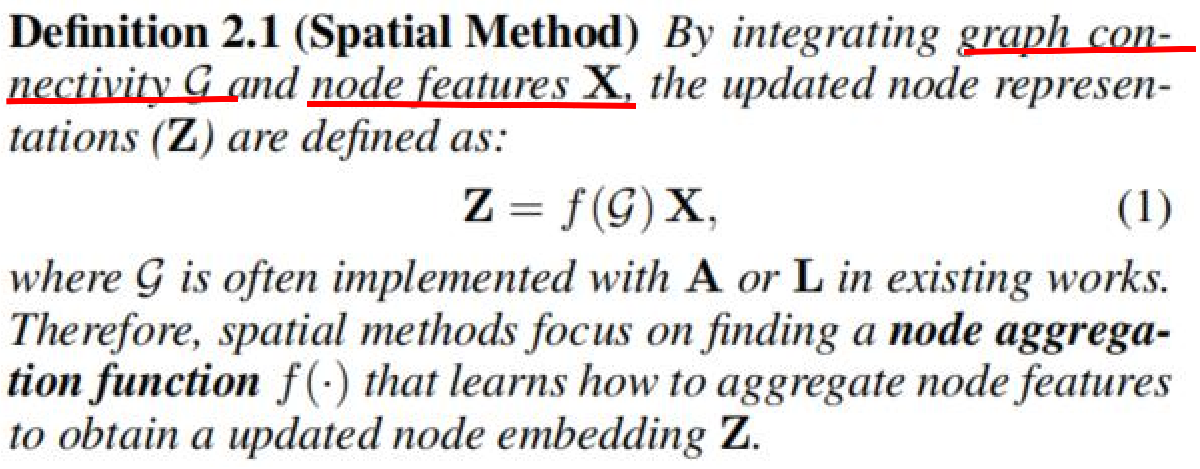
REPRESENTATION OF LEARNING PROCESS



Definition 2.1 (Spatial Method) *By integrating graph connectivity $\mathcal{G}$ and node features $\mathbf{X}$, the updated node representations ($\mathbf{Z}$) are defined as:*

$$\mathbf{Z} = f(\mathcal{G})\,\mathbf{X}, \qquad (1)$$

*where $\mathcal{G}$ is often implemented with $\mathbf{A}$ or $\mathbf{L}$ in existing works. Therefore, spatial methods focus on finding a **node aggregation function** $f(\cdot)$ that learns how to aggregate node features to obtain a updated node embedding $\mathbf{Z}$.*

concatenate the node’ s current representation

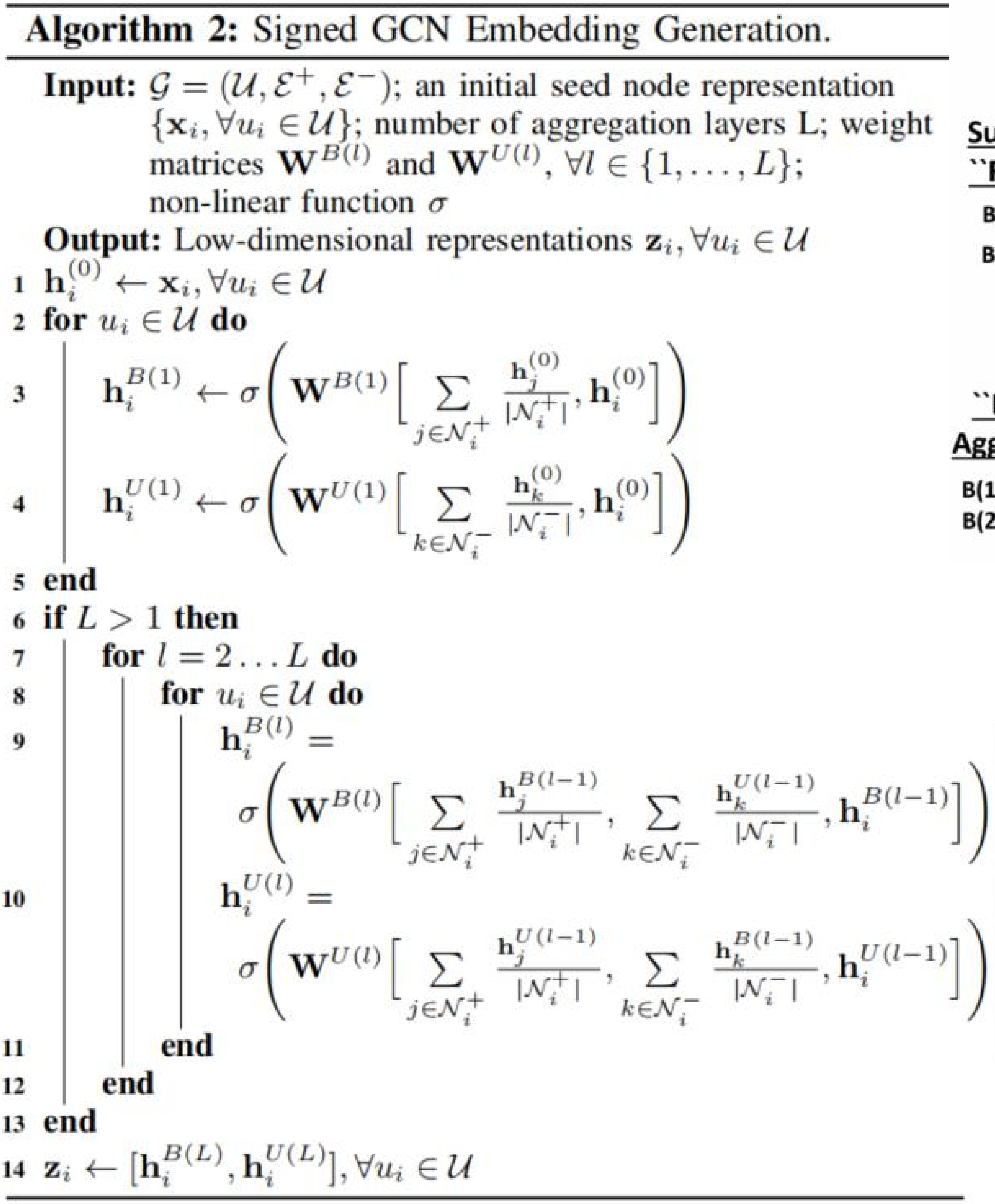
$$\mathrm{z}(v_i) = A(v_i)H(v_i) + \sum_{j\in \mathrm{N}(i)} B(u_j)H(u_j)$$

aggregate the neighboring feature vectors

$$Z = \left(\mathrm{a}I + b\widetilde{A}\right)X \qquad \widetilde{A}\ \text{denotes the normalized A}$$

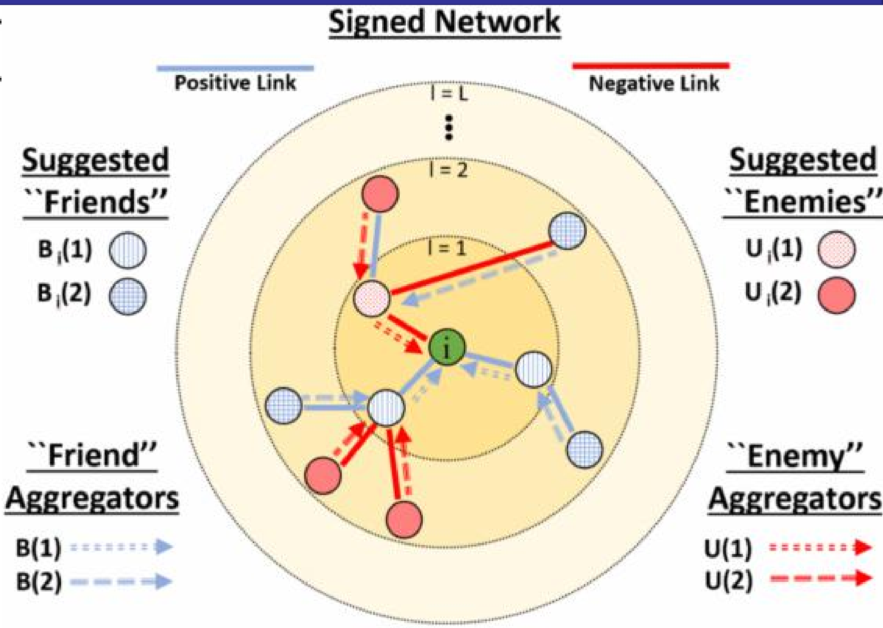
Bridging the Gap between Spatial and Spectral Domains:A Survey on Graph Neural Networks/2021

