# A. Proofs

## A.1. Proof of Theorem 2

*Proof.* For simplicity, we assume the signal vector $\mathbf{x}$ to be non-negative. Note that we can convert $\mathbf{x}$ into a non-negative input layer $\mathbf{H}^{(0)}$ by a linear transformation. We consider a weaker version of GCNII by fixing $\alpha_\ell = 0.5$ and fixing the weight matrix $(1-\beta_\ell)\mathbf{I}_n + \beta_\ell \mathbf{W}^{(\ell)}$ to be $\gamma_\ell \mathbf{I}_n$, where $\gamma_\ell$ is a learnable parameter. We have

$$\mathbf{H}^{(l+1)} = \sigma\left(\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\left(\mathbf{H}^{(\ell)}+\mathbf{x}\right)\gamma_\ell \mathbf{I}_n\right).$$

Since the input feature $\mathbf{x}$ is non-negative, we can remove the ReLU operation:

$$\begin{aligned}\mathbf{H}^{(\ell+1)} &= \gamma_\ell \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}\left(\mathbf{H}^{(\ell)}+\mathbf{x}\right)\\ &= \gamma_\ell\left(\left(\mathbf{I}_n - \tilde{\mathbf{L}}\right)\cdot\left(\mathbf{H}^{(\ell)}+\mathbf{x}\right)\right).\end{aligned}$$

Consequently, we can express the final representation as

$$\mathbf{H}^{(K-1)} = \left(\sum_{\ell=0}^{K-1}\left(\prod_{k=K-\ell-1}^{K-1}\gamma_k\right)\left(\mathbf{I}_n-\tilde{\mathbf{L}}\right)^\ell\right)\mathbf{x}. \quad (8)$$

On the other hand, a polynomial filter of graph $\tilde{G}$ can be expressed as

$$\left(\sum_{k=0}^{K-1}\theta_k\tilde{\mathbf{L}}^k\right)\mathbf{x} = \left(\sum_{i=0}^{k}\theta_k\left(\mathbf{I}_n-\left(\mathbf{I}_n-\tilde{\mathbf{L}}\right)\right)^k\right)\mathbf{x}$$
$$= \left(\sum_{k=0}^{K-1}\theta_k\left(\sum_{\ell=0}^{k}(-1)^\ell\binom{k}{\ell}\left(\mathbf{I}_n-\tilde{\mathbf{L}}\right)^\ell\right)\right)\mathbf{x}.$$

Switching the order of summation follows that a $K$-order polynomial fiter $\left(\sum_{k=0}^{K-1}\theta_k\tilde{\mathbf{L}}^k\right)\mathbf{x}$ can be expressed as

$$\left(\sum_{k=0}^{K-1}\theta_k\tilde{\mathbf{L}}^k\right)\mathbf{x}=\left(\sum_{\ell=0}^{K-1}\left(\sum_{k=\ell}^{K-1}\theta_k(-1)^\ell\binom{k}{\ell}\right)\left(\mathbf{I}_n-\tilde{\mathbf{L}}\right)^\ell\right)\mathbf{x}. \quad (9)$$

To show that GCNII can express an arbitrary $K$-order polynomial filter, we need to prove that there exists a solution $\gamma_\ell$, $\ell = 0,\dots,K-1$ such that the corresponding coefficients of $\left(\mathbf{I}_n-\tilde{\mathbf{L}}\right)^\ell$ in equations (8) and (9) are equivalent. More precisely, we need to show the following equation system

$$\prod_{k=K-\ell-1}^{K-1}\gamma_k = \sum_{k=\ell}^{K-1}\theta_k(-1)^\ell\binom{k}{\ell},\ k=0,\dots,K-1,$$

has a solution $\gamma_\ell$, $\ell = 0,\dots,K-1$. Since the left-hand side is a partial product of $\gamma_k$ from $K-\ell-1$ to $K-1$, we can solve the equation system by

