)$. Set $\gamma \triangleq \min\{\gamma_0, \gamma_0/c\}$. Denote $\tilde{\gamma} = \gamma/\sqrt{m}$. Then

$$\begin{aligned}
\left\|\boldsymbol{z}^{\ell+1} - \boldsymbol{z}^{\ell}\right\| =& \left\|\sigma\left(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z}^{\ell} + \boldsymbol{\phi}\right) - \sigma\left(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z}^{\ell-1} - \boldsymbol{\phi}\right)\right\| \\
\leq& \tilde{\gamma}\left\|\boldsymbol{A}\boldsymbol{z}^{\ell} - \boldsymbol{A}\boldsymbol{z}^{\ell-1}\right\|, \quad \sigma \text{ is 1-Lipschitz continuous} \\
=& \tilde{\gamma}\left\|\boldsymbol{A}(\boldsymbol{z}^{\ell} - \boldsymbol{z}^{\ell-1})\right\| \\
\leq& \tilde{\gamma}\|\boldsymbol{A}\|\|\boldsymbol{z}^{\ell} - \boldsymbol{z}^{\ell-1}\|, \\
\leq& \tilde{\gamma}c\sqrt{m}\|\boldsymbol{z}^{\ell} - \boldsymbol{z}^{\ell-1}\|, \quad \|\boldsymbol{A}\| \leq c\sqrt{m} \\
=& \gamma_0\|\boldsymbol{z}^{\ell} - \boldsymbol{z}^{\ell-1}\|.
\end{aligned}$$

Applying the above argument ℓ times, we obtain

$$\|\boldsymbol{z}^{\ell+1} - \boldsymbol{z}^{\ell}\| \leq \gamma_0^{\ell}\|\boldsymbol{z}^1 - \boldsymbol{z}^0\| = \gamma_0^{\ell}\|\boldsymbol{z}^1\| = \gamma_0^{\ell}\|\sigma(\boldsymbol{\phi})\| \leq \gamma_0^{\ell}\|\boldsymbol{\phi}\|,$$

where we use the fact $\boldsymbol{z}^0 = \boldsymbol{0}$. For any positive integers p, q with $p \leq q$, we have

$$\begin{aligned}
\|\boldsymbol{z}^p - \boldsymbol{z}^q\| \leq& \|\boldsymbol{z}^p - \boldsymbol{z}^{p+1}\| + \cdots + \|\boldsymbol{z}^{q-1} - \boldsymbol{z}^q\| \\
\leq& \gamma_0^p\|\boldsymbol{\phi}\| + \cdots + \gamma^q\|\boldsymbol{\phi}\| \\
\leq& \gamma_0^p\|\boldsymbol{\phi}\|\left(1 + \gamma + \gamma^2 + \cdots\right) \\
=& \frac{\gamma_0^p}{1-\gamma_0}\|\boldsymbol{\phi}\|.
\end{aligned}$$

Since $\gamma \in (0,1)$, we have $\|\boldsymbol{z}^p - \boldsymbol{z}^q\| \to 0$ as $p \to \infty$. Hence, $\{\boldsymbol{z}^{\ell}\}_{\ell=1}^{\infty}$ is a Cauchy sequence. Since $\mathbb{R}^m$ is complete, the equilibrium point $\boldsymbol{z}^*$ is the limit of the sequence $\{\boldsymbol{z}^{\ell}\}_{\ell=1}^{\infty}$, so that $\boldsymbol{z}$ exists and is unique. Moreover, let $q \to \infty$, then we obtain $\|\boldsymbol{z}^p - \boldsymbol{z}^*\| \leq \frac{\gamma^p}{1-\gamma}\|\boldsymbol{\phi}\|$, so that the fixed-point iteration converges to $\boldsymbol{z}$ linearly.

Let $p = 0$ and $q = \ell$, then we obtain $\|\boldsymbol{z}^{\ell}\| \leq \frac{1}{1-\gamma_0}\|\boldsymbol{\phi}\|$.

□

A.2 PROOF OF LEMMA 3.3

Proof. (i) To simplify the notations, we denote $\boldsymbol{D} \triangleq \mathbf{diag}(\sigma'(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z} + \boldsymbol{\phi}))$, and $\boldsymbol{E} \triangleq \mathbf{diag}(\sigma'(\boldsymbol{W}\boldsymbol{x}))$. The differential of f is given by

$$\begin{aligned}
df =& d(\boldsymbol{z} - \tilde{\gamma}\sigma(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z} + \boldsymbol{\phi})) \\
=& d\boldsymbol{z} - \boldsymbol{D}d(\tilde{\gamma}\boldsymbol{A}\boldsymbol{z} + \boldsymbol{\phi}) \\
=& \left[I_m - \tilde{\gamma}\boldsymbol{D}\boldsymbol{A}\right]d\boldsymbol{z} - \tilde{\gamma}\boldsymbol{D}(d\boldsymbol{A})\boldsymbol{z} - \boldsymbol{D}d\boldsymbol{\phi}.
\end{aligned}$$

Taking vectorization on both sides yields

$$\begin{aligned}
\text{vec}\,(df) =& \left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}\boldsymbol{A}\right]\text{vec}\,(d\boldsymbol{z}) - \text{vec}\,(\tilde{\gamma}\boldsymbol{D}d\boldsymbol{A}\boldsymbol{z}) - \boldsymbol{D}\text{vec}\,(d\boldsymbol{\phi}) \\
=& \left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}\boldsymbol{A}\right]\text{vec}\,(d\boldsymbol{z}) - \tilde{\gamma}[\boldsymbol{z}^T \otimes \boldsymbol{D}]\text{vec}\,(d\boldsymbol{A}) - \boldsymbol{D}\text{vec}\,(d\boldsymbol{\phi}).
\end{aligned}$$

Therefore, the partial derivative of f with respect to $\boldsymbol{z}$, $\boldsymbol{A}$, and $\boldsymbol{\phi}$ are given by

$$\begin{aligned}
\frac{\partial f}{\partial \boldsymbol{z}} &= \left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}\boldsymbol{A}\right]^T \\
\frac{\partial f}{\partial \boldsymbol{A}} &= -\tilde{\gamma}\left[\boldsymbol{z}^T \otimes \boldsymbol{D}\right]^T \\
\frac{\partial f}{\partial \boldsymbol{\phi}} &= -\boldsymbol{D}^T.
\end{aligned}$$

It follows from the definition of the feature vector $\boldsymbol{\phi}$ in equation 4 that

$$d\boldsymbol{\phi} = \frac{1}{\sqrt{m}} d\sigma(\boldsymbol{W}\boldsymbol{x}) = \frac{1}{\sqrt{m}} \boldsymbol{E}(d\boldsymbol{W})\boldsymbol{x} = \frac{1}{\sqrt{m}} \left[\boldsymbol{x}^T \otimes \boldsymbol{E}\right] \operatorname{vec}(\boldsymbol{W}).$$

Thus, the partial derivative of $\boldsymbol{\phi}$ with respect to $\boldsymbol{W}$ is given by

$$\frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{W}} = \frac{1}{\sqrt{m}} \left[\boldsymbol{x}^T \otimes \boldsymbol{E}\right]^T. \tag{23}$$

By using the chain rule, we obtain the partial derivative of f with respect to $\boldsymbol{W}$ as follows

$$\frac{\partial f}{\partial \boldsymbol{W}} = \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{W}} \frac{\partial f}{\partial \boldsymbol{\phi}} = -\frac{1}{\sqrt{m}} \left[\boldsymbol{x}^T \otimes \boldsymbol{E}\right]^T \boldsymbol{D}^T.$$

(ii) Let $\boldsymbol{v}$ be an arbitrary vector, and $\boldsymbol{u}$ be an arbitrary unit vector. The reverse triangle inequality implies that

$$\begin{aligned}
\|(\boldsymbol{I}_m - \tilde{\gamma}\, \mathbf{diag}(\sigma'(\boldsymbol{v}))\boldsymbol{A})\, \boldsymbol{u}\| \geq & \|\boldsymbol{u}\| - \|\tilde{\gamma}\, \mathbf{diag}(\sigma'(\boldsymbol{v}))\boldsymbol{A}\boldsymbol{u}\| \\
\geq & \|\boldsymbol{u}\| - \tilde{\gamma}\|\, \mathbf{diag}(\sigma'(\boldsymbol{v}))\|\|\boldsymbol{A}\|\|\boldsymbol{u}\| \\
\overset{(a)}{\geq} & (1-\gamma_0)\|\boldsymbol{u}\| \\
= & 1-\gamma_0 \\
> & 0,
\end{aligned}$$

where (a) is due to $|\sigma'(v)| \leq 1$ and $\|\boldsymbol{A}\|_{\text{op}} \leq c\sqrt{m}$. Therefore, taking infimum on the left-hand side over all unit vector $\boldsymbol{u}$ yields the desired result.