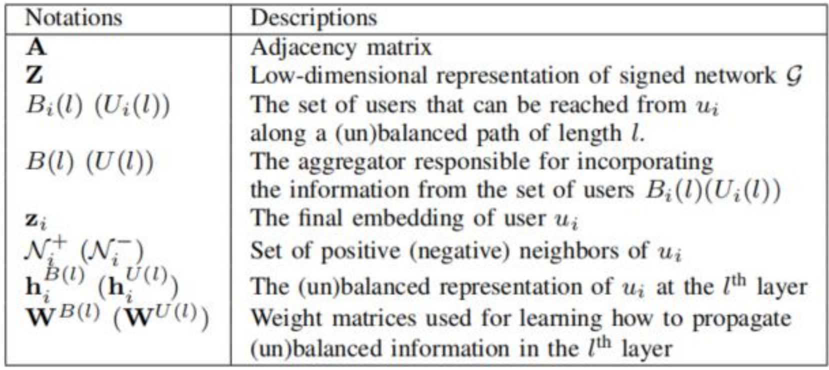
SGNN



Algorithm 2: Signed GCN Embedding Generation.

```
Input: $\mathcal{G} = (\mathcal{U}, \mathcal{E}^+, \mathcal{E}^-)$; an initial seed node representation
       $\{\mathbf{x}_i, \forall u_i \in \mathcal{U}\}$; number of aggregation layers L; weight
       matrices $\mathbf{W}^{B(l)}$ and $\mathbf{W}^{U(l)}$, $\forall l \in \{1, \dots, L\}$;
       non-linear function $\sigma$
Output: Low-dimensional representations $\mathbf{z}_i, \forall u_i \in \mathcal{U}$
$\mathbf{h}_i^{(0)} \leftarrow \mathbf{x}_i, \forall u_i \in \mathcal{U}$
for $u_i \in \mathcal{U}$ do
    $\mathbf{h}_i^{B(1)} \leftarrow \sigma\left(\mathbf{W}^{B(1)}\left[\sum_{j \in \mathcal{N}_i^+} \frac{\mathbf{h}_j^{(0)}}{|\mathcal{N}_i^+|}, \mathbf{h}_i^{(0)}\right]\right)$
    $\mathbf{h}_i^{U(1)} \leftarrow \sigma\left(\mathbf{W}^{U(1)}\left[\sum_{k \in \mathcal{N}_i^-} \frac{\mathbf{h}_k^{(0)}}{|\mathcal{N}_i^-|}, \mathbf{h}_i^{(0)}\right]\right)$
end
if $L > 1$ then
    for $l = 2 \dots L$ do
        for $u_i \in \mathcal{U}$ do
            $\mathbf{h}_i^{B(l)} =$
              $\sigma\left(\mathbf{W}^{B(l)}\left[\sum_{j \in \mathcal{N}_i^+} \frac{\mathbf{h}_j^{B(l-1)}}{|\mathcal{N}_i^+|}, \sum_{k \in \mathcal{N}_i^-} \frac{\mathbf{h}_k^{U(l-1)}}{|\mathcal{N}_i^-|}, \mathbf{h}_i^{B(l-1)}\right]\right)$
            $\mathbf{h}_i^{U(l)} =$
              $\sigma\left(\mathbf{W}^{U(l)}\left[\sum_{j \in \mathcal{N}_i^+} \frac{\mathbf{h}_j^{U(l-1)}}{|\mathcal{N}_i^+|}, \sum_{k \in \mathcal{N}_i^-} \frac{\mathbf{h}_k^{B(l-1)}}{|\mathcal{N}_i^-|}, \mathbf{h}_i^{U(l-1)}\right]\right)$
        end
    end
end
$\mathbf{z}_i \leftarrow [\mathbf{h}_i^{B(L)}, \mathbf{h}_i^{U(L)}], \forall u_i \in \mathcal{U}$
```





Notations	Descriptions
$\mathbf{A}$	Adjacency matrix
$\mathbf{Z}$	Low-dimensional representation of signed network $\mathcal{G}$
$B_i(l)$ $(U_i(l))$	The set of users that can be reached from u_i along a (un)balanced path of length l.
$B(l)$ $(U(l))$	The aggregator responsible for incorporating the information from the set of users $B_i(l)(U_i(l))$
$\mathbf{z}_i$	The final embedding of user u_i
$\mathcal{N}_i^+$ $(\mathcal{N}_i^-)$	Set of positive (negative) neighbors of u_i
$\mathbf{h}_i^{B(l)}$ $(\mathbf{h}_i^{U(l)})$	The (un)balanced representation of u_i at the l^{th} layer
$\mathbf{W}^{B(l)}$ $(\mathbf{W}^{U(l)})$	Weight matrices used for learning how to propagate (un)balanced information in the l^{th} layer

SGNN

The first term incorporates an additional layer for performing a weighted multinomial logistic regression (MLG) classifier. Here we wish to classify whether a pair of node embeddings are from users with a positive, negative, or no link between them.

$$
\begin{aligned}
&\mathcal{L}(\theta^{W}, \theta^{MLG}) = \\
&-\frac{1}{\mathcal{M}} \sum_{(u_i, u_j, s) \in \mathcal{M}} \omega_s \log \frac{\exp\left([\mathbf{z}_i, \mathbf{z}_j]\theta_s^{MLG}\right)}{\sum_{q \in \{+,-,?\}} \exp\left([\mathbf{z}_i, \mathbf{z}_j]\theta_q^{MLG}\right)} \\
&+ \lambda \Bigg[\frac{1}{|\mathcal{M}_{(+,?)}|} \sum_{\substack{(u_i, u_j, u_k) \\ \in \mathcal{M}_{(+,?)}}} \max\left(0, (||\mathbf{z}_i - \mathbf{z}_j||_2^2 - ||\mathbf{z}_i - \mathbf{z}_k||_2^2)\right) \\
&+ \frac{1}{|\mathcal{M}_{(-,?)}|} \sum_{\substack{(u_i, u_j, u_k) \\ \in \mathcal{M}_{(-,?)}}} \max\left(0, (||\mathbf{z}_i - \mathbf{z}_k||_2^2 - ||\mathbf{z}_i - \mathbf{z}_j||_2^2)\right) \Bigg] \\
&+ Reg(\theta^{W}, \theta^{MLG}) \qquad (7)
\end{aligned}
$$

SINGLE LAYER/2-LAYERS

$$\left|E_{SGD}\left[l\left(A_S, y\right) - l\left(A_{S^i}, y\right)\right]\right| \leq \alpha_l g_\lambda E_{SGD}\left[\left|\Delta\theta\right|\right]$$

↓

$$\mathbf{E}_{\mathrm{SGD}}\left[\left|\Delta\theta_{t+1}\right|\right] \leq \left(1 - \frac{1}{m}\right)\mathbf{E}_{\mathrm{SGD}}\left[\left|\left(\theta_{S,t} - \eta\nabla\ell\left(f(\mathbf{x}, \theta_{S,t}), y\right)\right) - \left(\theta_{S^i,t} - \eta\nabla\ell\left(f(\mathbf{x}, \theta_{S^i,t}), y\right)\right)\right|\right] + \left(\frac{1}{m}\right)\mathbf{E}_{\mathrm{SGD}}\left[\left|\left(\theta_{S,t} - \eta\nabla\ell\left(f(\mathbf{x}', \theta_{S,t}), y'\right)\right) - \left(\theta_{S^i,t} - \eta\nabla\ell\left(f(\mathbf{x}'', \theta_{S^i,t}), y''\right)\right)\right|\right]$$

↓

$$\mathbf{E}_{\mathrm{SGD}}\left[\left|\Delta\theta_T\right|\right] \leq \frac{2\eta\nu_\ell\alpha_\sigma \mathrm{g}_\lambda}{m}\sum_{t=1}^{T}\left(1 + \eta\nu_\ell\nu_\sigma \mathrm{g}_\lambda^2\right)^{t-1}$$

↓

$$\mathbf{E}_{\mathrm{SGD}}[R(A_S) - R_{emp}(A_S)] \leq \frac{1}{m}\mathcal{O}\left((\lambda_G^{\max})^{2T}\right) + \left(\mathcal{O}\left((\lambda_G^{\max})^{2T}\right) + M\right)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

$$\left|\mathbf{E}_{sgd}[\ell(\mathcal{A}^S, h) - \ell(\mathcal{A}^S, h)]\right|$$