$$\gamma_{K-\ell-1} = \sum_{k=\ell}^{K-1}\theta_k(-1)^\ell\binom{k}{\ell}\Bigg/\sum_{k=\ell-1}^{K-1}\theta_k(-1)^{\ell-1}\binom{k}{\ell-1}, \quad (10)$$

for $\ell = 1,\dots,K-1$ and $\gamma_{K-1} = \sum_{k=0}^{K-1}\theta_k$. Note that the above solution may fail when $\sum_{k=\ell-1}^{K-1}\theta_k(-1)^{\ell-1}\binom{k}{\ell-1} = 0$. In this case, we can set $\gamma_{K-\ell-1}$ sufficiently large so that equation (10) is still a good approximation. We also note that this case is rare because it implies that the $K$-order filter ignores all features from the $\ell$-hop neighbors. This proves that a $K$-layer GCNII can express the $K$-th order polynomial filter $\left(\sum_{i=0}^{k}\theta_i\mathbf{L}^i\right)\mathbf{x}$ with arbitrary coefficients $\theta$. □

## A.2. Proof of Theorem 1

To prove Theorem 1, we need the following *Cheeger Inequality* (Chung, 2007) for lazy random walks.

**Lemma 1** ((Chung, 2007)). *Let $\mathbf{p}_i^{(K)} = \left(\frac{\mathbf{I}_n+\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1}}{2}\right)^K\mathbf{e_i}$ is the $K$-th transition probability vector from node $i$ on connected self-looped graph $\tilde{G}$. Let $\lambda_{\tilde{G}}$ denote the spectral gap of $\tilde{G}$. The $j$-th entry of $\mathbf{p}_i^{(K)}$ can be bounded by*

$$\left|\mathbf{p}_i^{(K)}(j) - \frac{d_j+1}{2m+n}\right| \le \sqrt{\frac{d_j+1}{d_i+1}}\left(1-\frac{\lambda_{\tilde{G}}^2}{2}\right)^K.$$

*Proof of Theorem 1.* Note that $\mathbf{I}_n = \tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{D}}^{1/2}$, we have

$$\begin{aligned}\mathbf{h}^{(K)} &= \left(\frac{\mathbf{I}_n+\tilde{\mathbf{D}}^{-1/2}\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1/2}}{2}\right)^K\cdot\mathbf{x}\\ &= \left(\tilde{\mathbf{D}}^{-1/2}\left(\frac{\mathbf{I}_n+\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1}}{2}\right)\tilde{\mathbf{D}}^{1/2}\right)^K\cdot\mathbf{x}\\ &= \tilde{\mathbf{D}}^{-1/2}\left(\frac{\mathbf{I}_n+\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1}}{2}\right)^K\cdot\left(\tilde{\mathbf{D}}^{1/2}\mathbf{x}\right).\end{aligned}$$

We express $\tilde{\mathbf{D}}^{1/2}\mathbf{x}$ as linear combination of standard basis:

$$\tilde{\mathbf{D}}^{1/2}\mathbf{x} = (\mathbf{D}+\mathbf{I}_n)^{1/2}\mathbf{x} = \sum_{i=1}^{n}\left(\mathbf{x}(i)\sqrt{d_i+1}\right)\cdot\mathbf{e_i},$$

it follows that

$$\begin{aligned}\mathbf{h}^{(K)} &= \tilde{\mathbf{D}}^{-1/2}\left(\frac{\mathbf{I}_n+\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1}}{2}\right)^K\cdot\sum_{i=1}^{n}\left(\mathbf{x}(i)\sqrt{d_i+1}\right)\cdot\mathbf{e_i}\\ &= \sum_{i=1}^{n}\mathbf{x}(i)\sqrt{d_i+1}\cdot\tilde{\mathbf{D}}^{-1/2}\left(\frac{\mathbf{I}_n+\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1}}{2}\right)^K\cdot\mathbf{e_i}.\end{aligned}$$