(iii) Since $f(\boldsymbol{z}^*, \boldsymbol{A}, \boldsymbol{W}) = 0$, taking implicit differentiation of f with respect to $\boldsymbol{A}$ at $\boldsymbol{z}^*$ gives us

$$\left(\frac{\partial \boldsymbol{z}}{\partial \boldsymbol{A}}\bigg|_{z=z^*}\right)\left(\frac{\partial f}{\partial \boldsymbol{z}}\bigg|_{z=z^*}\right) + \left(\frac{\partial f}{\partial \boldsymbol{A}}\bigg|_{\boldsymbol{z}=\boldsymbol{z}^*}\right) = 0$$

The results in part (i)-(ii) imply the smallest eigenvalue of $\frac{\partial f}{\partial \boldsymbol{z}}\big|_{\boldsymbol{z}^*}$ is strictly positive, so that it is invertible. Therefore, we have

$$\frac{\partial \boldsymbol{z}^*}{\partial \boldsymbol{A}} = -\left(\frac{\partial f}{\partial \boldsymbol{A}}\bigg|_{\boldsymbol{z}=\boldsymbol{z}^*}\right)\left(\frac{\partial f}{\partial \boldsymbol{z}}\bigg|_{z=z^*}\right)^{-1} = \tilde{\gamma}\left[\boldsymbol{z}^T \otimes \boldsymbol{D}\right]^T \left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}\boldsymbol{A}\right]^{-T} \tag{24}$$

Similarly, we obtain the partial derivative of $\boldsymbol{z}^*$ with respect to $\boldsymbol{W}$ as follows

$$\frac{\partial \boldsymbol{z}^*}{\partial \boldsymbol{W}} = -\left(\frac{\partial f}{\partial \boldsymbol{W}}\bigg|_{\boldsymbol{z}=\boldsymbol{z}^*}\right)\left(\frac{\partial f}{\partial \boldsymbol{z}}\bigg|_{z=z^*}\right)^{-1} = \frac{1}{\sqrt{m}}\left[\boldsymbol{x}^T \otimes \boldsymbol{E}\right]^T \boldsymbol{D}^T \left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}\boldsymbol{A}\right]^{-T} \tag{25}$$

To further simplify the notation, we denote $\boldsymbol{z}$ to be the equilibrium point $\boldsymbol{z}^*$ by omitting the superscribe, *i.e.*, $\boldsymbol{z} = \boldsymbol{z}^*$. Let $\hat{y} = \boldsymbol{u}^T\boldsymbol{z} + \boldsymbol{v}^T\boldsymbol{\phi}$ be the prediction for the training data $(\boldsymbol{x}, \boldsymbol{y})$. The differential of $\hat{y}$ is given by

$$d\hat{y} = d\left(\boldsymbol{u}^T\boldsymbol{z} + \boldsymbol{v}^T\boldsymbol{\phi}\right) = \boldsymbol{u}^T d\boldsymbol{z} + \boldsymbol{z} d\boldsymbol{u} + \boldsymbol{v}^T d\boldsymbol{\phi} + \boldsymbol{\phi}^T d\boldsymbol{v}.$$

The partial derivative of $\hat{y}$ with respect to $\boldsymbol{u}$, $\boldsymbol{v}$, $\boldsymbol{z}$, and $\boldsymbol{\phi}$ are given by

$$\frac{\partial \hat{y}}{\partial \boldsymbol{z}} = \boldsymbol{u}, \quad \frac{\partial \hat{y}}{\partial \boldsymbol{u}} = \boldsymbol{z}, \quad \frac{\partial \hat{y}}{\partial \boldsymbol{v}} = \boldsymbol{\phi}, \quad \frac{\partial \hat{y}}{\partial \boldsymbol{\phi}} = \boldsymbol{v}, \tag{26}$$

Let $\ell = \frac{1}{2}(\hat{y} - y)^2$. Then $\partial\ell/\partial\hat{y} = (\hat{y} - y)$. By chain rule, we have

$$\frac{\partial \ell}{\partial \boldsymbol{u}} = \frac{\partial \hat{y}}{\partial \boldsymbol{u}}\frac{\partial \ell}{\partial \hat{y}} = \boldsymbol{z}(\hat{y} - y) \tag{27}$$

$$\frac{\partial \ell}{\partial \boldsymbol{\phi}} = \frac{\partial \hat{y}}{\partial \boldsymbol{v}}\frac{\partial \ell}{\partial \hat{y}} = \boldsymbol{\phi}(\hat{y} - y). \tag{28}$$

By using equation 24-equation 25 and chain rule, we obtain

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{A}} &= \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{A}}\frac{\partial \ell}{\partial \boldsymbol{z}} \\ &= \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{A}}\frac{\partial \hat{y}}{\partial \boldsymbol{z}}\frac{\partial \ell}{\partial \hat{y}} = \tilde{\gamma}(\hat{y}-y)\left[\boldsymbol{z}^T \otimes \boldsymbol{D}\right]^T \left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}\boldsymbol{A}\right]^{-T}\boldsymbol{u},\end{aligned} \tag{29}$$

and

$$\begin{aligned}\frac{\partial \ell}{\partial \boldsymbol{W}} &= \frac{\partial \boldsymbol{z}}{\partial \boldsymbol{W}}\frac{\partial \hat{y}}{\partial \boldsymbol{z}}\frac{\partial \ell}{\partial \hat{y}} + \frac{\partial \boldsymbol{\phi}}{\partial \boldsymbol{W}}\frac{\partial \hat{y}}{\partial \boldsymbol{\phi}}\frac{\partial \ell}{\partial \hat{y}} \\ &= \frac{1}{\sqrt{m}}(\hat{y}-y)[\boldsymbol{x}^T \otimes \boldsymbol{E}]^T \left[\boldsymbol{D}^T(I_m - \tilde{\gamma}\boldsymbol{D}\boldsymbol{A})^{-T}\boldsymbol{u} + \boldsymbol{v}\right].\end{aligned} \tag{30}$$

Since $L = \sum_{i=1}^n \ell_i$ with $\ell_i = \ell(\hat{y}_i, y_i)$, we have $dL = \sum_{i=1}^n d\ell_i$ and $\partial L/\partial \ell_i = 1$. Therefore, we obtain

$$\frac{\partial L}{\partial \boldsymbol{A}} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \boldsymbol{A}} = \sum_{i=1}^n \tilde{\gamma}(\hat{y}_i - y_i)\left[\boldsymbol{z}_i^T \otimes \boldsymbol{D}_i\right]^T \left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A}\right]^{-T}\boldsymbol{u} \tag{31}$$

$$\frac{\partial L}{\partial \boldsymbol{W}} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \boldsymbol{W}} = \sum_{i=1}^n \frac{1}{\sqrt{m}}(\hat{y}_i - y_i)[\boldsymbol{x}_i^T \otimes \boldsymbol{E}_i]^T \left[\boldsymbol{D}_i^T(I_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A})^{-T}\boldsymbol{u} + \boldsymbol{v}\right] \tag{32}$$

$$\frac{\partial L}{\partial \boldsymbol{u}} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \boldsymbol{u}} = \sum_{i=1}^n (\hat{y}_i - y_i)\boldsymbol{z}_i \tag{33}$$

$$\frac{\partial L}{\partial \boldsymbol{v}} = \sum_{i=1}^n \frac{\partial \ell_i}{\partial \boldsymbol{v}} = \sum_{i=1}^n (\hat{y}_i - y_i)\boldsymbol{\phi}_i \tag{34}$$

□

A.3 PROOF OF LEMMA 3.4

Proof. Let $\boldsymbol{z}_i$ denote the i-th equilibrium point of $\boldsymbol{x}_i$. By using equation 24, 25, 31 and 32, we obtain the dynamics of the equilibrium point $\boldsymbol{z}_i$ as follows

$$\begin{aligned}\frac{d\boldsymbol{z}_i}{dt} &= \left(\frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{A}}\right)^T \frac{d\mathrm{vec}\left(\boldsymbol{A}\right)}{dt} + \left(\frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{W}}\right)^T \frac{d\mathrm{vec}\left(\boldsymbol{W}\right)}{dt} \\ &= \left(\frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{A}}\right)^T \left(-\frac{\partial L}{\partial \boldsymbol{A}}\right) + \left(\frac{\partial \boldsymbol{z}_i}{\partial \boldsymbol{W}}\right)^T \left(-\frac{\partial L}{\partial \boldsymbol{W}}\right) \\ &= -\tilde{\gamma}^2 \sum_{j=1}^n (\hat{y}_j - y_j)\left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A}\right]^{-1}\left[\boldsymbol{z}_i^T \otimes \boldsymbol{D}_i\right]\left[\boldsymbol{z}_j^T \otimes \boldsymbol{D}_j\right]^T \left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A}\right]^{-T}\boldsymbol{u} \\ &\quad - \frac{1}{m}\sum_{j=1}^n (\hat{y}_j - y_j)\left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A}\right]^{-1}\boldsymbol{D}_i\left[\boldsymbol{x}_i^T \otimes \boldsymbol{E}_i\right][\boldsymbol{x}_j^T \otimes \boldsymbol{E}_j]^T \left[\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u} + \boldsymbol{v}\right] \\ &= -\tilde{\gamma}^2 \sum_{j=1}^n (\hat{y}_j - y_j)\left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A}\right]^{-1}\boldsymbol{D}_i\boldsymbol{D}_j^T\left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A}\right]^{-T}\boldsymbol{u}\boldsymbol{z}_i^T\boldsymbol{z}_j \\ &\quad - \frac{1}{m}\sum_{j=1}^n (\hat{y}_j - y_j)\left[\boldsymbol{I}_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A}\right]^{-1}\boldsymbol{D}_i\boldsymbol{E}_i\boldsymbol{E}_j^T\left[\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u} + \boldsymbol{v}\right]\boldsymbol{x}_i^T\boldsymbol{x}_j.\end{aligned}$$

By using equation 23 and 32, we obtain the dynamics of the feature vector $\boldsymbol{\phi}_i$

$$\begin{aligned}\frac{d\boldsymbol{\phi}_i}{dt} &= \left(\frac{\partial \boldsymbol{\phi}_i}{\partial \boldsymbol{W}}\right)^T \frac{d\text{vec}\left(\boldsymbol{W}\right)}{dt} \\ &= \left(\frac{\partial \boldsymbol{\phi}_i}{\partial \boldsymbol{W}}\right)^T \left(-\frac{\partial L}{\partial W}\right) \\ &= -\frac{1}{m}\sum_{j=1}^{n}(\hat{y}_i - y_i)\boldsymbol{E}_i\boldsymbol{E}_j^T[\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u} + \boldsymbol{v}]\boldsymbol{x}_i^T\boldsymbol{x}_j.\end{aligned}$$