↓

$$\mathbf{E}_{sgd}[\|\Delta\boldsymbol{W}_1\| + \|\Delta\boldsymbol{W}_2\|]$$

↓

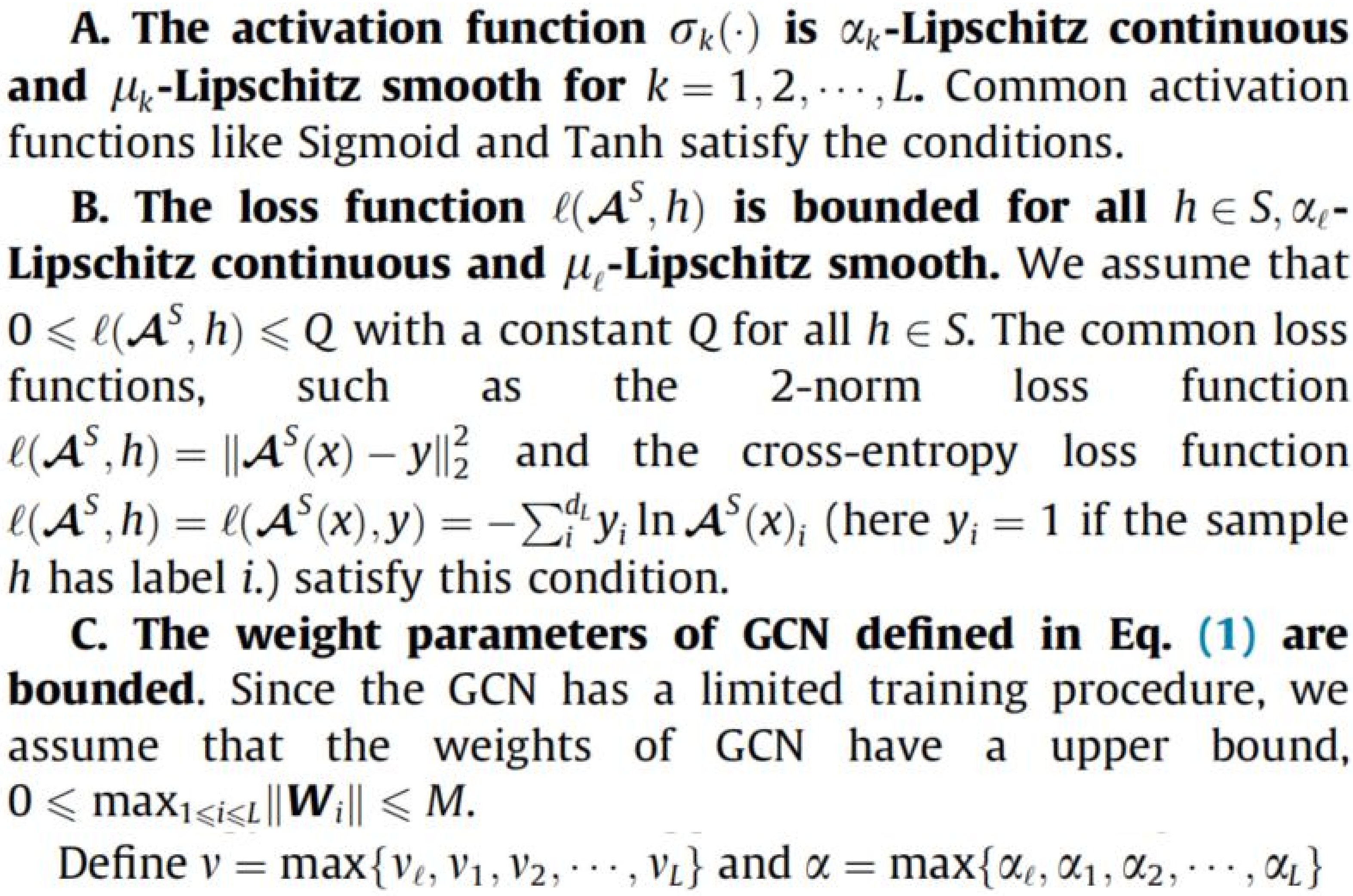
$$\begin{aligned}\mathbf{E}_{sgd}\left[\|\Delta\boldsymbol{W}_{j,t+1}\|\right] \leqslant \left(1 - \frac{1}{m}\right)\mathbf{E}_{sgd}\Big[\Big\| & \left(\boldsymbol{W}_{j,t}^{S} - \eta\nabla_{\boldsymbol{w}_{j,t}}\ell\left(\mathcal{A}_t^S(x_i), y_i\right)\right) \\ & - \left(\boldsymbol{W}_{j,t}^{S^i} - \eta\nabla_{\boldsymbol{w}_{j,t}}\ell\left(\mathcal{A}_t^{S^i}(x_i), y_i\right)\right)\Big\|\Big] \\ & + \left(\frac{1}{m}\right)\mathbf{E}_{sgd}\Big[\Big\|\left(\boldsymbol{W}_t^{S} - \eta\nabla_{\boldsymbol{w}_{j,t}}\ell\left(\mathcal{A}_t^S(x_i), y_i\right)\right) \\ & - \left(\boldsymbol{W}_{j,t}^{S^i} - \eta\nabla_{\boldsymbol{w}_{j,t}}\ell\left(\mathcal{A}_t^{S^i}(x_i'), y_i'\right)\right)\Big\|\Big]\end{aligned}$$

↓

$$\begin{aligned}\mathbf{E}_{sgd}[R(\mathcal{A}^S) - R_{emp}(\mathcal{A}^S)] \leqslant & \frac{1}{m}\mathcal{O}((K_1 + K_2)^T) \\ & + \left(\mathcal{O}((K_1 + K_2)^T) + Q\right)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}, \qquad (8)\end{aligned}$$

where $K_1 = \nu^3\alpha^2\|f(\mathbf{A})\|^2\|f(\mathbf{A})\mathbf{X}\|^3M^3$, *and* $K_2 = \nu^2\alpha^2\|f(\mathbf{A})\|^2\|f(\mathbf{A})\mathbf{X}\|^2M^2$. *Q is the upper bound of the loss function defined in assumption B, and M is the upper bound of the weights defined in assumption C.*

Assumptions



A. The activation function $\sigma_k(\cdot)$ is α_k-Lipschitz continuous and μ_k-Lipschitz smooth for $k = 1, 2, \cdots, L$. Common activation functions like Sigmoid and Tanh satisfy the conditions.

B. The loss function $\ell(\mathcal{A}^S, h)$ is bounded for all $h \in S, \alpha_\ell$-Lipschitz continuous and μ_ℓ-Lipschitz smooth. We assume that $0 \leqslant \ell(\mathcal{A}^S, h) \leqslant Q$ with a constant Q for all $h \in S$. The common loss functions, such as the 2-norm loss function $\ell(\mathcal{A}^S, h) = \|\mathcal{A}^S(x) - y\|_2^2$ and the cross-entropy loss function $\ell(\mathcal{A}^S, h) = \ell(\mathcal{A}^S(x), y) = -\sum_i^{d_L} y_i \ln \mathcal{A}^S(x)_i$ (here $y_i = 1$ if the sample h has label i.) satisfy this condition.

C. The weight parameters of GCN defined in Eq. (1) are bounded. Since the GCN has a limited training procedure, we assume that the weights of GCN have a upper bound, $0 \leqslant \max_{1 \leqslant i \leqslant L} \|\boldsymbol{W}_i\| \leqslant M$.

Define $v = \max\{v_\ell, v_1, v_2, \cdots, v_L\}$ and $\alpha = \max\{\alpha_\ell, \alpha_1, \alpha_2, \cdots, \alpha_L\}$

Lemma 1

$$|\mathbf{E}_{sgd}[\ell(\mathcal{A}^{S},h)-\ell(\mathcal{A}^{S},h)]|\leqslant\alpha_\ell\mathbf{E}_{sgd}[\|\mathbf{Z}_2(\mathbf{X},\boldsymbol{W}^{S})-\mathbf{Z}_2(\mathbf{X},\boldsymbol{W}^{S^i})\|]$$

$$\left.\begin{aligned}&=\alpha_\ell\mathbf{E}_{sgd}\left[\|\sigma_2\left(f(\boldsymbol{A})\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S})\boldsymbol{W}_2^{S}\right)-\sigma_2\left(f(\boldsymbol{A})\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S^i})\boldsymbol{W}_2^{S^i}\right)\|\right]\\&\leqslant\alpha_\ell\alpha_2\mathbf{E}_{sgd}\left[\|f(\boldsymbol{A})\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S})\boldsymbol{W}_2^{S}-f(\boldsymbol{A})\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S^i})\boldsymbol{W}_2^{S^i}\|\right]\\&\leqslant\alpha_\ell\alpha_2\|f(\boldsymbol{A})\|\mathbf{E}_{sgd}\left[\|\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S})\boldsymbol{W}_2^{S}-\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S^i})\boldsymbol{W}_2^{S^i}\|\right]\end{aligned}\right\}\text{Lipschitz condition}$$

$$\begin{aligned}&\leqslant\alpha_\ell\alpha_2\|f(\boldsymbol{A})\|\mathbf{E}_{sgd}\Big[\|\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S})\boldsymbol{W}_2^{S}-\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S})\boldsymbol{W}_2^{S^i}\\&+\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S})\boldsymbol{W}_2^{S^i}-\sigma_1(f(\boldsymbol{A})\boldsymbol{X}\boldsymbol{W}_1^{S^i})\boldsymbol{W}_2^{S^i}\|\Big]\quad\text{Separation}\end{aligned}$$

$$\begin{aligned}&\leqslant\alpha_\ell\alpha_2\|f(\boldsymbol{A})\|\mathbf{E}_{sgd}\left[|\alpha_1\|f(\boldsymbol{A})\boldsymbol{X}\|\|\boldsymbol{W}_1^{S}\|\|\Delta\boldsymbol{W}_2\|+\alpha_1\|f(\boldsymbol{A})\boldsymbol{X}\|\|\Delta\boldsymbol{W}_1\|\|\boldsymbol{W}_2^{S^i}\||\right]\\&=\alpha_\ell\alpha_2\|f(\boldsymbol{A})\|\alpha_1\|f(\boldsymbol{A})\boldsymbol{X}\|\mathbf{E}_{sgd}\left[|\|\boldsymbol{W}_1^{S}\|\|\Delta\boldsymbol{W}_2\|+\|\Delta\boldsymbol{W}_1\|\|\boldsymbol{W}_2^{S^i}\||\right]\\&\leqslant 2M\alpha^3\|f(\boldsymbol{A})\|\|f(\boldsymbol{A})\boldsymbol{X}\|\underline{\mathbf{E}_{sgd}[\|\Delta\boldsymbol{W}_1\|+\|\Delta\boldsymbol{W}_2\|]}\end{aligned}$$

$$0\leqslant\max_{1\leqslant i\leqslant L}\|\boldsymbol{W}_i\|\leqslant M.$$

$$\nu=\max\{\nu_\ell,\nu_1,\nu_2,\cdots,\nu_L\}$$

$$\alpha=\max\{\alpha_\ell,\alpha_1,\alpha_2,\cdots,\alpha_L\}$$

Step 2(Lemma 2/3/4)

