A Training Configurations

Data statistics. We summarize the data statistics in our experiments in Table 1.

Table 1: Dataset statistics of the three learning tasks in our experiments.

Learning Task	Dataset	Nodes	Edges	Train/Dev/Test Nodes	Split Ratio (%)
Semi-supervised	Cora	2,708	5,429	140/500/1,000	5.2/18.5/36.9
	Citeseer	3,327	4,732	120/500/1,000	3.6/15.0/30.1
	Pubmed	19,717	44,338	60/500/1,000	0.3/2.5/5.1
Fully-supervised	Cora	2,708	5,429	1624/541/543	60.0/20.0/20.0
	Citeseer	3,327	4,732	1996/665/666	60.0/20.0/20.0
	Pubmed	19,717	44,338	11830/3943/3944	60.0/20.0/20.0
Inductive (large-scale)	Reddit	233K	11.6M	152K/24K/55K	65.2/10.3/23.6

Training hyper-parameters. For both fully and semi-supervised node classification tasks on the citation networks, Cora, Citeseer and Pubmed, we train our DGC following the hyper-parameters in SGC [4]. Specifically, we train DGC for 100 epochs using Adam [2] with learning rate 0.2. For weight decay, as in SGC, we tune this hyperparameter on each dataset using hyperopt [1] for 10,000 trails. For the large-scale inductive learning task on the Reddit network, we also follow the protocols of SGC [4], where we use L-BFGS [3] optimizer for 2 epochs with no weight decay.

B Omitted Proofs

B.1 Proof of Theorem 1

Theorem 1. *The heat kernel $\mathbf{H}_t = e^{-t\mathbf{L}}$ admits the following eigen-decomposition,*

$$\mathbf{H}_t = \mathbf{U} \begin{pmatrix} e^{-\lambda_1 t} & 0 & \cdots & 0 \\ 0 & e^{-\lambda_2 t} & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & e^{-\lambda_n t} \end{pmatrix} \mathbf{U}^\top. \tag{1}$$

As a result, with $\lambda_i \geq 0$, we have

$$\lim_{t\to\infty} e^{-\lambda_i t} = \begin{cases} 0, & \text{if } \lambda_i > 0 \\ 1, & \text{if } \lambda_i = 0 \end{cases}, \ i = 1, \dots, n. \tag{2}$$

Proof. With the eigen-decomposition of the Laplacian $\mathbf{L} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^\top$, the heat kernel can be written equivalently as

$$\mathbf{H}_t = e^{-t\mathbf{L}} = \sum_{k=0}^{\infty} \frac{t^k}{k!}(-\mathbf{L})^k = \sum_{k=0}^{\infty} \frac{t^k}{k!} \left[\mathbf{U}(-\mathbf{\Lambda})\mathbf{U}^\top\right]^k = \mathbf{U}\left[\sum_{k=0}^{\infty} \frac{t^k}{k!}(-\mathbf{\Lambda})^k\right]\mathbf{U}^T = \mathbf{U}e^{-t\mathbf{\Lambda}}\mathbf{U}^T, \tag{3}$$

which corresponds to the eigen-decomposition of the heat kernel with eigen-vectors in the orthogonal matrix $\mathbf{U}$ and eigven-values in the diagonal matrix $e^{-t\mathbf{\Lambda}}$. Now it is easy to see the limit behavior of the heat kernel as $t \to \infty$ from the spectral domain. □

B.2 Proof of Theorem 2

Theorem 2. *For the general initial value problem*

$$\begin{cases} \frac{d\mathbf{X}_t}{dt} & = -\mathbf{L}\mathbf{X}_t, \\ \mathbf{X}_0 & = \mathbf{X}, \end{cases} \tag{4}$$

with any finite terminal time T, the numerical error of the forward Euler method

$$\hat{\mathbf{X}}_T^{(K)} = \left(\mathbf{I} - \frac{T}{K}\mathbf{L}\right)^K \mathbf{X}_0. \tag{5}$$

with K propagation steps can be upper bounded by

$$\|\mathbf{e}_T^{(K)}\| \leq \frac{T\|\mathbf{L}\|\|\mathbf{X}_0\|}{2K}\left(e^{T\|\mathbf{L}\|}-1\right). \tag{6}$$