By chain rule, the dynamics of the prediction $\hat{y}_i$ is given by

$$\begin{aligned}\frac{d\hat{y}_i}{dt} &= \left(\frac{\partial \hat{y}_i}{\partial \boldsymbol{z}_i}\right)^T \frac{d\boldsymbol{z}_i}{dt} + \left(\frac{\partial \hat{y}_i}{\partial \boldsymbol{\phi}_i}\right)^T \frac{d\boldsymbol{\phi}_i}{dt} + \left(\frac{\partial \hat{y}_i}{\partial \boldsymbol{u}}\right)^T \frac{d\boldsymbol{u}}{dt} + \left(\frac{\partial \hat{y}_i}{\partial \boldsymbol{v}}\right)^T \frac{d\boldsymbol{v}}{dt} \\ &= -\tilde{\gamma}^2\sum_{j=1}^{n}(\hat{y}_j - y_j)\left[\boldsymbol{u}^T(I_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A})^{-1}\boldsymbol{D}_i\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u}\right](\boldsymbol{z}_i^T\boldsymbol{z}_j) \\ &\quad - \frac{1}{m}\sum_{j=1}^{n}(\hat{y}_j - y_j)\left[\left(\boldsymbol{D}_i^T(I_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A})^{-1}\boldsymbol{u} + \boldsymbol{v}\right)^T \boldsymbol{E}_i\boldsymbol{E}_j^T\left(\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u} + \boldsymbol{v}\right)\right](\boldsymbol{x}_i^T\boldsymbol{x}_j) \\ &\quad - \sum_{j=1}^{n}(\hat{y}_j - y_j)(\boldsymbol{z}_i^T\boldsymbol{z}_j) \\ &\quad - \sum_{j=1}^{n}(\hat{y}_j - y_j)(\boldsymbol{\phi}_i^T\boldsymbol{\phi}_j).\end{aligned}$$

Define the matrices $\boldsymbol{M}(t) \in \mathbb{R}^{n\times n}$ and $\boldsymbol{Q}(t) \in \mathbb{R}^{n\times n}$ as follows

$$\begin{aligned}\boldsymbol{M}(t)_{ij} &\triangleq \frac{1}{m}\boldsymbol{u}^T(I_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A})^{-1}\boldsymbol{D}_i\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u}, \\ \boldsymbol{Q}(t)_{ij} &\triangleq \frac{1}{m}\left(\boldsymbol{D}_i^T(I_m - \tilde{\gamma}\boldsymbol{D}_i\boldsymbol{A})^{-1}\boldsymbol{u} + \boldsymbol{v}\right)^T \boldsymbol{E}_i\boldsymbol{E}_j^T\left(\boldsymbol{D}_j^T(I_m - \tilde{\gamma}\boldsymbol{D}_j\boldsymbol{A})^{-T}\boldsymbol{u} + \boldsymbol{v}\right).\end{aligned}$$

Let $\boldsymbol{X} \in \mathbb{R}^{n\times d}$, $\boldsymbol{\Phi}(t) \in \mathbb{R}^{n\times m}$, and $\boldsymbol{Z}(t) \in \mathbb{R}^{n\times m}$ be the matrices whose rows are the training data $\boldsymbol{x}_i$, feature vectors $\boldsymbol{\phi}_i$, and equilibrium points $\boldsymbol{z}_i$ at time t, respectively. The dynamics of the prediction vector $\hat{\boldsymbol{y}}$ is given by

$$\frac{d\hat{\boldsymbol{y}}}{dt} = -\left[\left(\gamma^2\boldsymbol{M}(t) + \boldsymbol{I}_n\right)\circ\boldsymbol{Z}(t)\boldsymbol{Z}(t)^T + \boldsymbol{Q}(t)\circ\boldsymbol{X}\boldsymbol{X}^T + \boldsymbol{\Phi}(t)\boldsymbol{\Phi}(t)^T\right](\hat{\boldsymbol{y}}(t) - \boldsymbol{y}).$$

□

A.4 Proof of Lemma 3.5

A.4.1 Review of Hermite Expansions

To make the paper self-contained, we review the necessary background about the Hermite polynomials in this section. One can find each result in this section from any standard textbooks about functional analysis such as MacCluer (2008); Kreyszig (1978), or most recent literature (Nguyen & Mondelli, 2020, Appendix D) and (Oymak & Soltanolkotabi, 2020, Appendix H).

We consider an L^2-space defined by $L^2(\mathbb{R}, dP)$, where dP is the *Gaussian measure*, that is,

$$dP = p(x)dx, \quad \text{where} \quad p(x) = \frac{1}{\sqrt{2\pi}}e^{-\frac{x^2}{2}}.$$

Thus, $L^2(\mathbb{R}, dP)$ is a collection of functions f for which

$$\int_{-\infty}^{\infty}|f(x)|^2\,dP(x) = \int_{-\infty}^{\infty}|f(x)|^2\,p(x)dx = \mathbb{E}_{x\sim N(0,1)}\,|f(x)|^2 < \infty.$$

Lemma A.1. The relu activation $\sigma \in L^2(\mathbb{R}, dP)$.

Proof. Note that

$$\int_{-\infty}^{\infty} |\sigma(x)|^2 p(x)dx \leq \int_{-\infty}^{\infty} |x|^2 p(x)dx = \mathbb{E}_{x\sim N(0,1)} |x|^2 = \mathrm{Var}(x) = 1.$$

□

For any functions $f, g \in L^2(\mathbb{R}, dP)$, we define an *inner product*

$$\langle f, g\rangle := \int_{-\infty}^{\infty} f(x)g(x)dP(x) = \int_{-\infty}^{\infty} f(x)g(x)p(x)dx = \mathbb{E}_{x\sim N(0,1)}[f(x)g(x)].$$

Furthermore, the induced norm $\|\cdot\|$ is given by

$$\|f\|^2 = \langle f, f\rangle = \int_{-\infty}^{\infty} |f(x)|^2 dP(x) = \mathbb{E}_{x\sim N(0,1)} |f(x)|^2 .$$

This L^2 space has an orthonormal basis with respect to the inner product defined above, called *normalized probabilist's Hermite polynomials* $\{h_n(x)\}_{n=0}^{\infty}$ that are given by

$$h_n(x) = \frac{1}{\sqrt{n!}}(-1)^n e^{x^2/2} D^n(e^{-x^2/2}), \quad \text{where} \quad D^n(e^{-x^2/2}) = \frac{d^n}{dx^n} e^{-x^2/2}.$$

Lemma A.2. The *normalized probabilist's Hermite polynomials* is an orthonormal basis of $L^2(\mathbb{R}, dP)$: $\langle h_m, h_n\rangle = \delta_{mn}$.

Proof. Note that $D^n(e^{-x^2/2}) = e^{-x^2/2}P_n(x)$ for a polynomial with degree of n and leading term is $(-1)^n x^n$. Thus, we can consider $h_n(x) = \frac{1}{\sqrt{n!}}(-1)^n P_n(x)$.

Assume $m < n$

$$\begin{aligned}
\langle h_n, h_m\rangle =& \mathbb{E}_{x\sim N(0,1)}[h_n(x)h_m(x)] \\
=& \int_{-\infty}^{\infty} h_n(x)h_m(x)\frac{1}{\sqrt{2\pi}}e^{-x^2/2}dx, \\
=& \frac{1}{\sqrt{2\pi}\sqrt{n!}}(-1)^n \int_{-\infty}^{\infty} D^n(e^{-x^2/2})h_m(x)dx, \quad \text{rewrite } h_n(x) \text{ by its definition} \\
=& \frac{1}{\sqrt{2\pi}\sqrt{n!}\sqrt{m!}}(-1)^{n+m} \int_{-\infty}^{\infty} D^n(e^{-x^2/2})P_m(x)dx, \quad \text{rewrite } h_m \text{ by the polynomial form} \\
=& \frac{1}{\sqrt{2\pi}\sqrt{n!}\sqrt{m!}}(-1)^{2n+m} \int_{-\infty}^{\infty} e^{-x^2/2}D_n[P_m(x)]dx, \quad \text{integration by parts } n \text{ times}
\end{aligned}$$

There is no boundary terms because the super exponential decay of $e^{-x^2/2}$ at infinity. Since $m < n$, then $D_n(P_m) = 0$ so that $\langle h_m, h_n\rangle$=0. If $m = n$, then $D_n(P_m) = (-1)^n n!$. Thus, $\langle h_n, h_n\rangle = 1$. □

Remark: Since $\{h_n\}$ is an orthonormal basis, for every $f \in L^2(\mathbb{R}, dP)$, we have

$$f(x) = \sum_{n=0}^{\infty} \langle f, h_n\rangle h_n(x)$$

in the sense that

$$\lim_{N\to\infty} \left\| f(x) - \sum_{n=0}^{N} \langle f, h_n\rangle h_n(x) \right\|^2 = \lim_{N\to\infty} \mathbb{E}_{x\sim N(0,1)} \left| f(x) - \sum_{n=0}^{N} \langle f, h_n\rangle h_n(x) \right|^2 = 0$$

Lemma A.3. $f \in L^2(\mathbb{R}, dP)$ if and only if $\sum_{n=0}^{\infty} |\langle f, h_n\rangle|^2 < \infty$.

Proof. Note that

$$
\begin{aligned}
\langle f, f\rangle &= \int_{-\infty}^{\infty} |f(x)|^2 \, dP(x) \\
&= \int_{-\infty}^{\infty} \left(\sum_{i=0}^{\infty} \langle f, h_i\rangle \, h_i(x) \right) \left(\sum_{j=0}^{\infty} \langle f, h_j\rangle \, h_j(x) \right) dP(x) \\
&= \sum_{i,j=0}^{\infty} \langle f, h_i\rangle \langle f, h_j\rangle \int_{-\infty}^{\infty} h_i(x) h_j(x) dP(x) \\
&= \sum_{i=1}^{\infty} |\langle f, h_i\rangle|^2 .
\end{aligned}
$$

□

Lemma A.4. Consider a Hilbert space H with inner product $\langle \cdot, \cdot \rangle$. If $\|f_n - f\| \to 0$ and $\|g_n - g\| \to 0$, then $\langle f, g\rangle = \lim_{n\to\infty} \langle f_n, g_n\rangle$.