Since only one sample is different in S and Si, two scenarios are needed to consider:

$$\|\Delta \boldsymbol{W}_{j,t+1}\| \leqslant \|\Delta \boldsymbol{W}_{j,t}\| + \eta \|\nabla_{\mathbf{w}_j} \ell\Big(\mathcal{A}_t^S(x_i), y_i\Big) - \nabla_{\mathbf{w}_j} \ell\Big(\mathcal{A}_t^{S^i}(x_i), y_i\Big)\|.$$

$$\|\Delta \boldsymbol{W}_{j,t+1}\| \leqslant \|\Delta \boldsymbol{W}_{j,t}\| + \eta \|\nabla_{\mathbf{w}_j} \ell\Big(\mathcal{A}_t^S(x_i), y_i\Big) - \nabla_{\mathbf{w}_j} \ell\Big(\mathcal{A}_t^{S^i}(x_i'), y_i'\Big)\|.$$

$$E_{SGD} = \left(1 - \frac{1}{\mathrm{m}}\right) E_1 + \frac{1}{m} E_2$$

For condition 1 and 2,we need to treat parameters of two layers respectively

$$\begin{aligned}
&= \mathbf{E}_{sgd}\Big[\|\Delta \boldsymbol{W}_{j,t}\|\Big] + \\
&(1 - \tfrac{1}{m})\eta \mathbf{E}_{sgd}\Big[\|\nabla_{\mathbf{w}_{j,t}} \ell\Big(\mathcal{A}_t^S(x_i), y_i\Big) - \nabla_{\mathbf{w}_{j,t}} \ell\Big(\mathcal{A}_t^{S^i}(x_i), y_i\Big)\|\Big] + \\
&(\tfrac{1}{m})\eta \mathbf{E}_{sgd}\Big[\|\Big(\nabla_{\mathbf{w}_{j,t}} \ell\Big(\mathcal{A}_t^S(x_i), y_i\Big) - \nabla_{\mathbf{w}_{j,t}} \ell\Big(\mathcal{A}_t^{S^i}(x_i'), y_i'\Big)\|\Big].
\end{aligned}$$

Lemma 3

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{w}_2}\ell\big(\mathcal{A}_t^{S}(x_i),y_i\big)-\nabla_{\boldsymbol{w}_2}\ell\big(\mathcal{A}_t^{S^i}(x_i),y_i\big)\|\\
&\left.\begin{aligned}
&\leqslant \nu_\ell\|\nabla_{\boldsymbol{w}_2}\boldsymbol{Z}_2(\boldsymbol{X},\boldsymbol{W}_t^S)-\boldsymbol{Z}_2(\boldsymbol{X},\boldsymbol{W}_t^{S^i})\|\\
&\leqslant \nu_\ell\|\boldsymbol{Z}_2^{S'}f(\boldsymbol{A})\boldsymbol{Z}_1^{S}-\boldsymbol{Z}_2^{S^{i\prime}}f(\boldsymbol{A})\boldsymbol{Z}_1^{S^i}\|\\
&\leqslant \nu_\ell\|\boldsymbol{Z}_2^{S'}f(\boldsymbol{A})\boldsymbol{Z}_1^{S}-\boldsymbol{Z}_2^{S'}f(\boldsymbol{A})\boldsymbol{Z}_1^{S^i}+\boldsymbol{Z}_2^{S'}f(\boldsymbol{A})\boldsymbol{Z}_1^{S^i}-\boldsymbol{Z}_2^{S^{i\prime}}f(\boldsymbol{A})\boldsymbol{Z}_1^{S^i}\|\\
&\leqslant \nu_\ell\nu_2\Big[\|(f(\boldsymbol{A})\alpha_1\|f(\boldsymbol{A})\boldsymbol{X}\|\|\boldsymbol{W}_1^S\|\|\boldsymbol{W}_2^S\|)\|f(\boldsymbol{A})\|\alpha_1\|f(\boldsymbol{A})\boldsymbol{X}\|\|\Delta\boldsymbol{W}_1\|)\|
\end{aligned}\right\}\text{Lipschitz condition}\\
&+(f(\boldsymbol{A})\alpha_1(\|f(\boldsymbol{A})\boldsymbol{X}\|\|\boldsymbol{W}_1^S\|\|\Delta\boldsymbol{W}_2\|+\|\Delta\boldsymbol{W}_1\|\|\boldsymbol{W}_2^{S^i}\|)f(\boldsymbol{A})\alpha_1\|f(\boldsymbol{A})\boldsymbol{X}\|\|\boldsymbol{W}_1^S\|\|\Big]\\
&\leqslant \nu_\ell\nu_2\alpha_1^2\|f(\boldsymbol{A})\|^2\|f(\boldsymbol{A})\boldsymbol{X}\|^2\Big[\|\boldsymbol{W}_1^S\|\|\boldsymbol{W}_2^S\|)\|\Delta\boldsymbol{W}_1\|\\
&+\|\boldsymbol{W}_1^S\|\|\Delta\boldsymbol{W}_2\|\|\boldsymbol{W}_1^S\|+\|\Delta\boldsymbol{W}_1\|\|\boldsymbol{W}_2^{S^i}\|\|\boldsymbol{W}_1^S\|\Big]\\
&\leqslant 2\nu^2\alpha^2M^2\|f(\boldsymbol{A})\|^2\|f(\boldsymbol{A})\boldsymbol{X}\|^2(\|\Delta\boldsymbol{W}_2\|+\|\Delta\boldsymbol{W}_1\|)\qquad\text{Brief Statement}
\end{aligned}
$$

$$
\begin{aligned}
&\|\nabla_{\boldsymbol{w}_1}\ell\big(\mathcal{A}_t^{S^i}(x_i),y_i\big)-\nabla_{\boldsymbol{w}_1}\ell\big(\mathcal{A}_t^{S}(x_i),y_i\big)\|\\
&\left.\begin{aligned}
&\leqslant \nu_\ell\|\boldsymbol{Z}_2^{S'}f(\boldsymbol{A})\boldsymbol{W}_2^S\boldsymbol{Z}_1^{S'}f(\boldsymbol{A})\boldsymbol{X}-\boldsymbol{Z}_2^{S^{i\prime}}f(\boldsymbol{A})\boldsymbol{W}_2^{S^i}\boldsymbol{Z}_1^{S^{i\prime}}f(\boldsymbol{A})\boldsymbol{X}\|\\
&\leqslant \nu_\ell\nu_1\nu_2\alpha_1^2\|f(\boldsymbol{A})\|^2\|f(\boldsymbol{A})\boldsymbol{X}\|^3\Big[\|\boldsymbol{W}_1^S\|^2\|\boldsymbol{W}_2^S\|\|\Delta\boldsymbol{W}_2\|+
\end{aligned}\right\}\text{Lipschitz condition}\\
&\|\boldsymbol{W}_1^S\|\|\boldsymbol{W}_2^S\|\|\boldsymbol{W}_2^{S^i}\|\|\Delta\boldsymbol{W}_1\|+\|\boldsymbol{W}_1^S\|^2\|\boldsymbol{W}_2^{S^i}\|\|\Delta\boldsymbol{W}_2\|+\|\boldsymbol{W}_1^S\|\|\boldsymbol{W}_2^{S^i}\|^2\|\Delta\boldsymbol{W}_1\|\Big]\\
&\leqslant 2\nu^3\alpha^2M^3\|f(\boldsymbol{A})\|^2\|f(\boldsymbol{A})\boldsymbol{X}\|^3(\|\Delta\boldsymbol{W}_2\|+\|\Delta\boldsymbol{W}_1\|)
\end{aligned}
$$