We note that $\left(\frac{\mathbf{I}_n+\tilde{\mathbf{A}}\tilde{\mathbf{D}}^{-1}}{2}\right)^K \cdot \mathbf{e_i} = \mathbf{p}_i^{(K)}$ is the $K$-th transition probability vector of a random walk from node $i$. By Lemma 1, the $j$-th entry of $\mathbf{p}_i^{(K)}$ can be bounded by

$$\left|\mathbf{p}_i^{(K)}(j) - \frac{d_j+1}{2m+n}\right| \leq \sqrt{\frac{d_j+1}{d_i+1}}\left(1-\frac{\lambda_{\tilde{G}}^2}{2}\right)^K,$$

or equivalently,

$$\mathbf{p}_i^{(K)}(j) = \frac{d_j+1}{2m+n} \pm \sqrt{\frac{d_j+1}{d_i+1}}\left(1-\frac{\lambda_{\tilde{G}}^2}{2}\right)^K.$$

Therefore, we can express the $j$-th entry of $\mathbf{h}^{(K)}$ as

$$\mathbf{h}^{(K)}(j) = \left(\sum_{i=1}^n \sqrt{d_i+1}\mathbf{x}(i)\cdot\tilde{\mathbf{D}}^{-1/2}\mathbf{p}_i^{(K)}\right)(j)$$

$$=\sum_{i=1}^n \sqrt{d_i+1}\mathbf{x}(i)\frac{1}{\sqrt{d_j+1}}\cdot\left(\frac{d_j+1}{2m+n}\pm\sqrt{\frac{d_j+1}{d_i+1}}\left(1-\frac{\lambda_{\tilde{G}}^2}{2}\right)^K\right)$$

$$=\sum_{i=1}^n \frac{\sqrt{(d_j+1)(d_i+1)}}{2m+n}\mathbf{x}(i) \pm \sum_{i=1}^n \mathbf{x}(i)\left(1-\frac{\lambda_{\tilde{G}}^2}{2}\right)^K.$$

This proves

$$\mathbf{h}^{(K)} = \frac{\left\langle\tilde{\mathbf{D}}^{1/2}\mathbf{1},\mathbf{x}\right\rangle}{2m+n}\tilde{\mathbf{D}}^{1/2}\mathbf{1} \pm \left(\sum_{i=1}^n x_i\right)\cdot\left(1-\frac{\lambda_{\tilde{G}}^2}{2}\right)^K\cdot\mathbf{1},$$

and the Theorem follows. □

## B. Hyper-parameters details

Table 6 summarizes the training configuration of GCNII for semi-supervised. $L_{2_d}$ and $L_{2_c}$ denote the weight decay for dense layer and convolutional layer respectively. The searching hyper-parameters include numbers of layers, hidden dimension, dropout, $\lambda$ and $L_{2_c}$ regularization.

Table 7 summarizes the training configuration of all model for full-supervised. We use the full-supervised hyper-parameter setting from DropEdge for JKNet and IncepGCN on citation networks. For other cases, grid search was performed over the following search space: layers (4, 8, 16, 32 ,64), dropedge (0.1, 0.2, 0.3, 0.4, 0.5, 0.6, 0.7, 0.8, 0.9), $\alpha_\ell$ (0.1, 0.2, 0.3, 0.4, 0.5), $\lambda$ (0.5, 1, 1.5), $L_2$ regularization (1e-3, 5e-4, 1e-4, 5e-5, 1e-5, 5e-6, 1e-6).

*Table 6.* The hyper-parameters for Table 2.

| Dataset | Hyper-parameters |
|---|---|
| Cora | layers: 64, $\alpha_\ell$: 0.1, lr: 0.01, hidden: 64, $\lambda$: 0.5, dropout: 0.6, $L_{2_c}$: 0.01, $L_{2_d}$: 0.0005 |
| Citeseer | layers: 32, $\alpha_\ell$: 0.1, lr: 0.01, hidden: 256, $\lambda$: 0.6, dropout: 0.7, $L_{2_c}$: 0.01, $L_{2_d}$: 0.0005 |
| Pubmed | layers: 16, $\alpha_\ell$: 0.1, lr: 0.01, hidden: 256, $\lambda$: 0.4, dropout: 0.5, $L_{2_c}$: 0.0005, $L_{2_d}$: 0.0005 |