Proof. Observe that

$$
\begin{aligned}
|\langle f, g\rangle - \langle f_n, g_n\rangle| &\leq |\langle f, g\rangle - \langle f_n, g\rangle| + |\langle f_n, g\rangle - \langle f_n, g_n\rangle| \\
&\leq \|f\| \|g - g_n\| + \|f_n\| \|g - g_n\|.
\end{aligned}
$$

Let $n \to \infty$, then the continuity of $\|\cdot\|$ implies the desired result. □

Lemma A.5. Let $\{h_n(x)\}$ be the normalized probabilist's Hermite polynomials. For any fixed number t, we have

$$
e^{xt - t^2/2} = \sum_{n=0}^{\infty} \frac{t^n}{\sqrt{n!}} h_n(x). \tag{35}
$$

Proof. First show $f(x) = e^{xt - t^2/t} \in H \triangleq L^2(\mathbb{R}, dP)$.

$$
\begin{aligned}
\langle f, f\rangle &= \mathbb{E}_{x\sim N(0,1)} |f(x)|^2 \\
&= \int_{-\infty}^{\infty} e^{2xt - t^2} \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\
&= e^{t^2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} \exp\left\{ -\frac{(x-2t)^2}{2} \right\} dx, \quad x \sim N(2t, 1) \\
&= e^{t^2} < \infty.
\end{aligned}
$$

Thus $f(x) \in H$. Then $f(x) = \sum_{n=0}^{\infty} \langle f, h_n\rangle \, h_n(x)$. Note that

$$
\begin{aligned}
\langle f, h_n\rangle &= \mathbb{E}_{x\sim N(0,1)}[f(x) h_n(x)] \\
&= \int_{-\infty}^{\infty} e^{xt - t^2/2} \cdot \frac{1}{\sqrt{n!}} (-1)^n e^{x^2/2} D_n(e^{-x^2/2}) \cdot \frac{1}{\sqrt{2\pi}} e^{-x^2/2} dx \\
&= \frac{1}{\sqrt{n!}} (-1)^n \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{xt - t^2/2} \cdot D_n(e^{-x^2/2}) dx, \quad \text{integration by parts } n \text{ times} \\
&= \frac{1}{\sqrt{n!}} (-1)^{2n} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{xt - t^2/2} t^n \cdot e^{-x^2/2} dx \\
&= \frac{t^n}{\sqrt{n!}} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(x-t)^2/2} dx, \quad x \sim N(t, 1) \\
&= \frac{t^n}{\sqrt{n!}}
\end{aligned}
$$

□

Lemma A.6. Let $\boldsymbol{a}, \boldsymbol{b} \in \mathbb{R}^d$ with $\|\boldsymbol{a}\| = \|\boldsymbol{b}\| = 1$, then

$$\mathbb{E}_{w \sim N(\mathbf{0}, \boldsymbol{I}_d)}[h_n(\langle \boldsymbol{a}, \boldsymbol{w} \rangle) h_m(\langle \boldsymbol{b}, \boldsymbol{w} \rangle)] = \langle \boldsymbol{a}, \boldsymbol{b} \rangle^n \delta_{mn}.$$

Proof. Given fixed numbers s and t, we define two functions $f(\boldsymbol{w}) = e^{\langle \boldsymbol{a}, \boldsymbol{w} \rangle t - t^2/2}$ and $g(\boldsymbol{w}) = e^{\langle \boldsymbol{b}, \boldsymbol{w} \rangle s - s^2/2}$. Let $x = \langle \boldsymbol{a}, \boldsymbol{w} \rangle$ and $y = \langle \boldsymbol{b}, w \rangle$. Then we have

$$\begin{aligned} f(\boldsymbol{w}) =& e^{\langle \boldsymbol{a}, \boldsymbol{w} \rangle t - t^2/2} = e^{xt - t^2/2} = \sum_{n=0}^{\infty} \frac{t^n}{\sqrt{n!}} h_n(x) = \sum_{n=0}^{\infty} \frac{t^n}{\sqrt{n!}} h_n(\langle \boldsymbol{a}, \boldsymbol{w} \rangle), \\ g(\boldsymbol{w}) =& e^{\langle \boldsymbol{b}, \boldsymbol{w} \rangle s - s^2/2} = e^{ys - s^2/2} = \sum_{n=0}^{\infty} \frac{s^n}{\sqrt{n!}} h_n(y) = \sum_{n=0}^{\infty} \frac{s^n}{\sqrt{n!}} h_n(\langle \boldsymbol{b}, \boldsymbol{w} \rangle). \end{aligned}$$

Define a Hilbert space $H_d = L^2(\mathbb{R}^d, dP)$, where dP is the *multivariate Gaussian measure*, equipped with inner product $\langle f, g \rangle \triangleq \mathbb{E}_{\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{I}_d)}[f(\boldsymbol{w}) g(\boldsymbol{w})]$. Clearly, $f, g \in H_d$. Define sequences $\{f_N\}$ and $\{g_N\}$ as follows

$$f_N(\boldsymbol{w}) = \sum_{n=0}^{N} \frac{t^n}{\sqrt{n!}} h_n(\langle \boldsymbol{a}, \boldsymbol{w} \rangle) \quad \text{and} \quad g_N(\boldsymbol{w}) = \sum_{n=0}^{N} \frac{s^n}{\sqrt{n!}} h_n(\langle \boldsymbol{b}, \boldsymbol{w} \rangle).$$

Since $\|f - f_N\| \to 0$ and $\|g - g_N\| \to 0$, we have

$$\begin{aligned} \mathbb{E}_{\boldsymbol{w} \sim N(0, I_d)}[f(\boldsymbol{w}) g(\boldsymbol{w})] &= \langle f, g \rangle \\ &= \lim_{N \to \infty} \langle f_N, g_N \rangle \\ &= \lim_{N \to \infty} \mathbb{E}_{\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{I}_d)}[f_N(\boldsymbol{w}) g_N(\boldsymbol{w})] \\ &= \lim_{N \to \infty} \sum_{n,m=0}^{N} \frac{t^n s^m}{\sqrt{n!}\sqrt{m!}} \mathbb{E}_{\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{I}_d)}[h_n(\langle \boldsymbol{a}, \boldsymbol{w} \rangle) g_n(\langle \boldsymbol{b}, \boldsymbol{w} \rangle)] \end{aligned}$$

Note that the LHS is also given by

$$\begin{aligned} \mathbb{E}_{\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{I}_d)}[f(\boldsymbol{w}) g(\boldsymbol{w})] =& e^{-t^2/2 - s^2/2} \mathbb{E}_{\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{I}_d)}[e^{\langle \boldsymbol{a}, \boldsymbol{w} \rangle t + \langle \boldsymbol{b}, \boldsymbol{w} \rangle s}] \\ =& e^{-t^2/2 - s^2/2} \mathbb{E}_{\boldsymbol{w} \sim N(\mathbf{0}, \boldsymbol{I}_d)}[e^{\sum_{i=1}^{d} \boldsymbol{w}_i (a_i t + b_i s)}] \\ =& e^{-t^2/2 - s^2/2} \prod_{i=1}^{d} \mathbb{E}_{w_i \sim N(0,1)}[e^{\boldsymbol{w}_i (a_i t + b_i s)}] \\ =& e^{-t^2/2 - s^2/2} \prod_{i=1}^{d} M_{\boldsymbol{w}_i}(a_i t + b_i s) \\ =& e^{\langle \boldsymbol{a}, \boldsymbol{b} \rangle st} \\ =& \sum_{n=0}^{\infty} \frac{\langle \boldsymbol{a}, \boldsymbol{b} \rangle^n (st)^n}{n!}. \end{aligned}$$

Since s and t are arbitrary numbers, matching the coefficients yields

$$\mathbb{E}_{\boldsymbol{w} \sim N(0, \boldsymbol{I}_d)}[h_n(\langle a, \boldsymbol{w} \rangle) h_m(\langle b, \boldsymbol{w} \rangle)] = \langle \boldsymbol{a}, \boldsymbol{b} \rangle^n \delta_{mn}.$$

□

A.4.2 LOWER BOUND THE SMALLEST EIGENVALUES OF $\boldsymbol{G}^{\infty}$

The result in this subsection is similar to the results in (Nguyen & Mondelli, 2020, Appendix D) and (Oymak & Soltanolkotabi, 2020, Appendix H). The key difference is the assumptions made on the training data. In particular, Oymak & Soltanolkotabi (2020) assumes the training data is δ-separable, *i.e.*, $\min\{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|, \|\boldsymbol{x}_i + \boldsymbol{x}_j\|\} \geq \delta > 0$ for all $i \neq j$, and Nguyen & Mondelli (2020) assumes the data $\boldsymbol{x}_i$ follows some sub-Gaussian random variable, while we assume no two data are parallel to each other, *i.e.*, $\boldsymbol{x}_i \not\parallel \boldsymbol{x}_j$ for all $i \neq j$.

Lemma A.7. Given an activation function σ, if $\sigma \in L^2(\mathbb{R}, dP)$ and $\|\boldsymbol{x}_i\| = 1$ for all $i \in [n]$, then

$$\boldsymbol{G}^{\infty} = \sum_{k=0}^{\infty} |\langle \sigma, h_k \rangle|^2 \underbrace{\left(\boldsymbol{X}\boldsymbol{X}^T \circ \cdots \circ \boldsymbol{X}\boldsymbol{X}^T\right)}_{k \text{ times}} \tag{36}$$

where $\circ$ is elementwise product.