Lemma 4

$$K_1 = v^3 \alpha^2 \|f(\mathbf{A})\|^2 \|f(\mathbf{A})\mathbf{X}\|^3 M^3$$

$$\begin{aligned}
&\|\nabla_{\boldsymbol{w}_1} \ell\left(\mathcal{A}_t^S(x_i), y_i\right) - \nabla_{\boldsymbol{w}_1} \ell\left(\mathcal{A}_t^{S^i}(x_i), y_i'\right)\| \\
&\leqslant v_\ell v_1 v_2 \alpha_1^2 \|f(\boldsymbol{A})\|\|^2 \|f(\boldsymbol{A})\boldsymbol{X}\|^3 \|\boldsymbol{W}_1^S\|^2 \|\boldsymbol{W}_2^S\|^2 \\
&+ v_\ell v_1 v_2 \alpha_1^2 \|f(\boldsymbol{A})\|\|^2 \|f(\boldsymbol{A})\boldsymbol{X}'\|^3 \|\boldsymbol{W}_1^{S^i}\|^2 \|\boldsymbol{W}_2^{S^i}\|^2 \\
&\leqslant 2K_1 M
\end{aligned}$$

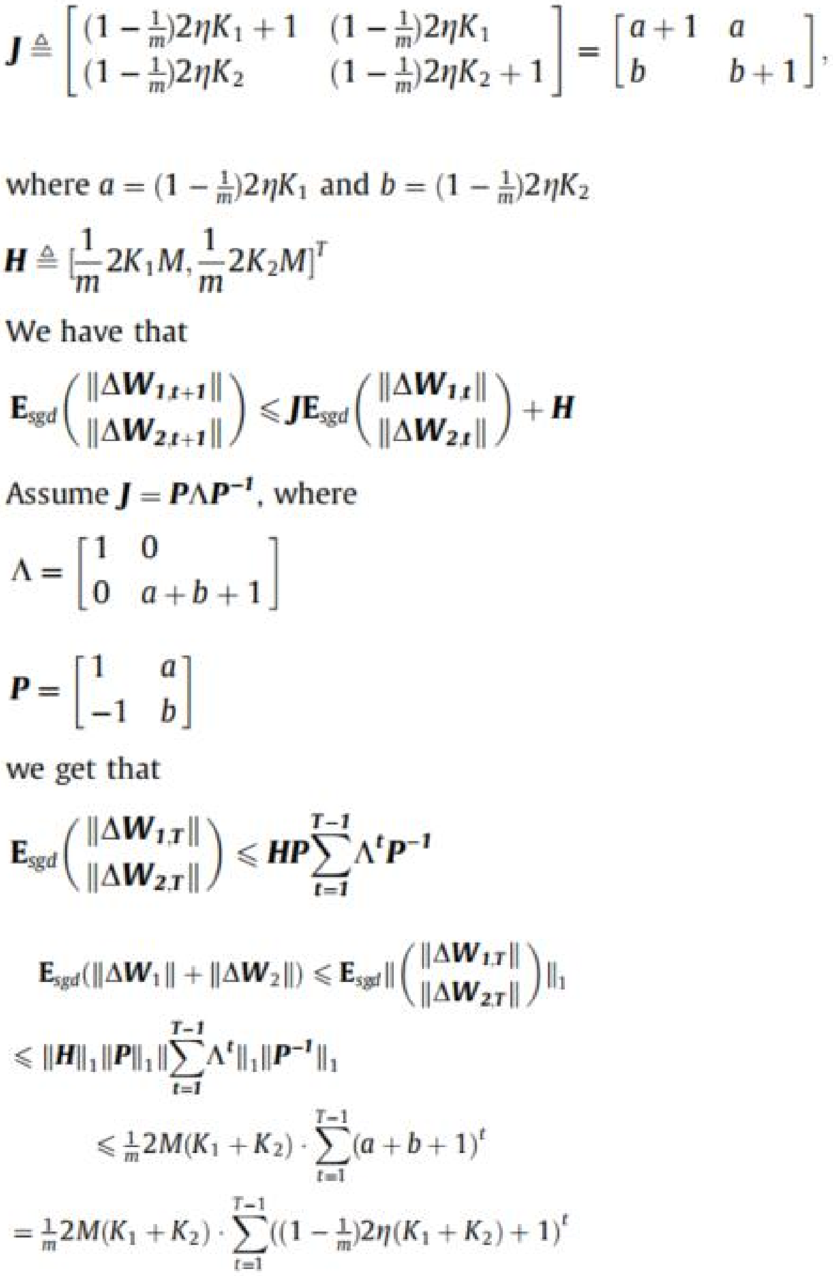
$$K_2 = v^2 \alpha^2 \|f(\mathbf{A})\|^2 \|f(\mathbf{A})\mathbf{X}\|^2 M^2$$

$$\begin{aligned}
&\|\nabla_{\boldsymbol{w}_2} \ell\left(\mathcal{A}_t^S(x_i), y_i\right) - \nabla_{\boldsymbol{w}_2} \ell\left(\mathcal{A}_t^{S^i}(x_i), y_i'\right)\| \\
&\leqslant v_\ell v_1 v_2 \alpha_1^2 \|f(\boldsymbol{A})\|\|^2 \|f(\boldsymbol{A})\boldsymbol{X}\|^3 \|\boldsymbol{W}_2^S\|^2 \|\boldsymbol{W}_2^S\|^2 \\
&+ v_\ell v_1 v_2 \alpha_1^2 \|f(\boldsymbol{A})\|\|^2 \|f(\boldsymbol{A})\boldsymbol{X}'\|^3 \|\boldsymbol{W}_1^{S^i}\|^2 \|\boldsymbol{W}_2^{S^i}\|^2 \\
&\leqslant 2K_2 M
\end{aligned}$$

Using the results above

$$\begin{aligned}
&\mathbf{E}_{sgd}(\|\Delta \boldsymbol{W}_{1,t+1}\| - \|\Delta \boldsymbol{W}_{1,t}\|) \leqslant (1 - \tfrac{1}{m}) 2\eta K_1 (\mathbf{E}_{sgd}(\|\Delta \boldsymbol{W}_{1,t}\| + \|\Delta \boldsymbol{W}_{2,t}\|)) + \tfrac{1}{m} 2K_2 M \\
&\mathbf{E}_{sgd}(\|\Delta \boldsymbol{W}_{2,t+1}\| - \|\Delta \boldsymbol{W}_{2,t}\|) \leqslant (1 - \tfrac{1}{m}) 2\eta K_2 (\mathbf{E}_{sgd}(\|\Delta \boldsymbol{W}_{1.t}\| + \|\Delta \boldsymbol{W}_{2,t}\|))
\end{aligned}$$

PROOF



$$\boldsymbol{J} \triangleq \begin{bmatrix} (1-\frac{1}{m})2\eta K_1 + 1 & (1-\frac{1}{m})2\eta K_1 \\ (1-\frac{1}{m})2\eta K_2 & (1-\frac{1}{m})2\eta K_2 + 1 \end{bmatrix} = \begin{bmatrix} a+1 & a \\ b & b+1 \end{bmatrix},$$

where $a = (1-\frac{1}{m})2\eta K_1$ and $b = (1-\frac{1}{m})2\eta K_2$

$$\boldsymbol{H} \triangleq [\frac{1}{m}2K_1M, \frac{1}{m}2K_2M]^T$$

We have that

$$\mathbf{E}_{sgd}\begin{pmatrix} \|\Delta \boldsymbol{W}_{1,t+1}\| \\ \|\Delta \boldsymbol{W}_{2,t+1}\| \end{pmatrix} \leqslant \boldsymbol{J}\mathbf{E}_{sgd}\begin{pmatrix} \|\Delta \boldsymbol{W}_{1,t}\| \\ \|\Delta \boldsymbol{W}_{2,t}\| \end{pmatrix} + \boldsymbol{H}$$

Assume $\boldsymbol{J} = \boldsymbol{P}\Lambda\boldsymbol{P}^{-1}$, where

$$\Lambda = \begin{bmatrix} 1 & 0 \\ 0 & a+b+1 \end{bmatrix}$$

$$\boldsymbol{P} = \begin{bmatrix} 1 & a \\ -1 & b \end{bmatrix}$$

we get that

$$\mathbf{E}_{sgd}\begin{pmatrix} \|\Delta \boldsymbol{W}_{1,T}\| \\ \|\Delta \boldsymbol{W}_{2,T}\| \end{pmatrix} \leqslant \boldsymbol{H}\boldsymbol{P}\sum_{t=1}^{T-1}\Lambda^t\boldsymbol{P}^{-1}$$

$$\mathbf{E}_{sgd}(\|\Delta \boldsymbol{W}_1\| + \|\Delta \boldsymbol{W}_2\|) \leqslant \mathbf{E}_{sgd}\|\begin{pmatrix} \|\Delta \boldsymbol{W}_{1,T}\| \\ \|\Delta \boldsymbol{W}_{2,T}\| \end{pmatrix}\|_1$$

$$\leqslant \|\boldsymbol{H}\|_1\|\boldsymbol{P}\|_1\|\sum_{t=1}^{T-1}\Lambda^t\|_1\|\boldsymbol{P}^{-1}\|_1$$

$$\leqslant \tfrac{1}{m}2M(K_1+K_2)\cdot\sum_{t=1}^{T-1}(a+b+1)^t$$

$$= \tfrac{1}{m}2M(K_1+K_2)\cdot\sum_{t=1}^{T-1}((1-\tfrac{1}{m})2\eta(K_1+K_2)+1)^t$$

