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.

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

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 \( H_t = e^{-tL} \) admits the following eigen-decomposition,

\[
H_t = U \begin{pmatrix}
  e^{-\lambda_1 t} & 0 & \cdots & 0 \\
  0 & e^{-\lambda_2 t} & \cdots & 0 \\
  \vdots & \vdots & \ddots & \vdots \\
  0 & 0 & \cdots & e^{-\lambda_n t}
\end{pmatrix} U^T.
\]

As a result, with \( \lambda_i \geq 0 \), we have

\[
\lim_{t \to \infty} e^{-\lambda_i t} = \begin{cases} 
0, & \text{if } \lambda_i > 0 \\
1, & \text{if } \lambda_i = 0 \end{cases}, \quad i = 1, \ldots, n.
\]

Proof. With the eigen-decomposition of the Laplacian \( L = U\Lambda U^T \), the heat kernel can be written equivalently as

\[
H_t = e^{-tL} = \sum_{k=0}^{\infty} \frac{t^k}{k!} (-L)^k = U \left[ (\Lambda U^T)^k \right] U^T = U e^{-t\Lambda} U^T,
\]

which corresponds to the eigen-decomposition of the heat kernel with eigen-vectors in the orthogonal matrix \( U \) and eigen-values in the diagonal matrix \( e^{-t\Lambda} \). 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

\[
\begin{align*}
\frac{dX_t}{dt} &= -LX_t, \\
X_0 &= X,
\end{align*}
\]

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

\[
\hat{X}_T^K = \left( I - \frac{T}{K}L \right)^K X_0.
\]
with $K$ propagation steps can be upper bounded by
\[
\|e_T^{(K)}\| \leq \frac{T\|L\|\|X_0\|}{2K} \left(e^{T\|L\|} - 1\right).
\] (6)

**Proof.** Consider a general Euler forward scheme for our initial problem
\[
\dot{X}^{(k+1)} = \dot{X}^{(k)} - hL\dot{X}^{(k)}, \quad k = 0, 1, \ldots, K-1, \quad X^{(0)} = X,
\] (7)
where $\dot{X}^{(k)}$ denotes the approximated $X$ at step $k$, $h$ denotes the step size and the terminal time $T = Kh$. We denote the global error at step $k$ as
\[
e_k = X^{(k)} - \hat{X}^{(k)},
\] (8)
and the truncation error of the Euler forward finite difference (Eqn. [7]) at step $k$ as
\[
T^{(k)} = \frac{X^{(k+1)} - X^{(k)}}{h} + LX^{(k)}.
\] (9)
We continue by noting that Eqn. (9) can be written equivalently as
\[
X^{(k+1)} = X^{(k)} + h \left(T^{(k)} - LX^{(k)}\right).
\] (10)
Taking the difference of Eqn. (10) and (7), we have
\[
e^{(k+1)} = (1 - hL)e^{(k)} + hT^{(k)},
\] (11)
whose norm can be upper bounded as
\[
\|e^{(k+1)}\| \leq (1 + h\|L\|)\|e^{(k)}\| + h\|T^{(k)}\|.
\] (12)
Let $M = \max_{0 \leq k \leq K-1} \|T^{(k)}\|$ be the upper bound on global truncation error, we have
\[
\|e^{(k+1)}\| \leq (1 + h\|L\|)\|e^{(k)}\| + hM.
\] (13)
By induction, and noting that $1 + h\|L\| \leq e^{h\|L\|}$ and $e^{(0)} = X^{(0)} - \dot{X}^{(0)} = 0$, we have
\[
\|e^{(K)}\| \leq M \left((1 + h\|L\|)^n - 1\right) \leq M \left(e^{kh\|L\|} - 1\right).
\] (14)
Now we note that $\frac{dX^{(k)}}{dt} = -LX^{(k)}$ and applying Taylor’s theorem, there exists $\delta \in [nh, (k+1)h]$ such that the truncation error $T^{(k)}$ in Eqn. (9) follows
\[
T^{(k)} = \frac{1}{2h}L^2X_\delta.
\] (15)
Thus the truncation error can be bounded by
\[
\|T^{(k)}\| \leq \frac{1}{2h}\|L\|^2\|X_\delta\| \leq \frac{1}{2h}\|L\|^2\|X_0\|,
\] (16)
because
\[
\|X_\delta\| = \|e^{-\delta L}X_0\| \leq \|X_0\|, \forall \delta \geq 0.
\] (17)
Together with the fact $T = Kh$, we have
\[
\|e^{(K)}\| \leq \frac{\|L\|^2\|X_0\|}{2h\|L\|} \left(e^{Kh\|L\|} - 1\right) = \frac{T\|L\|\|X_0\|}{2K} \left(e^{r\|L\|} - 1\right),
\] (18)
which completes the proof. \qed
B.3 Proof of Theorem 3

For the ground-truth data generation process
\[ Y = X_c W_e + \sigma_y \varepsilon_y, \varepsilon_y \sim \mathcal{N}(0, I); \] (19)

together with the data corruption process,
\[ \frac{d\tilde{X}_t}{dt} = L\tilde{X}_t, \text{ where } \tilde{X}_0 = X_c \text{ and } \tilde{X}_{T'} = X. \] (20)

and the final state \( X \) 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 \( W \) as \[ R(W) = \mathbb{E}_{p(x, y)} \| Y - X_c W \|^2. \] (21)

and that of the corrupted regression problem as \[ \tilde{R}(W) = \mathbb{E}_{p(\tilde{x}, \tilde{y})} \| \tilde{Y} - [S^{(T/K)}]^K \tilde{X} W \|^2. \] (22)

Supposing that \( \mathbb{E}\|X_c\|^2 = M < \infty \), we have the following upper bound on the latter risk:
\[ \tilde{R}(W) \leq R(W) + \| W \|^2 \left[ \sigma_y^2 + (M + \sigma_y^2) \| T^* L \|^2 \cdot \left( \| e^{T^* L} - e^{-T^* L} \|^2 \right) + E \| e_{T^*}^{(k)} \|^2 \right]. \] (23)

**Proof.** Given the fact that \( X_c = e^{-T^* L} X \), we can decompose the corrupted population risk as follows
\[ \tilde{R}(W) = \mathbb{E}_{p(\tilde{x}, \tilde{y})} \| \tilde{Y} - [S^{(T/K)}]^K \tilde{X} W \|^2 \]
\[ = \mathbb{E}_{p(x, y)} \| Y - X_c W + \left( e^{-T^* L} - [S^{(T/K)}]^K \right) X W \|^2 \]
\[ \leq \mathbb{E}_{p(x, y)} \| Y - X_c W \|^2 + \| W \|^2 \mathbb{E}_{p(x, y)} \left( \left( e^{-T^* L} - S^{(T/K)} \right)^K \right) X + \left( e^{-T^* L} - e^{-T^* L} \right) X \|^2 \]
\[ \leq \mathbb{E}_{p(x, y)} \| Y - X_c W \|^2 + \| W \|^2 \mathbb{E}_{p(x, y)} \left( \left( e^{-T^* L} - S^{(T/K)} \right)^K \right) X + \left( e^{-T^* L} - e^{-T^* L} \right) e^{T^* L} X_0 \|^2 \]
\[ \leq R(W) + \| W \|^2 \left( E \| e_{T^*}^{(k)} \|^2 + \| e^{-T^* L} \|^2 \| e^{-T^* L} - e^{-T^* L} \|^2 \right), \] (24)

which completes the proof. \( \square \)

**References**