Proof. Observe

$$\begin{aligned} \boldsymbol{G}_{ij}^{\infty} =& \mathbb{E}_{\boldsymbol{w} \sim N(0, \boldsymbol{I}_d)} \left[\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle)\sigma(\langle \boldsymbol{w}, \boldsymbol{x}_j \rangle)\right] \\ =& \sum_{k,\ell=0}^{\infty} \langle \sigma, h_k \rangle \langle \sigma, h_\ell \rangle \mathbb{E}_{\boldsymbol{w} \sim N(0, I_d)} \left[h_k(\langle \boldsymbol{w}, \boldsymbol{x}_i \rangle) h_\ell(\langle \boldsymbol{w}, \boldsymbol{x}_j \rangle)\right] \\ =& \sum_{k,\ell=0}^{\infty} \langle \sigma, h_k \rangle \langle \sigma, h_\ell \rangle \cdot \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^k \delta_{k\ell} \\ =& \sum_{k=0}^{\infty} \langle \sigma, h_k \rangle^2 \langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^k \end{aligned}$$

$\square$

Note that the tensor product of $\boldsymbol{x}_i$ and $\boldsymbol{x}_i$ is $\boldsymbol{x}_i \otimes \boldsymbol{x}_i \in \mathbb{R}^{d^2 \times 1}$, so that

$$\langle \boldsymbol{x}_i, \boldsymbol{x}_j \rangle^k = \left\langle \underbrace{\boldsymbol{x}_i \otimes \cdots \otimes \boldsymbol{x}_i}_{k \text{ times}}, \underbrace{\boldsymbol{x}_j \otimes \cdots \otimes \boldsymbol{x}_j}_{k \text{ times}} \right\rangle$$

Here we introduce the *(row-wise) Khatri–Rao product* of two matrices $\boldsymbol{A} \in \mathbb{R}^{k \times m}$, $\boldsymbol{B} \in \mathbb{R}^{k \times n}$. Then

$$\boldsymbol{A} * \boldsymbol{B} = \begin{bmatrix} \boldsymbol{A}_{1*} \otimes \boldsymbol{B}_{1*} \\ \vdots \\ \boldsymbol{A}_{k*} \otimes \boldsymbol{B}_{k*} \end{bmatrix} \in \mathbb{R}^{k \times mn}$$

where $\boldsymbol{A}_{i*}$ indicates the i-th row of matrix $\boldsymbol{A}$. Therefore, the i-th row of $\boldsymbol{X} * \cdots * \boldsymbol{X} \triangleq \boldsymbol{X}^{*n}$ is $\boldsymbol{x}_i \otimes \cdots \otimes \boldsymbol{x}_i$. As a result, we obtain a more compact form of equation 36 as follows

$$\boldsymbol{G}^{\infty} = \sum_{k=0}^{\infty} |\langle \sigma, h_k \rangle|^2 (\boldsymbol{X}^{*k})(\boldsymbol{X}^{*k})^T. \tag{37}$$

Lemma A.8. If $\sigma(x)$ is a nonlinear function and $|\sigma(x)| \leq |x|$ and , then

$$\sup\{n : \langle \sigma, h_n \rangle > 0\} = \infty.$$

Proof. It is equivalent to show $\sigma(x)$ is not a finite linear combination of polynomials. We prove by contradiction. Suppose $\sigma(x) = a_0 + a_1 x + \cdots + a_n x^n$. Since $\sigma(0) = 0 = a_0$, then $\sigma(x) = a_1 x + \cdots + a_n x^n$. Observe that

$$\begin{aligned} \lim_{x \to \infty} \frac{|\sigma(x)|}{|x|} &= \lim_{x \to \infty} \frac{|a_1 x + \cdots + a_n x^n|}{|x|} \\ &= \lim_{x \to \infty} \left|a_1 + \cdots + a_n x^{n-1}\right|, \\ &= \infty \end{aligned}$$

which contradicts $\frac{|\sigma(x)|}{|x|} \leq 1$ for all $x \neq 0$. $\square$

Lemma A.9. If $\boldsymbol{x}_i \nparallel \boldsymbol{x}_j$ for all $i \neq j$, then there exists $k_0 > 0$ such that $\lambda_{\min}\left[(\boldsymbol{X}^{*k})(\boldsymbol{X}^{*k})^T\right] > 0$ for all $k \geq k_0$. Therefore, $\lambda_{\min}(\boldsymbol{G}^{\infty}) > 0$.

Proof. To simplify the notation, denote $\boldsymbol{K} = (\boldsymbol{X}^{*k})^T \in \mathbb{R}^{kd\times n}$. Since $x_i \nparallel \boldsymbol{x}_j$ and $\|\boldsymbol{x}_i\| = 1$, then let $\delta \triangleq \max\{|\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle|\} = \max\{|\cos\theta_{ij}|\}$ and $\delta \in (0, 1)$, where θ_{ij} is the angle between $\boldsymbol{x}_i$ and $\boldsymbol{x}_j$. For any unit vector $\boldsymbol{v} \in \mathbb{R}^n$, we have

$$
\begin{aligned}
\boldsymbol{v}^T(\boldsymbol{X}^{*k})(\boldsymbol{X}^{*k})^T\boldsymbol{v} =& \|\boldsymbol{K}\boldsymbol{v}\|^2 \\
=& \left\|\sum_{i=1}^n v_i \boldsymbol{K}_{*i}\right\|^2 \\
=& \sum_{i=1}^n \sum_{j=1}^n v_i v_j \left\langle \boldsymbol{K}_{*i}, \boldsymbol{K}_{*j}\right\rangle \\
=& \sum_{i=1}^n \sum_{j=1}^n v_i v_j \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j\right\rangle^k \\
=& \sum_{i=1}^n v_i^2 \|\boldsymbol{x}_i\|^{2k} + \sum_{i\neq j} v_i v_j \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j\right\rangle^k \\
=& 1 + \sum_{i\neq j} v_i v_j \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j\right\rangle^k,
\end{aligned}
$$

where the last equality is because $\|\boldsymbol{x}_i\| = 1$ and $\|\boldsymbol{v}\| = 1$. Note that

$$
\begin{aligned}
\left|\sum_{i\neq j} v_i v_j \left\langle \boldsymbol{x}_i, \boldsymbol{x}_j\right\rangle^k\right| \leq& \sum_{i\neq j} |v_i|\,|v_j|\,|\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle|^k \\
\leq& \delta^k \sum_{i\neq j} |v_i|\,|v_j|, \quad \text{by } |\langle \boldsymbol{x}_i, \boldsymbol{x}_j\rangle| \leq \delta \\
\leq& \delta^k \left(\sum_{i=1}^n |v_i|\right)^2 \\
\leq& n\delta^k, \quad \text{by Cauchy-Schwart inequlity.}
\end{aligned}
$$

By inverse triangle inequality, we have

$$
\|\boldsymbol{K}\boldsymbol{v}\|^2 \geq 1 - n\delta^k.
$$

Choose $k_0 \geq \log n / \log(1/\delta)$, then $\lambda_{\min}\{(\boldsymbol{X}^{*k})(\boldsymbol{X}^{*k})^T\} > 0$ for all $k \geq k_0$. $\square$

A.5 Proof of Lemma 3.6

Proof. By using the concentration inequality for standard Gaussian random variables, we have

$$\begin{aligned}
\mathbb{P}\left\{\|\boldsymbol{G}(0)-\boldsymbol{G}^{\infty}\|_2 \geq \frac{\lambda_0}{4}\right\} \leq & \mathbb{P}\left\{\|\boldsymbol{G}(0)-\boldsymbol{G}^{\infty}\|_F \geq \frac{\lambda_0}{4}\right\} \\
= & \mathbb{P}\left\{\|\boldsymbol{G}(0)-\boldsymbol{G}^{\infty}\|_F^2 \geq \left(\frac{\lambda_0}{4}\right)^2\right\} \\
= & \mathbb{P}\left\{\sum_{i,j=1}^{n}\left|\boldsymbol{G}_{ij}(0)-\boldsymbol{G}_{ij}^{\infty}\right|^2 \geq \left(\frac{\lambda_0}{4}\right)^2\right\} \\
\leq & \sum_{i,j=1}^{n}\mathbb{P}\left\{\left|\boldsymbol{G}_{ij}(0)-\boldsymbol{G}_{ij}^{\infty}\right|^2 \geq \left(\frac{\lambda_0}{4n}\right)^2\right\} \\
= & \sum_{i,j=1}^{n}\mathbb{P}\left\{\left|\boldsymbol{G}_{ij}(0)-\boldsymbol{G}_{ij}^{\infty}\right| \geq \frac{\lambda_0}{4n}\right\} \\
\leq & n^2 2\exp\left\{-\frac{2m(\lambda_0/4n)^2}{2^2}\right\} \\
\leq & \delta,
\end{aligned}$$

where we use the fact $\|\boldsymbol{X}\|_2 \leq \|\boldsymbol{X}\|_F$, and $\mathbb{P}\{\sum_{i=1}^n x_i \geq \varepsilon\} \leq \sum_{i=1}^n \mathbb{P}\{x_i \geq \varepsilon/n\}$. □

A.6 Proof of Lemma 3.7

Proof. By using the 1-Lipschitz continuity of $\sigma(x)$, we have

$$\begin{aligned}
\|\boldsymbol{G}-\boldsymbol{G}(0)\| = & \frac{1}{m}\|\sigma(\boldsymbol{X}\boldsymbol{W}^T)\sigma(\boldsymbol{X}\boldsymbol{W}^T)^T - \sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)\sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)^T\| \\
\leq & \frac{1}{m}\|\sigma(\boldsymbol{X}\boldsymbol{W}^T)\sigma(\boldsymbol{X}\boldsymbol{W}^T)^T - \sigma(\boldsymbol{X}\boldsymbol{W}^T)\sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)^T\| \\
& + \frac{1}{m}\|\sigma(\boldsymbol{X}\boldsymbol{W}^T)\sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)^T - \sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)\sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)^T\| \\
= & \frac{1}{m}\|\sigma(\boldsymbol{X}\boldsymbol{W}^T)\|\|\sigma(\boldsymbol{X}\boldsymbol{W}^T) - \sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)\| \\
& + \frac{1}{m}\|\sigma(\boldsymbol{X}\boldsymbol{W}^T) - \sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)\|\|\sigma(\boldsymbol{X}\boldsymbol{W}(0)^T)\| \\
\leq & \frac{1}{m}\|\boldsymbol{X}\|\|\boldsymbol{W}\|\|\boldsymbol{X}\|\|\boldsymbol{W}-\boldsymbol{W}(0)\| + \frac{1}{m}\|\boldsymbol{X}\|\|\boldsymbol{W}-\boldsymbol{W}(0)\|\|\boldsymbol{X}\|\|\boldsymbol{W}(0)\| \\
\leq & \frac{4c}{\sqrt{m}}\|\boldsymbol{X}\|^2\|\boldsymbol{W}-\boldsymbol{W}(0)\| \\
\leq & \frac{\lambda_0}{4}.
\end{aligned}$$

□

A.7 Proof of Lemma 3.8

Proof. It suffices to show the result hold for $\gamma = \min\{\gamma_0, \gamma_0/2c\}$, where $\gamma_0 = 1/2$. We prove by the induction. Suppose that for $0 \leq s \leq t$, the followings hold