PROOF

$$E_{SGD}\left[\left\|\Delta W_{1,\mathrm{t}+1}\right\|\right]=E_{SGD}\left[\left\|\Delta W_{1,t}\right\|\right]+\left(1-\frac{1}{m}\right)\eta E\left[\left\|A_{1,1}\right\|\right]+\frac{1}{m}\eta E\left[\left\|A_{1,2}\right\|\right]$$

$$E_{SGD}\left[\left\|\Delta W_{2,\mathrm{t}+1}\right\|\right]=E_{SGD}\left[\left\|\Delta W_{2,t}\right\|\right]+\left(1-\frac{1}{m}\right)\eta E\left[\left\|A_{2,1}\right\|\right]+\frac{1}{m}\eta E\left[\left\|A_{2,2}\right\|\right]$$

$$A_{j,condition} : j/condition = 1,2$$

$$A_{1,1} \le 2v^3\alpha^2M^3\left\|f(A)\right\|^2\left\|f(A)X\right\|^3\left(\left\|\Delta W_{2,t}\right\|+\left\|\Delta W_{1,t}\right\|\right)$$

$$A_{1,2} \le 2K_1M \qquad K_1 = 2v^3\alpha^2M^3\left\|f(A)\right\|^2\left\|f(A)X\right\|^3$$

$$A_{2,1} \le 2v^2\alpha^2M^2\left\|f(A)\right\|^2\left\|f(A)X\right\|^2\left(\left\|\Delta W_{2,t}\right\|+\left\|\Delta W_{1,t}\right\|\right)$$

$$A_{2,2} \le 2K_2M \qquad K_2 = 2v^2\alpha^2M^2\left\|f(A)\right\|^2\left\|f(A)X\right\|^2$$

↓

$$E_{SGD}\left(\left\|\Delta W_{1,t+1}\right\|\right)\le\left(1-\frac{1}{m}\right)2\eta K_1\left(E_{SGD}\left(\left\|\Delta W_{1,t}\right\|+\left\|\Delta W_{2,t}\right\|\right)\right)+\frac{1}{m}2K_1M+E_{SGD}\left(\left\|\Delta W_{1,t}\right\|\right)$$

$$E_{SGD}\left(\left\|\Delta W_{2,t+1}\right\|\right)\le\left(1-\frac{1}{m}\right)2\eta K_2\left(E_{SGD}\left(\left\|\Delta W_{1,t}\right\|+\left\|\Delta W_{2,t}\right\|\right)\right)+\frac{1}{m}2K_2M+E_{SGD}\left(\left\|\Delta W_{2,t}\right\|\right)$$

PROOF

$$E_{SGD}\left(\|\Delta W_{1,t+1}\|\right)\leq\left(1-\frac{1}{m}\right)2\eta K_1\left(E_{SGD}\left(\|\Delta W_{1,t}\|+\|\Delta W_{2,t}\|\right)\right)+\frac{1}{m}2K_1M+E_{SGD}\left(\|\Delta W_{1,t}\|\right)$$

$$E_{SGD}\left(\|\Delta W_{2,t+1}\|\right)\leq\left(1-\frac{1}{m}\right)2\eta K_2\left(E_{SGD}\left(\|\Delta W_{1,t}\|+\|\Delta W_{2,t}\|\right)\right)+\frac{1}{m}2K_2M+E_{SGD}\left(\|\Delta W_{2,t}\|\right)$$

→

$$\mathbf{E}_{sgd}\begin{pmatrix}\|\boldsymbol{\Delta W_{1,t+1}}\|\\ \|\boldsymbol{\Delta W_{2,t+1}}\|\end{pmatrix}\leqslant \boldsymbol{J}\mathbf{E}_{sgd}\begin{pmatrix}\|\boldsymbol{\Delta W_{1,t}}\|\\ \|\boldsymbol{\Delta W_{2,t}}\|\end{pmatrix}+\boldsymbol{H}$$

$$\boldsymbol{J}\triangleq\begin{bmatrix}(1-\frac{1}{m})2\eta K_1+1 & (1-\frac{1}{m})2\eta K_1\\ (1-\frac{1}{m})2\eta K_2 & (1-\frac{1}{m})2\eta K_2+1\end{bmatrix}=\begin{bmatrix}a+1 & a\\ b & b+1\end{bmatrix}$$

$$\boldsymbol{H}\triangleq[\frac{1}{m}2K_1M,\frac{1}{m}2K_2M]^T$$

In the last step, we only need to deal with the first - order recursion of E_t

$$E_{SGD}\begin{pmatrix}\|\Delta W_{1,T}\|\\ \|\Delta W_{2,T}\|\end{pmatrix}\leq J^T E_{SGD}\begin{pmatrix}\|\Delta W_{1,1}\|\\ \|\Delta W_{2,1}\|\end{pmatrix}+H\sum_{t=0}^{T-1}J^t$$

$$E_{SGD}\begin{pmatrix}\|\Delta W_{1,T}\|\\ \|\Delta W_{2,T}\|\end{pmatrix}\leq P\wedge^T P^{-1}E_{SGD}\begin{pmatrix}\|\Delta W_{1,1}\|\\ \|\Delta W_{2,1}\|\end{pmatrix}+H\sum_{t=0}^{T-1}P\wedge^t P^{-1}$$

$$\mathbf{E}_{sgd}\begin{pmatrix}\|\boldsymbol{\Delta W_{1,T}}\|\\ \|\boldsymbol{\Delta W_{2,T}}\|\end{pmatrix}\leqslant \boldsymbol{HP}\sum_{t=1}^{T-1}\boldsymbol{\Lambda^t P^{-1}} \qquad (27)$$

PROOF

$$\begin{aligned}
\mathbf{E}_{sgd}(\|\Delta \boldsymbol{W}_1\| + \|\Delta \boldsymbol{W}_2\|) &\leqslant \mathbf{E}_{sgd}\left\|\begin{pmatrix} \|\Delta \boldsymbol{W}_{1,T}\| \\ \|\Delta \boldsymbol{W}_{2,T}\| \end{pmatrix}\right\|_1 \\
&\leqslant \|\boldsymbol{H}\|_1 \|\boldsymbol{P}\|_1 \left\|\sum_{t=1}^{T-1} \Lambda^t\right\|_1 \|\boldsymbol{P}^{-1}\|_1 \\
&\leqslant \tfrac{1}{m} 2M(K_1 + K_2) \cdot \sum_{t=1}^{T-1} (a + b + 1)^t \\
&= \tfrac{1}{m} 2M(K_1 + K_2) \cdot \sum_{t=1}^{T-1} ((1 - \tfrac{1}{m}) 2\eta (K_1 + K_2) + 1)^t
\end{aligned} \tag{28}$$

Finish Proof

$$E_{\text{sgd}}\left[\|\Delta W_1\| + \|\Delta W_2\|\right] \leq \frac{1}{m} 2M(K_1 + K_2) \cdot \sum_{t=1}^{T-1} \left(\left(1 - \frac{1}{m}\right) 2\eta (K_1 + K_2) + 1\right)^t$$

$$\Big\downarrow$$

$$\begin{aligned}
\mathbf{E}_{sgd}[R(\mathcal{A}^S) - R_{emp}(\mathcal{A}^S)] \leqslant{} & \frac{1}{m} \mathcal{O}((K_1 + K_2)^T) \\
& + \left(\mathcal{O}((K_1 + K_2)^T) + Q\right) \sqrt{\frac{\log \frac{1}{\delta}}{2m}},
\end{aligned}$$

PROMOTION

For L-layers,

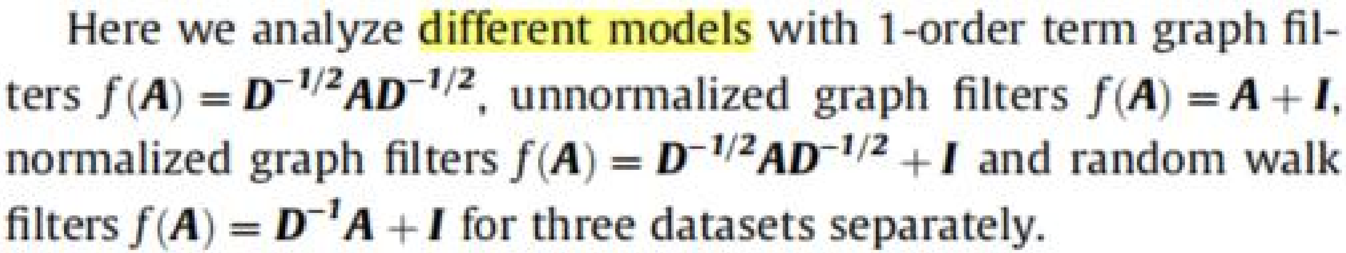
$$|\mathbf{E}_{sgd}[\ell(\mathcal{A}^{S},h)-\ell(\mathcal{A}^{S},h)]|\leqslant\mathcal{O}\left(\mathbf{E}_{sgd}\left[\sum_{n=1}^{L}\|\Delta\boldsymbol{W}_n\|\right]\right)$$

$$\mathbf{E}_{sgd}\left[\|\Delta\boldsymbol{W}_{j,t+1}\|\right]\leqslant\mathbf{E}_{sgd}\left[\|\Delta\boldsymbol{W}_{j,t}\|\right]+(1-\frac{1}{m})2\eta K_j(\mathbf{E}_{sgd}(\sum_{n=1}^{L}\|\Delta\boldsymbol{W}_{n,t}\|))+\frac{1}{m}2K_jM,$$