*Table 7.* The hyper-parameters for Table 5.

| Dataset | Method | Hyper-parameters |
|---|---|---|
| Cora | APPNP | $\alpha$: 0.1, $L_2$: 0.0005, lr: 0.01, hidden: 64, dropout: 0.5 |
| | GCNII | layers: 64, $\alpha_\ell$: 0.2, lr: 0.01, hidden: 64, $\lambda$: 0.5, dropout: 0.5, $L_2$: 0.0001 |
| Cite. | APPNP | $\alpha$: 0.5, $L_2$: 0.0005, lr: 0.01, hidden: 64, dropout: 0.5 |
| | GCNII | layers: 64, $\alpha_\ell$: 0.5, lr: 0.01, hidden: 64, $\lambda$: 0.5, dropout: 0.5, $L_2$: 5e-6 |
| Pubm. | APPNP | $\alpha$: 0.4, $L_2$: 0.0001, lr: 0.01, hidden: 64, dropout: 0.5 |
| | GCNII | layers: 64, $\alpha_\ell$: 0.1, lr: 0.01, hidden: 64, $\lambda$: 0.5, dropout: 0.5, $L_2$: 5e-6 |
| Cham. | APPNP | $\alpha$: 0.1, $L_2$: 1e-6, lr: 0.01, hidden: 64, dropout: 0.5 |
| | JKNet | layers: 32, lr: 0.01, hidden: 64, dropedge: 0.7, dropout: 0.5, $L_2$: 0.0001 |
| | IncepGCN | layers: 8, lr: 0.01, hidden: 64, dropedge: 0.9, dropout: 0.5, $L_2$: 0.0005 |
| | GCNII | layers: 8, $\alpha_\ell$: 0.2, lr: 0.01, hidden: 64, $\lambda$: 1.5, dropout: 0.5, $L_2$: 0.0005 |
| Corn. | APPNP | $\alpha$: 0.5, $L_2$: 0.005, lr: 0.01, hidden: 64, dropout: 0.5 |
| | JKNet | layers: 4, lr: 0.01, hidden: 64, dropedge: 0.5, dropout: 0.5, $L_2$: 5e-5 |
| | IncepGCN | layers: 16, lr: 0.01, hidden: 64, dropedge: 0.7, dropout: 0.5, $L_2$: 5e-5 |
| | GCNII | layers: 16, $\alpha_\ell$: 0.5, lr: 0.01, hidden: 64, $\lambda$: 1, dropout: 0.5, $L_2$: 0.001 |
| Texa. | APPNP | $\alpha$: 0.5, $L_2$: 0.001, lr: 0.01, hidden: 64, dropout: 0.5 |
| | JKNet | layers: 32, lr: 0.01, hidden: 64, dropedge: 0.8, dropout: 0.5, $L_2$: 5e-5 |
| | IncepGCN | layers: 8, lr: 0.01, hidden: 64, dropedge: 0.8, dropout: 0.5, $L_2$: 5e-6 |
| | GCNII | layers: 32, $\alpha_\ell$: 0.5, lr: 0.01, hidden: 64, $\lambda$: 1.5, dropout: 0.5, $L_2$: 0.0001 |
| Wisc. | APPNP | $\alpha$: 0.5, $L_2$: 0.005, lr: 0.01, hidden: 64, dropout: 0.5 |
| | JKNet | layers: 8, lr: 0.01, hidden: 64, dropedge: 0.8, dropout: 0.5, $L_2$: 5e-5 |
| | IncepGCN | layers: 8, lr: 0.01, hidden: 64, dropedge: 0.7, dropout: 0.5, $L_2$: 0.0001 |
| | GCNII | layers: 16, $\alpha_\ell$: 0.5, lr: 0.01, hidden: 64, $\lambda$: 1, dropout: 0.5, $L_2$: 0.0005 |