(i) $\lambda_{\min}(\boldsymbol{G}(s)) \geq \frac{\lambda_0}{2}$,

(ii) $\|\boldsymbol{u}(s)\| \leq \frac{16c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0)-\boldsymbol{y}\|$,

(iii) $\|\boldsymbol{v}(s)\| \leq \frac{8c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0)-\boldsymbol{y}\|$,

(iv) $\|\boldsymbol{W}(s)\| \leq 2c\sqrt{m}$,

(v) $\|\boldsymbol{A}(s)\| \leq 2c\sqrt{m}$,

(vi) $\|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\|^2 \leq \exp\{-\lambda_0 s\}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2$,

Since $\lambda_{\min}(\boldsymbol{G}(s)) \geq \frac{\lambda_0}{2}$, we have

$$\begin{aligned}\frac{d}{dt}\|\hat{\boldsymbol{y}}(t) - \boldsymbol{y}\|^2 =& -2(\hat{\boldsymbol{y}}(t) - \boldsymbol{y})^T \boldsymbol{H}(t)(\hat{\boldsymbol{y}}(t) - \boldsymbol{y}) \\ \leq& -\lambda_0\|\hat{\boldsymbol{y}}(t) - \boldsymbol{y}\|^2\end{aligned}$$

Solving the ordinary differential equation yields

$$\|\hat{\boldsymbol{y}}(t) - \boldsymbol{y}\|^2 \leq \exp\{-\lambda_0 t\}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2.$$

By using the inductive hypothesis $\|\boldsymbol{W}(s)\| \leq 2c\sqrt{m}$, we have

$$\|\boldsymbol{\phi}_i(s)\| = \left\|\frac{1}{\sqrt{m}}\sigma(\boldsymbol{W}(s)\boldsymbol{x}_i)\right\| \leq \frac{1}{\sqrt{m}}\|\boldsymbol{W}(s)\|\|\boldsymbol{x}_i\| \leq 2c.$$

It follows from Lemma 3.2 with $\gamma_0 = 1/2$ that

$$\|\boldsymbol{z}_i^*(s)\| \leq 2\|\boldsymbol{\phi}_i(s)\| \leq 4c.$$

Note that

$$\begin{aligned}\|\nabla_{\boldsymbol{v}} L(s)\| \leq& \sum_{i=1}^n |\hat{y}_i(s) - y_i| \, \|\boldsymbol{\phi}_i(s)\| \\ \leq& 2c\sum_{i=1}^n |\hat{y}_i(s) - y_i| \\ \leq& 2c\sqrt{n}\|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\| \\ \leq& 2c\sqrt{n}\exp\{-\lambda_0 s/2\}\|\boldsymbol{y}(0) - \boldsymbol{y}\|\end{aligned}$$

and so

$$\begin{aligned}\|\boldsymbol{v}(t) - \boldsymbol{v}(0)\| \leq& \int_0^t \|\nabla_{\boldsymbol{v}} L(s)\| ds \\ \leq& 2c\sqrt{n}\|\boldsymbol{y}(0) - \boldsymbol{y}\| \int_0^t \exp\{-\lambda_0 s/2\} ds \\ \leq& \frac{4c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|.\end{aligned}$$

Since $\boldsymbol{v}_i(0)$ follows symmetric Bernoulli distribution with $\pm 1/\sqrt{m}$, then $\|\boldsymbol{v}(0)\|^2 = 1$ and we obtain

$$\|\boldsymbol{v}(t)\| \leq \|\boldsymbol{v}(t) - \boldsymbol{v}(0)\| + \|\boldsymbol{v}(0)\| \leq \frac{8c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|.$$

Note that

$$\begin{aligned}\|\nabla_{\boldsymbol{u}} L(s)\| \leq& \sum_{i=1}^n |\hat{y}_i(s) - y_i| \, \|\boldsymbol{z}_i^*\| \\ \leq& 4c\sqrt{n}\|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\| \\ \leq& 4c\sqrt{n}\exp\{-\lambda_0 s/2\}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|\end{aligned}$$

so that

$$\|\boldsymbol{u}(t) - \boldsymbol{u}(0)\| \leq \int_0^t \|\nabla_{\boldsymbol{u}} L(s)\| ds \leq \frac{8c\sqrt{n}}{\lambda_0}$$

Since $\boldsymbol{u}_i(0)$ follows symmetric Bernoulli distribution with $\pm 1/\sqrt{m}$, then $\|\boldsymbol{u}(0)\|^2 = 1$ and we obtain

$$\|\boldsymbol{u}(t)\| \leq \|\boldsymbol{u}(t) - \boldsymbol{u}(0)\| + \|\boldsymbol{u}(0)\| \leq \frac{16c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|$$

Note that

$$\begin{aligned}
\|\nabla_{\boldsymbol{W}} L(s)\| \leq & \sum_{i=1}^n \frac{1}{\sqrt{m}} \left|\hat{y}_i(s) - y_i\right| \|\boldsymbol{E}_i(s)\| \left(\|\boldsymbol{U}_i(s)^{-1}\boldsymbol{u}(s)\| + \|\boldsymbol{v}(s)\|\right) \|\boldsymbol{x}_i\| \\
\leq & \frac{64c\sqrt{n}}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \sum_{i=1}^n |\hat{y}_i(s) - y_i| \\
\leq & \frac{64cn}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\| \\
\leq & \frac{64cn}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \exp\{-\lambda_0 s/2\}
\end{aligned}$$

so that

$$\begin{aligned}
\|\boldsymbol{W}(t) - \boldsymbol{W}(0)\| \leq & \int_0^t \|\nabla_{\boldsymbol{W}} L(s)\| ds \\
\leq & \frac{128cn}{\lambda_0^2\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \\
\leq & \frac{\sqrt{m}\lambda_0}{16c\|\boldsymbol{X}\|^2} \\
\leq & R.
\end{aligned}$$

so that we obtain

$$\|\boldsymbol{W}(t)\| \leq \|\boldsymbol{W}(t) - \boldsymbol{W}(0)\| + \|\boldsymbol{W}(0)\| \leq 2c\sqrt{m},$$

provided $c > 0$ is chosen to be large enough, *i.e.*, $c \gtrsim \sqrt{\lambda_0}/\|\boldsymbol{X}\|$. Moreover, it follows from Lemma 3.7 that $\lambda_{\min}\{\boldsymbol{G}(t)\} \geq \frac{\lambda_0}{2}$.

Note that

$$\begin{aligned}
\|\nabla_{\boldsymbol{A}} L(s)\| \leq & \sum_{i=1}^n \frac{\gamma}{\sqrt{m}} \left|\hat{y}_i(s) - y_i\right| \|\boldsymbol{D}_i\| \|\boldsymbol{U}_i(s)^{-1}\| \|\boldsymbol{u}(s)\| \|\boldsymbol{z}_i^*\| \\
\leq & \frac{32c\sqrt{n}}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \sum_{i=1}^n |\hat{y}_i(s) - y_i| \\
\leq & \frac{32cn}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\| \\
\leq & \frac{32cn}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \exp\{-\lambda_0 s/2\},
\end{aligned}$$

so that

$$\begin{aligned}
\|\boldsymbol{A}(t) - \boldsymbol{A}(0)\| \leq & \int_0^t \|\nabla_{\boldsymbol{A}} L(s)\| ds \\
\leq & \frac{64cn}{\lambda_0^2\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2
\end{aligned}$$

Then

$$\|\boldsymbol{A}(t)\| \leq \|\boldsymbol{A}(t) - \boldsymbol{A}(0)\| + \|\boldsymbol{A}(0)\| \leq 2c\sqrt{m}.$$

□

A.8 DISCRETE TIME ANALYSIS

In this section, we prove the result for discrete time analysis or result for gradient descent. Assume $\|\boldsymbol{A}(0)\| \leq c\sqrt{m}$ and $\|\boldsymbol{W}(0)\| \leq c\sqrt{m}$. Further, we assume $\lambda_{\min}(\boldsymbol{G}(0)) \geq \frac{3}{4}\lambda_0$ and we assume $m = \Omega\left(\frac{c^2 n\|\boldsymbol{X}\|^2}{\lambda_0^3}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2\right)$ and choose $0 < \gamma \leq \min\{1/2, 1/4c\}$. Moreover, we assume the stepsize $\alpha = \mathcal{O}\left(\lambda_0/n^2\right)$. We make the inductive hypothesis as follows for all $0 \leq s \leq k$

(i) $\lambda_{\min}(\boldsymbol{G}(s)) \geq \frac{\lambda_0}{2}$,
(ii) $\|\boldsymbol{u}(s)\| \leq \frac{32c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|$,
(iii) $\|\boldsymbol{v}(s)\| \leq \frac{16c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|$,
(iv) $\|\boldsymbol{W}(s)\| \leq 2c\sqrt{m}$,
(v) $\|\boldsymbol{A}(s)\| \leq 2c\sqrt{m}$,
(vi) $\|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\|^2 \leq (1 - \alpha\lambda_0/2)^s\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2$.