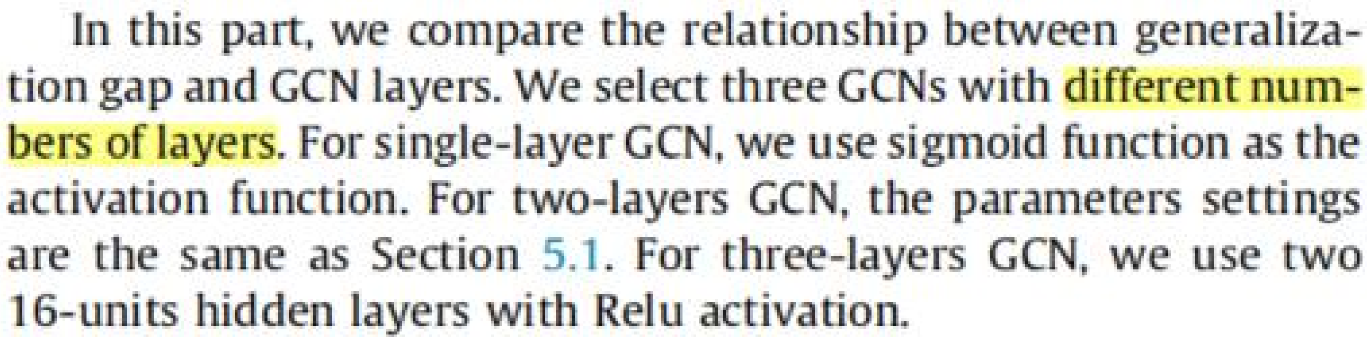
$$\mathbf{E}_{sgd}\left[\sum_{n=1}^{L}\|\Delta\boldsymbol{W}_n\|\right]\leqslant\mathcal{O}((K_1+K_2+\cdots+K_L)^T)$$

$$\mathbf{E}_{sgd}[R(\mathcal{A}^S)-R_{emp}(\mathcal{A}^S)]\leqslant\frac{1}{m}\mathcal{O}\left((K_1+K_2+\cdots+K_L)^T\right)$$
$$+\left(\mathcal{O}((K_1+K_2+\cdots+K_L)^T)+Q\right)\sqrt{\frac{\log\frac{1}{\delta}}{2m}}$$

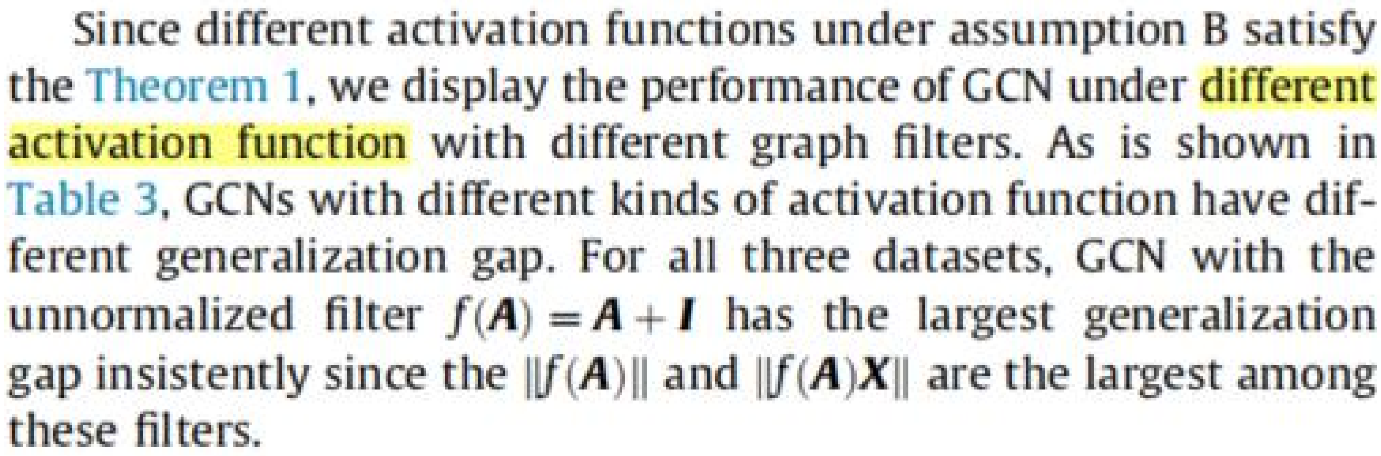
EXPERIMENTS



Here we analyze different models with 1-order term graph filters $f(\boldsymbol{A}) = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2}$, unnormalized graph filters $f(\boldsymbol{A}) = \boldsymbol{A} + \boldsymbol{I}$, normalized graph filters $f(\boldsymbol{A}) = \boldsymbol{D}^{-1/2}\boldsymbol{A}\boldsymbol{D}^{-1/2} + \boldsymbol{I}$ and random walk filters $f(\boldsymbol{A}) = \boldsymbol{D}^{-1}\boldsymbol{A} + \boldsymbol{I}$ for three datasets separately.



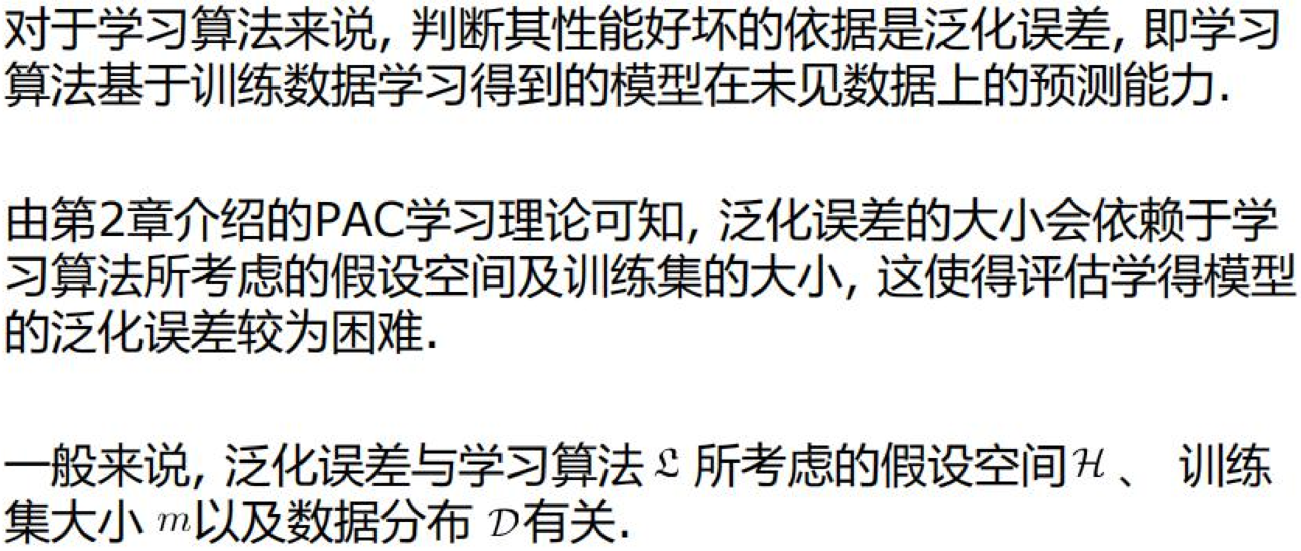
In this part, we compare the relationship between generalization gap and GCN layers. We select three GCNs with different numbers of layers. For single-layer GCN, we use sigmoid function as the activation function. For two-layers GCN, the parameters settings are the same as Section 5.1. For three-layers GCN, we use two 16-units hidden layers with Relu activation.



Since different activation functions under assumption B satisfy the Theorem 1, we display the performance of GCN under different activation function with different graph filters. As is shown in Table 3, GCNs with different kinds of activation function have different generalization gap. For all three datasets, GCN with the unnormalized filter $f(\boldsymbol{A}) = \boldsymbol{A} + \boldsymbol{I}$ has the largest generalization gap insistently since the $\|f(\boldsymbol{A})\|$ and $\|f(\boldsymbol{A})\boldsymbol{X}\|$ are the largest among these filters.