Proof. Consider a general Euler forward scheme for our initial problem

$$\hat{\mathbf{X}}^{(k+1)} = \hat{\mathbf{X}}^{(k)} - h\mathbf{L}\hat{\mathbf{X}}_t, \quad k=0,1,\ldots,K-1, \quad \mathbf{X}^{(0)} = \mathbf{X}, \tag{7}$$

where $\hat{\mathbf{X}}^{(k)}$ denotes the approximated $\mathbf{X}$ at step k, h denotes the step size and the terminal time $T = Kh$. We denote the global error at step k as

$$\mathbf{e}_k = \mathbf{X}^{(k)} - \hat{\mathbf{X}}^{(k)}, \tag{8}$$

and the truncation error of the Euler forward finite difference (Eqn. (7)) at step k as

$$\mathbf{T}^{(k)} = \frac{\mathbf{X}^{(k+1)} - \mathbf{X}^{(k)}}{h} + \mathbf{L}\mathbf{X}^{(k)}. \tag{9}$$

We continue by noting that Eqn. (9) can be written equivalently as

$$\mathbf{X}^{(k+1)} = \mathbf{X}^{(k)} + h\left(\mathbf{T}^{(k)} - \mathbf{L}\mathbf{X}^{(k)}\right). \tag{10}$$

Taking the difference of Eqn. (10) and (7), we have

$$\mathbf{e}^{(k+1)} = (1-h\mathbf{L})\mathbf{e}^{(k)} + h\mathbf{T}^{(k)}, \tag{11}$$

whose norm can be upper bounded as

$$\left\|\mathbf{e}^{(k+1)}\right\| \leq (1+h\|\mathbf{L}\|)\left\|\mathbf{e}^{(k)}\right\| + h\left\|\mathbf{T}^{(k)}\right\|. \tag{12}$$

Let $M = \max_{0\leq k\leq K-1}\|\mathbf{T}^{(k)}\|$ be the upper bound on global truncation error, we have

$$\left\|\mathbf{e}^{(k+1)}\right\| \leq (1+h\|\mathbf{L}\|)\left\|\mathbf{e}^{(k)}\right\| + hM. \tag{13}$$

By induction, and noting that $1+h\|\mathbf{L}\| \leq e^{h\|\mathbf{L}\|}$ and $\mathbf{e}^{(0)} = \mathbf{X}^{(0)} - \hat{\mathbf{X}}^{(0)} = \mathbf{0}$, we have

$$\left\|\mathbf{e}^{(K)}\right\| \leq \frac{M}{\|\mathbf{L}\|}\left[(1+h\|\mathbf{L}\|)^n - 1\right] \leq \frac{M}{\|\mathbf{L}\|}\left(e^{Kh\|\mathbf{L}\|}-1\right). \tag{14}$$

Now we note that $\frac{d\mathbf{X}^{(k)}}{dt} = -\mathbf{L}\mathbf{X}^{(k)}$ and applying Taylor's theorem, there exists $\delta \in [nh, (k+1)h]$ such that the truncation error $\mathbf{T}^{(k)}$ in Eqn. (9) follows

$$\mathbf{T}^{(k)} = \frac{1}{2h}\mathbf{L}^2\mathbf{X}_\delta. \tag{15}$$

Thus the truncation error can be bounded by

$$\left\|\mathbf{T}^{(k)}\right\| = \frac{1}{2h}\|\mathbf{L}\|^2\|\mathbf{X}_\delta\| \leq \frac{1}{2h}\|\mathbf{L}\|^2\|\mathbf{X}_0\|, \tag{16}$$

because

$$\|\mathbf{X}_\delta\| = \left\|e^{-\delta\mathbf{L}}\mathbf{X}_0\right\| \leq \|\mathbf{X}_0\|, \ \forall\delta \geq 0. \tag{17}$$

Together with the fact $T = Kh$, we have

$$\left\|\mathbf{e}^{(K)}\right\| \leq \frac{\|\mathbf{L}\|^2\|\mathbf{X}_0\|}{2h\|\mathbf{L}\|}\left(e^{Kh\|\mathbf{L}\|}-1\right) = \frac{T\|\mathbf{L}\|\|\mathbf{X}_0\|}{2K}\left(e^{T\|\mathbf{L}\|}-1\right), \tag{18}$$

which completes the proof. □

B.3 Proof of Theorem 3

For the ground-truth data generation process

$$\mathbf{Y} = \mathbf{X}_c \mathbf{W}_c + \sigma_y \boldsymbol{\varepsilon}_y,\ \boldsymbol{\varepsilon}_y \sim \mathcal{N}(\mathbf{0}, \mathbf{I}); \tag{19}$$

together with the data corruption process,

$$\frac{d\widetilde{\mathbf{X}}_t}{dt} = \mathbf{L}\widetilde{\mathbf{X}}_t, \ \text{ where } \widetilde{\mathbf{X}}_0 = \mathbf{X}_c \ \text{ and } \ \widetilde{\mathbf{X}}_{T^*} = \mathbf{X}. \tag{20}$$

and the final state $\mathbf{X}$ denote the observed data. Then, we have the following bound its population risks.