Proof. By using the inductive hypothesis, we have for any $0 \leq s \leq k$

$$\|\boldsymbol{\phi}_i(s)\| = \|\frac{1}{\sqrt{m}}\sigma(\boldsymbol{W}(s)\boldsymbol{x}_i)\| \leq \frac{1}{\sqrt{m}}\|\boldsymbol{W}(s)\| \leq 2c$$

and

$$\|\boldsymbol{\Phi}(s)\| \leq \|\boldsymbol{\Phi}(s)\|_F = \left(\sum_{i=1}^n \|\boldsymbol{\phi}_i(s)\|^2\right)^{1/2} \leq 2c\sqrt{n}. \tag{38}$$

By using Lemma 3.2, we obtain the upper bound for the equilibrium point $\boldsymbol{z}_i(s)$ for any $0 \leq s \leq k$ as follows

$$\|\boldsymbol{z}_i(s)\| \leq \frac{1}{1-\gamma_0}\|\boldsymbol{\phi}_i(s)\| = 2\|\boldsymbol{\phi}_i(s)\| \leq 4c,$$

where the last inequality is because we choose $\gamma_0 = 1/2$, and

$$\|\boldsymbol{Z}(s)\| \leq \|\boldsymbol{Z}(s)\|_F = \left(\sum_{i=1}^n \|\boldsymbol{z}_i(s)\|^2\right)^{1/2} = 4c\sqrt{n}. \tag{39}$$

By using the upper bound of $\boldsymbol{\phi}_i(s)$, we obtain for any $0 \leq s \leq k$

$$\begin{aligned}
\|\nabla_{\boldsymbol{v}} L(s)\| \leq& \sum_{i=1}^n |\hat{y}_i(s) - y_i|\,\|\boldsymbol{\phi}_i(s)\| \\
\leq& 2c\sum_{i=1}^n |\hat{y}_i(s) - y_i| \\
\leq& 2c\sqrt{n}\|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\| \\
\leq& 2c\sqrt{n}(1 - \alpha\lambda_0/2)^{s/2}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|.
\end{aligned}$$

Let $\beta := \sqrt{1 - \alpha\lambda_0/2}$. Then the upper bound of $\|\nabla_{\boldsymbol{v}} L(s)\|$ can be written as

$$\|\nabla_{\boldsymbol{v}} L(s)\| \leq 2c\sqrt{n}\beta^s\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|, \tag{40}$$

and

$$\begin{aligned}
\|\boldsymbol{v}(k+1) - \boldsymbol{v}(0)\| \leq& \sum_{s=0}^k \|\boldsymbol{v}(s+1) - \boldsymbol{v}(s)\| = \alpha\sum_{s=0}^k \|\nabla_{\boldsymbol{v}} L(s)\| \\
\leq& \alpha \cdot 2c\sqrt{n}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \sum_{s=0}^k \beta^s \\
=& \frac{2(1-\beta^2)}{\lambda_0} \cdot 2c\sqrt{n}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|\frac{1-\beta^{k+1}}{1-\beta} \\
\leq& \frac{8c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|,
\end{aligned}$$

where the last inequality we use the facts $\beta < 1$. By triangle inequality, we obtain

$$\|\boldsymbol{v}(k+1)\| \leq \|\boldsymbol{v}(k+1) - \boldsymbol{v}(0)\| + \|\boldsymbol{v}(0)\| \leq \frac{16c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|,$$

which proves the result (iii). Similarly, we can upper bound the gradient of $\boldsymbol{u}$

$$\|\nabla_{\boldsymbol{u}} L(s)\| \leq \sum_{i=1}^{n} |\hat{y}_i(s) - y_i| \, \|\boldsymbol{z}_i\| \leq 4c\sqrt{n}\|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\| \leq 4c\sqrt{n}\beta^s\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \tag{41}$$

so that

$$\|\boldsymbol{u}(k+1) - \boldsymbol{u}(0)\| \leq \frac{16c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}} - \boldsymbol{y}\|,$$

and

$$\|\boldsymbol{u}(k)\| \leq \|\boldsymbol{u}(k) - \boldsymbol{u}(0)\| + \|\boldsymbol{u}(0)\| \leq \frac{32c\sqrt{n}}{\lambda_0}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|.$$

The result (ii) is also obtained.

By using the inductive hypothesis, we can upper bound the gradient of $\boldsymbol{W}$ as follows

$$\begin{aligned}
\|\nabla_{\boldsymbol{W}} L(s)\| \leq & \sum_{i=1}^{n} \frac{1}{\sqrt{m}} |\hat{y}_i(s) - y_i| \, \|\boldsymbol{E}_i(s)\| \left(\|\boldsymbol{U}_i(s)^{-1}\boldsymbol{u}(s)\| + \|\boldsymbol{v}(s)\|\right) \|\boldsymbol{x}_i\| \\
\leq & \frac{128c\sqrt{n}}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \sum_{i=1}^{n} |\hat{y}_i(s) - y_i| \\
\leq & \frac{128cn}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\| \\
\leq & \frac{128cn}{\lambda_0\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^s
\end{aligned} \tag{42}$$

so that

$$\begin{aligned}
\|\boldsymbol{W}(k+1) - \boldsymbol{W}(0)\| \leq & \alpha \sum_{s=0}^{k} \|\nabla_{\boldsymbol{W}} L(s)\| \\
\leq & \alpha \cdot \frac{128cn}{\lambda_0^2\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \sum_{s=0}^{k} \beta^s \\
\leq & \frac{512cn}{\lambda_0^2\sqrt{m}}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \\
\leq & \frac{\sqrt{m}\lambda_0}{16c\|\boldsymbol{X}\|^2} \\
\leq & R,
\end{aligned}$$

where the third inequality holds is because m is large, *i.e.*, $m = \Theta(\frac{c^2 n\|\boldsymbol{X}\|^2}{\lambda_0^3}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2)$. To simplify the notation, we assume

$$m = \frac{Cc^2 n\|\boldsymbol{X}\|^2}{\lambda_0^3}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \tag{43}$$

for some large number $C > 0$. Moreover, we obtain

$$\|\boldsymbol{W}(k+1)\| \leq \|\boldsymbol{W}(k+1) - \boldsymbol{W}(0)\| + \|\boldsymbol{W}(0)\| \leq 2c\sqrt{m},$$

provided $c > 0$ is chosen to be large enough, *i.e.*, $c \gtrsim \sqrt{\lambda_0}/\|\boldsymbol{X}\|$. Therefore, it follows from Lemma 3.7 that $\lambda_{\min}\{\boldsymbol{G}(k+1)\} \geq \frac{\lambda_0}{2}$. Thus, the results (i) and (iv) are established.

By using similar argument, we can upper bound the gradient of $\boldsymbol{A}$ as follows Note that

$$\begin{aligned}\|\nabla_{\boldsymbol{A}} L(s)\| \leq & \sum_{i=1}^{n} \frac{\gamma}{\sqrt{m}} \left|\hat{y}_i(s) - y_i\right| \|\boldsymbol{D}_i\| \|\boldsymbol{U}_i(s)^{-1}\| \|\boldsymbol{u}(s)\| \|\boldsymbol{z}_i^*\| \\ \leq & \frac{64c\sqrt{n}}{\lambda_0\sqrt{m}} \|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \sum_{i=1}^{n} \left|\hat{y}_i(s) - y_i\right| \\ \leq & \frac{64cn}{\lambda_0\sqrt{m}} \|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \|\hat{\boldsymbol{y}}(s) - \boldsymbol{y}\| \\ \leq & \frac{64cn}{\lambda_0\sqrt{m}} \|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^s,\end{aligned}$$

so that

$$\begin{aligned}\|\boldsymbol{A}(k+1) - \boldsymbol{A}(0)\| \leq & \alpha \sum_{s=0}^{k} \|\nabla_{\boldsymbol{A}} L(s)\| \\ \leq & \alpha \cdot \frac{64cn}{\lambda_0\sqrt{m}} \|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \sum_{s=0}^{k} \beta^s \\ \leq & \frac{256cn}{\lambda_0^2\sqrt{m}} \|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2\end{aligned}$$

Since $m \geq \frac{Cc^2 n\|\boldsymbol{X}\|^2}{\lambda_0^3}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2$ and $c, C > 0$ is large enough, we have

$$\|\boldsymbol{A}(k+1)\| \leq \|\boldsymbol{A}(k+1) - \boldsymbol{A}(0)\| + \|\boldsymbol{A}(0)\| \leq 2c\sqrt{m}.$$

Therefore, the result (v) is obtained and the equilibrium points $\boldsymbol{z}_i(k+1)$ exists for all $i \in [n]$.

To establish the result (vi), we need to derive the bounds between equilibrium points and feature vectors. Next, we will bound the difference between equilibrium points $\boldsymbol{z}_i(k+1)$ and $\boldsymbol{z}_i(k)$. For any $\ell \geq 1$, we have

$$\begin{aligned}\|\boldsymbol{z}_i^{\ell+1}(k+1) - \boldsymbol{z}_i^{\ell}(k)\| = & \|\sigma\left[\tilde{\gamma}\boldsymbol{A}(k+1)\boldsymbol{z}_i^{\ell}(k+1) + \boldsymbol{\phi}_i(k+1)\right] - \sigma\left[\tilde{\gamma}\boldsymbol{A}(k)\boldsymbol{z}_i^{\ell}(k) + \boldsymbol{\phi}_i(k)\right]\| \\ \leq & \|\tilde{\gamma}\boldsymbol{A}(k+1)\boldsymbol{z}_i^{\ell}(k+1) + \boldsymbol{\phi}_i(k+1) - \tilde{\gamma}\boldsymbol{A}(k)\boldsymbol{z}_i^{\ell}(k) - \boldsymbol{\phi}_i(k)\| \\ \leq & \tilde{\gamma}\|\boldsymbol{A}(k+1)\boldsymbol{z}_i^{\ell}(k+1) - \boldsymbol{A}(k)\boldsymbol{z}_i^{\ell}(k)\| + \|\boldsymbol{\phi}_i(k+1) - \boldsymbol{\phi}_i(k)\|,\end{aligned}$$

where the first term can be bounded as follows

$$\begin{aligned}& \tilde{\gamma}\|\boldsymbol{A}(k+1) - \boldsymbol{A}(k)\|\|\boldsymbol{z}_i^{\ell}(k+1)\| + \tilde{\gamma}\|\boldsymbol{A}(k)\|\|\boldsymbol{z}_i^{\ell}(k+1) - \boldsymbol{z}_i^{\ell}(k)\| \\ \leq & \tilde{\gamma}\alpha\|\nabla_{\boldsymbol{A}} L(k)\|(4c) + \tilde{\gamma}\|\boldsymbol{A}(k)\|\|\boldsymbol{z}_i^{\ell}(k+1) - \boldsymbol{z}_i^{\ell}(k)\| \\ \leq & \frac{64\alpha cn}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2\beta^k + (1/2)\|\boldsymbol{z}_i^{\ell}(k+1) - \boldsymbol{z}_i^{\ell}(k)\|,\end{aligned}$$

and the second term is bounded as follows

$$\begin{aligned}\frac{1}{\sqrt{m}}\|\sigma[\boldsymbol{W}(k+1)\boldsymbol{x}_i] - \sigma[\boldsymbol{W}(k)\boldsymbol{x}_i]\| \leq & \frac{1}{\sqrt{m}}\|\boldsymbol{W}(k+1) - \boldsymbol{W}(k)\|\|\boldsymbol{x}_i\| \\ \leq & \frac{\alpha}{\sqrt{m}}\|\nabla_{\boldsymbol{W}} L(k)\| \\ \leq & \frac{128\alpha cn}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^k.\end{aligned}$$