STABILITY OR GENERALIZATION

Chap 4

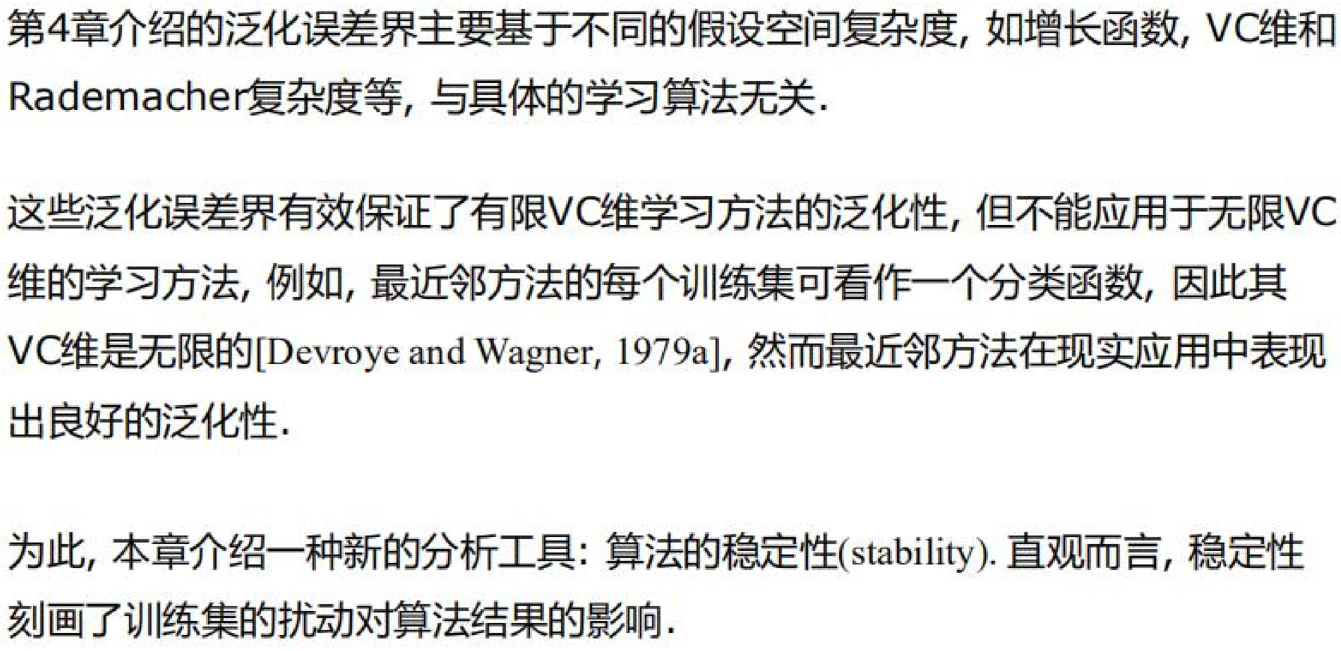


对于学习算法来说，判断其性能好坏的依据是泛化误差，即学习算法基于训练数据学习得到的模型在未见数据上的预测能力.

由第2章介绍的PAC学习理论可知，泛化误差的大小会依赖于学习算法所考虑的假设空间及训练集的大小，这使得评估学得模型的泛化误差较为困难.

一般来说，泛化误差与学习算法 $\mathfrak{L}$ 所考虑的假设空间 $\mathcal{H}$ 、 训练集大小 m 以及数据分布 $\mathcal{D}$ 有关.

Chap 5



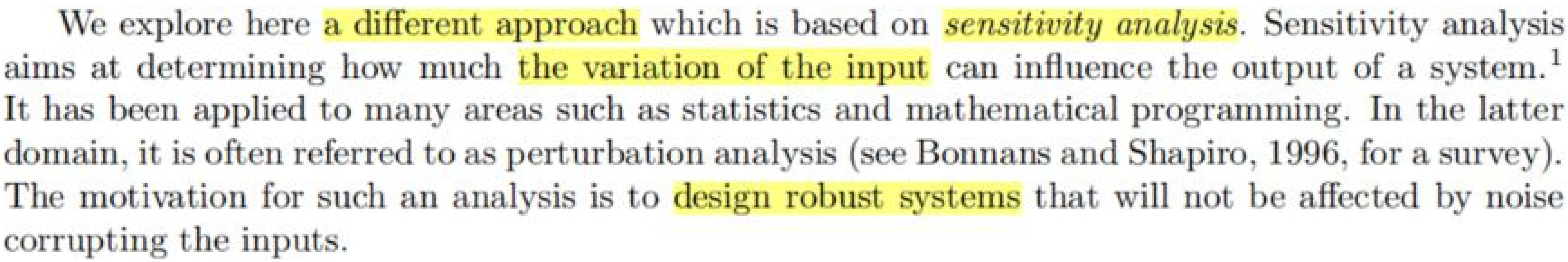
第4章介绍的泛化误差界主要基于不同的假设空间复杂度，如增长函数，VC维和Rademacher复杂度等，与具体的学习算法无关.

这些泛化误差界有效保证了有限VC维学习方法的泛化性，但不能应用于无限VC维的学习方法，例如，最近邻方法的每个训练集可看作一个分类函数，因此其VC维是无限的[Devroye and Wagner, 1979a]，然而最近邻方法在现实应用中表现出良好的泛化性.

为此，本章介绍一种新的分析工具：算法的稳定性(stability). 直观而言，稳定性刻画了训练集的扰动对算法结果的影响.

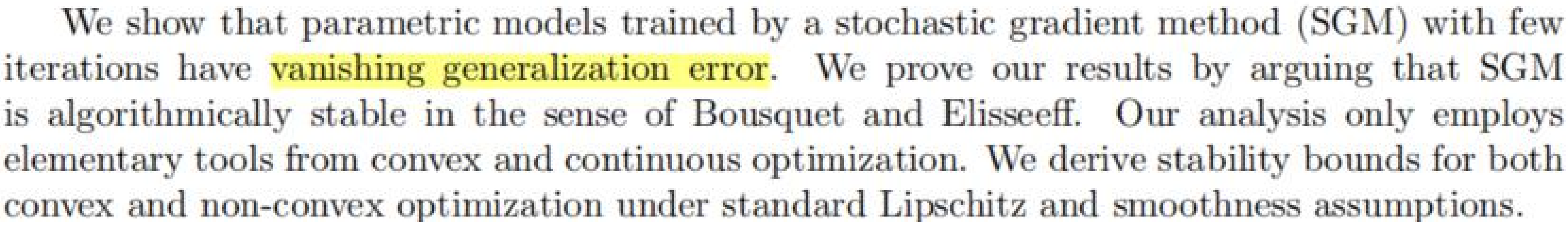
STABILITY OR GENERALIZATION

Stability and Generalization/2002



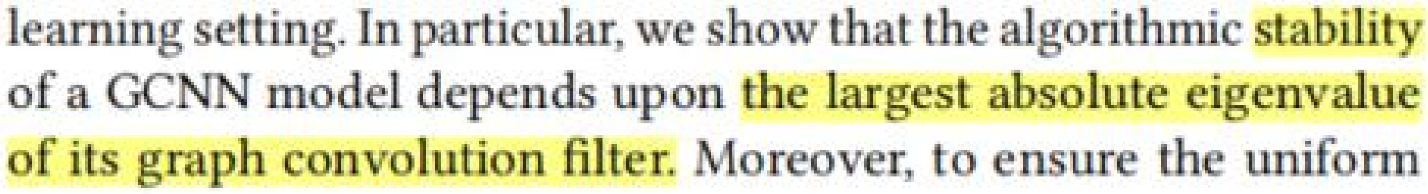
We explore here a different approach which is based on *sensitivity analysis*. Sensitivity analysis aims at determining how much the variation of the input can influence the output of a system.[1] It has been applied to many areas such as statistics and mathematical programming. In the latter domain, it is often referred to as perturbation analysis (see Bonnans and Shapiro, 1996, for a survey). The motivation for such an analysis is to design robust systems that will not be affected by noise corrupting the inputs.

Train faster, generalize better:Stability of stochastic gradient descent/2016



We show that parametric models trained by a stochastic gradient method (SGM) with few iterations have vanishing generalization error. We prove our results by arguing that SGM is algorithmically stable in the sense of Bousquet and Elisseeff. Our analysis only employs elementary tools from convex and continuous optimization. We derive stability bounds for both convex and non-convex optimization under standard Lipschitz and smoothness assumptions.

Stability and Generalization of Graph Convolutional Neural Networks



learning setting. In particular, we show that the algorithmic stability of a GCNN model depends upon the largest absolute eigenvalue of its graph convolution filter. Moreover, to ensure the uniform

REFERENCES

机器学习理论研究导引（研究生, 2023年春季）

1 课程信息

- 课程名称: 机器学习理论研究导引
- 授课对象: 南京大学人工智能学院、计算机科学与技术系, 研究生
- 上课时间: 2023年春季学期, 每周三, 10:10 – 12:00
- 上课地点: 仙I-524
- 授课老师: 王魏

2 课程资料

序号	时间	课件
01	03月01日	第1章 预备知识
02	03月08日	第2章 可学性
03	03月29日	第3章 复杂度
04	03月29日	第4章 泛化性
05	04月19日	第5章 稳定性
06	05月17日	第6章 一致性
07	05月17日	第7章 收敛率
08	06月06日	第8章 遗憾界

https://www.lamda.nju.edu.cn/mlt2023/index.html

Signed Graph Convolutional Network / 2018

Bridging the Gap between Spatial and Spectral Domains: A Survey on Graph Neural Networks / 2021

The generalization error of graph convolutional networks may enlarge with more layers / 2021