Theorem 3. *Denote the population risk of the ground truth regression problem with weight* $\mathbf{W}$ *as*

$$R(\mathbf{W}) = \mathbb{E}_{p(\mathbf{X}_c, \mathbf{Y})} \|\mathbf{Y} - \mathbf{X}_c \mathbf{W}\|^2 . \tag{21}$$

and that of the corrupted regression problem as

$$\hat{R}(\mathbf{W}) = \mathbb{E}_{p(\hat{\mathbf{X}}, \mathbf{Y})} \left\|\mathbf{Y} - [\mathbf{S}^{(\hat{T}/K)}]^K \hat{\mathbf{X}}\mathbf{W}\right\|^2 . \tag{22}$$

Supposing that $\mathbb{E}\|\mathbf{X}_c\|^2 = M < \infty$*, we have the following upper bound on the latter risk:*

$$\hat{R}(\mathbf{W}) \leq R(\mathbf{W}) + \|\mathbf{W}\|^2 \left[\sigma_x^2 + (M + \sigma_x^2) \left\|e^{T^\star \mathbf{L}}\right\|^2 \cdot \left(\left\|e^{-T^\star \mathbf{L}} - e^{-\hat{T}\mathbf{L}}\right\|^2\right) + \mathbb{E}\left\|\mathbf{e}_{T^\star}^{(K)}\right\|^2\right]. \tag{23}$$

Proof. Given the fact that $\mathbf{X}_c = e^{-T^* \mathbf{L}}\mathbf{X}$, we can decompose the corrupted population risk as follows

$$\begin{aligned}
&\hat{R}(\mathbf{W}) = \mathbb{E}_{p(\hat{\mathbf{X}}, \mathbf{Y})} \left\|\mathbf{Y} - \left[\mathbf{S}^{(\hat{T}/K)}\right]^K \mathbf{X}\mathbf{W}\right\|^2 \\
=&\mathbb{E}_{p(\mathbf{X}, \mathbf{Y})} \left\|\mathbf{Y} - \mathbf{X}_c\mathbf{W} + \left(e^{-T^\star \mathbf{L}} - \left[\mathbf{S}^{(\hat{T}/K)}\right]^K\right)\mathbf{X}\mathbf{W}\right\|^2 \\
\leq&\mathbb{E}_{p(\mathbf{X}, \mathbf{Y})} \|\mathbf{Y} - \mathbf{X}_c\mathbf{W}\|^2 + + \|\mathbf{W}\|^2 \, \mathbb{E}_{p(\mathbf{X}, \mathbf{Y})} \left\|\left(\left[e^{-\hat{T}\mathbf{L}} - \mathbf{S}^{(\hat{T}/K)}\right]^K\right)\mathbf{X} + \left(e^{-T^\star \mathbf{L}} - e^{-\hat{T}\mathbf{L}}\right)\mathbf{X}\right\|^2 \\
\leq&\mathbb{E}_{p(\mathbf{X}, \mathbf{Y})} \|\mathbf{Y} - \mathbf{X}_c\mathbf{W}\|^2 + \|\mathbf{W}\|^2 \, \mathbb{E}_{p(\mathbf{X}, \mathbf{Y})} \left\|\mathbf{e}_{\hat{T}}^{(K)} + \left(e^{-T^\star \mathbf{L}} - e^{-\hat{T}\mathbf{L}}\right) e^{T^\star \mathbf{L}}\mathbf{X}_0\right\|^2 \\
\leq& R(\mathbf{W}) + \|\mathbf{W}\|^2 \left(\mathbb{E}\left\|\mathbf{e}_{\hat{T}}^{(K)}\right\|^2 + M\left\|e^{T^\star \mathbf{L}}\right\|^2 \left\|e^{-T^\star \mathbf{L}} - e^{-\hat{T}\mathbf{L}}\right\|^2\right),
\end{aligned} \tag{24}$$

which completes the proof. □

References

[1] James Bergstra, Brent Komer, Chris Eliasmith, Dan Yamins, and David D Cox. Hyperopt: a python library for model selection and hyperparameter optimization. *Computational Science & Discovery*, 8(1):014008, 2015.

[2] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *ICLR*, 2015.

[3] Dong C Liu and Jorge Nocedal. On the limited memory BFGS method for large scale optimization. *Mathematical programming*, 45(1):503–528, 1989.

[4] Felix Wu, Tianyi Zhang, Amauri Holanda de Souza Jr, Christopher Fifty, Tao Yu, and Kilian Q Weinberger. Simplifying graph convolutional networks. *ICML*, 2019.