Thus, we obtain

$$\begin{aligned}\|\boldsymbol{z}_i^{\ell+1}(k+1) - \boldsymbol{z}_i^{\ell}(k)\| \leq & (1/2)\|\boldsymbol{z}_i^{\ell}(k+1) - \boldsymbol{z}_i^{\ell}(k)\| + \frac{256\alpha cn}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^k \\ \leq & (1/2)^{\ell}\|\boldsymbol{z}_i^{1}(k+1) - \boldsymbol{z}_i^{1}(k)\| + \frac{256\alpha cn}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^k \cdot \sum_{j=0}^{\infty} 2^{-j} \\ \leq & (1/2)^{\ell}\|\boldsymbol{z}_i^{1}(k+1) - \boldsymbol{z}_i^{1}(k)\| + \frac{512\alpha cn}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^k.\end{aligned}$$

Let $\ell \to \infty$, then we obtain

$$\|\boldsymbol{z}_i(k+1) - \boldsymbol{z}_i(k)\| \le \frac{512\alpha cn}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^k.$$

By using the Cauchy-Schwartz inequality, we have

$$\|\boldsymbol{Z}(k+1) - \boldsymbol{Z}(k)\| \le \|\boldsymbol{Z}(k+1) - \boldsymbol{Z}(k)\|_F \le \frac{512\alpha cn^{3/2}}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^k. \tag{44}$$

In addition, we will also bound the difference in $\boldsymbol{\phi}_i(k+1)$ and $\boldsymbol{\phi}_i(k)$. Note that

$$\|\boldsymbol{\phi}_i(k+1) - \boldsymbol{\phi}_i(k)\| = \frac{1}{\sqrt{m}}\|\sigma[\boldsymbol{W}(k+1)\boldsymbol{x}_i] - \sigma[\boldsymbol{W}(k)\boldsymbol{x}_i]\| \le \frac{128\alpha cn}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^k,$$

so that

$$\|\boldsymbol{\Phi}(k+1) - \boldsymbol{\Phi}(k)\| \le \|\boldsymbol{\Phi}(k+1) - \boldsymbol{\Phi}(k)\|_F \le \frac{128\alpha cn^{3/2}}{\lambda_0 m}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \cdot \beta^k \tag{45}$$

Now, we are ready to establish the result (vi). Note that

$$\begin{aligned}\|\hat{\boldsymbol{y}}(k+1) - \boldsymbol{y}\|^2 =& \|\hat{\boldsymbol{y}}(k+1) - \hat{\boldsymbol{y}}(k) + \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\|^2 \\ =& \|\hat{\boldsymbol{y}}(k+1) - \hat{\boldsymbol{y}}(k)\|^2 + 2\left\langle \hat{\boldsymbol{y}}(k+1) - \hat{\boldsymbol{y}}(k), \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\right\rangle + \|\hat{\boldsymbol{y}}(k) - \boldsymbol{y}\|^2.\end{aligned}$$

In the rest of this proof, we will bound each term in the above inequality. By the prediction rule of $\hat{\boldsymbol{y}}$, we can bound the difference between $\hat{\boldsymbol{y}}(k+1)$ and $\hat{\boldsymbol{y}}(k)$ as follows

$$\begin{aligned}\|\hat{\boldsymbol{y}}(k+1) - \hat{\boldsymbol{y}}(k)\| =& \|\boldsymbol{Z}(k+1)\boldsymbol{u}(k+1) + \boldsymbol{\Phi}(k+1)\boldsymbol{v}(k+1) - \boldsymbol{Z}(k)\boldsymbol{u}(k) - \boldsymbol{\Phi}(k)\boldsymbol{v}(k)\| \\ \le& \|\boldsymbol{Z}(k+1)\boldsymbol{u}(k+1) - \boldsymbol{Z}(k)\boldsymbol{u}(k)\| + \|\boldsymbol{\Phi}(k+1)\boldsymbol{v}(k+1) - \boldsymbol{\Phi}(k)\boldsymbol{v}(k)\|,\end{aligned}$$

where the first term can be bounded as follows by using equation 39, 41, 43, 44, hypothesis (ii), and a large constant $C_0 > 0$

$$\begin{aligned}&\|\boldsymbol{Z}(k+1)\|\|\boldsymbol{u}(k+1) - \boldsymbol{u}(k)\| + \|\boldsymbol{Z}(k+1) - \boldsymbol{Z}(k)\|\|\boldsymbol{u}(k)\| \\ =& \alpha\|\boldsymbol{Z}(k+1)\|\|\nabla_{\boldsymbol{u}} L(k)\| + \|\boldsymbol{Z}(k+1) - \boldsymbol{Z}(k)\|\|\boldsymbol{u}(k)\| \\ \le& \alpha C_0 c^2 n\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \beta^k,\end{aligned}$$

and the second term is bounded as follows by using equation 38, 40, 45, 43, hypothesis (iii), and a large constant $C_0 > 0$

$$\begin{aligned}&\|\boldsymbol{\Phi}(k+1)\|\|\boldsymbol{v}(k+1) - \boldsymbol{v}(k)\| + \|\boldsymbol{\Phi}(k+1) - \boldsymbol{\Phi}(k)\|\|\boldsymbol{v}(k)\| \\ =& \alpha\|\boldsymbol{\Phi}(k+1)\|\|\nabla_{\boldsymbol{v}} L(k)\| + \|\boldsymbol{\Phi}(k+1) - \boldsymbol{\Phi}(k)\|\|\boldsymbol{v}(k)\| \\ \le& \alpha C_0 c^2 n\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \beta^k.\end{aligned}$$

Therefore, we have

$$\|\hat{\boldsymbol{y}}(k+1) - \hat{\boldsymbol{y}}(k)\| \le \alpha C_0 c^2 n\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \beta^k, \tag{46}$$

where the scalar 2 is absorbed in C_0 and the constant C_0 is difference from C.

Let $\boldsymbol{g} := \boldsymbol{Z}(k)\boldsymbol{u}(k+1) + \boldsymbol{\Phi}(k)\boldsymbol{v}(k+1)$. Then we have

$$\left\langle \hat{\boldsymbol{y}}(k+1) - \hat{\boldsymbol{y}}(k), \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\right\rangle = \left\langle \hat{\boldsymbol{y}}(k+1) - \boldsymbol{g}, \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\right\rangle + \left\langle \boldsymbol{g} - \hat{\boldsymbol{y}}(k), \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\right\rangle.$$

Let us bound each term individually. By using Cauchy-Schwartz inequality, we have

$$\begin{aligned}&\left\langle \hat{\boldsymbol{y}}(k+1) - \boldsymbol{g}, \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\right\rangle \\ =& \left\langle (\boldsymbol{Z}(k+1) - \boldsymbol{Z}(k))\boldsymbol{u}(k+1), \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\right\rangle + \left\langle (\boldsymbol{\Phi}(k+1) - \boldsymbol{\Phi}(k))\boldsymbol{v}(k+1), \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\right\rangle \\ \le& \left(\|\boldsymbol{Z}(k+1) - \boldsymbol{Z}(k)\|\|\boldsymbol{u}(k+1)\| + \|\boldsymbol{\Phi}(k+1) - \boldsymbol{\Phi}(k)\|\|\boldsymbol{v}(k+1)\|\right)\|\hat{\boldsymbol{y}}(k) - \boldsymbol{y}\| \\ \le& \alpha C_0 c^2 n\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\| \cdot \beta^k\|\hat{\boldsymbol{y}}(k) - \boldsymbol{y}\|, \quad \text{by equation 39, 41, 43, 44} \\ \le& \alpha C_0 c^2 n \cdot \beta^{2k}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2.\end{aligned} \tag{47}$$

By using $\nabla_{\boldsymbol{u}} L(k) = \boldsymbol{Z}(k)^T(\hat{\boldsymbol{y}}(k) - \boldsymbol{y})$, $\nabla_{\boldsymbol{v}} L(k) = \boldsymbol{\Phi}(k)^T(\hat{\boldsymbol{y}}(k) - \boldsymbol{y})$ and $\lambda_{\min}(\boldsymbol{G}(k)) \geq \lambda_0/2$, we get

$$\begin{aligned}\langle \boldsymbol{g} - \hat{\boldsymbol{y}}(k), \hat{\boldsymbol{y}}(k) - \boldsymbol{y}\rangle &= -\alpha(\hat{\boldsymbol{y}}(k) - \boldsymbol{y})^T \left[\boldsymbol{Z}(k)\boldsymbol{Z}(k)^T + \boldsymbol{\Phi}(k)\boldsymbol{\Phi}(k)^T\right](\hat{\boldsymbol{y}}(k) - \boldsymbol{y}) \\ &\leq -\frac{\alpha\lambda_0}{2}\|\hat{\boldsymbol{y}}(k) - \boldsymbol{y}\|^2. \end{aligned} \tag{48}$$

By combining the inequalities equation 46, 47, 48, we obtain

$$\begin{aligned}\|\hat{\boldsymbol{y}}(k+1) - \boldsymbol{y}\|^2 &\leq \left(1 - \alpha\left[\lambda_0 - C_0 c^2 n - \alpha C_0^2 c^4 n^2\right]\right)\beta^{2k}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \\ &\leq \left(1 - \frac{\alpha\lambda_0}{2}\right)\beta^{2k}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2 \\ &= \left(1 - \frac{\alpha\lambda_0}{2}\right)^{k+1}\|\hat{\boldsymbol{y}}(0) - \boldsymbol{y}\|^2, \end{aligned}$$

where the second inequality is due to $\alpha = \mathcal{O}\left(\frac{\lambda_0}{n^2}\right)$. This proves the result (vi) and complete the whole proof.

$\square$