Supplementary materials

ECHO: Adaptive Correction for Subgraph-wise Sampling with Lightweight Hyperparameter Search

Contents

A PROOF OF THEOREM 3.3

By the smoothness of $\mathcal{L}(\boldsymbol{\theta}_t)$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right] \leq \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t}\right)\right] + \mathbb{E}\left[\langle\nabla\mathcal{L}\left(\boldsymbol{\theta}_{t}\right), \boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_{t}\rangle\right] \\ +\frac{L_f}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_{t}\|^2\right]\end{aligned} \tag{1}$$

according to the update rule of subgraph-wise sampling training

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t} - \eta\nabla\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}_{t})$$

by taking the norm on the both side, we have

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_{t}\|^2\right] = \eta^2\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}_{t})\|^2\right] \tag{2}$$

We can give the upper bound for the second term on the right of Eq.1

$$\begin{aligned}&\mathbb{E}\left[\langle\nabla\mathcal{L}\left(\boldsymbol{\theta}_{t}\right), \boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_{t}\rangle\right] \\ =& -\eta\mathbb{E}\left[\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{t}), \nabla\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}_{t})\rangle\right] \\ =& -\eta\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_{t}), \mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}_{t})\right]\right\rangle \\ \underset{(a)}{=}& -\frac{\eta}{2}\left(\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{t}\right)\right\|^2 + \|\mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\theta}_{t}\right)\right]\|^2\right) \\ &+\frac{\eta}{2}\left(\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_{t}\right) - \mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\theta}_{t}\right)\right]\|^2\right) \\ \underset{(b)}{\leq}& -\frac{\eta}{2}\left(\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}^{t}\right)\right\|^2 + \|\mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\theta}_{t}\right)\right]\|^2 - 2\kappa_1^2 - 4\kappa_2^2\right)\end{aligned} \tag{3}$$

where (a) is due to $2\langle \mathrm{x}, \mathrm{y}\rangle = \|\mathrm{x}\|^2 + \|\mathrm{y}\|^2 - \|\mathrm{x}-\mathrm{y}\|^2$
where (b)

$$\begin{aligned}&\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_{t}\right) - \mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\theta}_{t}\right)\right]\|^2 \\ =&\mathbb{E}\left[\left\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_{t}\right) - \nabla\mathcal{L}_{\mathcal{P}}^{\text{full}}(\boldsymbol{\theta}_{t}) + \nabla\mathcal{L}_{\mathcal{P}}^{\text{full}}(\boldsymbol{\theta}_{t}) - \nabla\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\theta}_{t}\right)\right\|^2\right] \\ \leq& 2\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{P}}^{\text{full}}(\boldsymbol{\theta}_{t}) - \nabla\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}_{t})\right\|^2\right] \\ &+4\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{P}}^{\text{full}}(\boldsymbol{\theta}_{t}) - \nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_{t})\right\|^2\right] \\ &+4\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_{t}) - \mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_{t})\right]\right\|^2 \\ \underset{(c)}{\leq}& 2\kappa_1^2 + 4\kappa_2^2\end{aligned} \tag{4}$$

where (c) is due to $\nabla\mathcal{L}(\boldsymbol{\theta}_t) = \mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right]$ and follows the definition of κ_1^2, κ_2^2.

Combining Eq.1,2,3 gives us

$$\begin{aligned}&\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right] \\ \leq&\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t}\right)\right] + \frac{L_f}{2}\eta^2\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}_{t})\|^2\right] \\ &-\frac{\eta}{2}\left(\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_{t}\right)\|^2 + \|\mathbb{E}\left[\nabla\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\theta}_{t}\right)\right]\|^2 - 2\kappa_1^2 - 4\kappa_2^2\right)\end{aligned} \tag{5}$$

Organizing the above formula gives us

$$\begin{aligned}\|\nabla\mathcal{L}\left(\boldsymbol{\theta}_{t}\right)\|^2 \leq& \frac{2}{\eta}\left(\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t}\right)\right] - \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right]\right) \\ &-(1-L_f\eta)\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\theta}_{t}\right)\|^2\right] \\ &+2\kappa_1^2 + 4\kappa_2^2\end{aligned} \tag{6}$$

Summing over $t \in \{0, \cdots, T-1\}$ and dividing both side by T, we get,

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\left[\nabla\mathcal{L}(\boldsymbol{\theta}_t)\right]\right\|^2 \\ \leq&\frac{2}{T\eta}\sum_{t=0}^{T-1}(\mathcal{L}(\theta_t) - \mathcal{L}(\theta_{(t+1)})) - \frac{1}{T}\sum_{t=0}^{T-1}(1-L_f\eta)\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{P}}(\theta_t)\|^2\right] \\ &+2\kappa_1^2 + 4\kappa_2^2\end{aligned} \tag{7}$$

Since $\max_t\left(\nabla\mathcal{L}(\theta_t) - \nabla\mathcal{L}(\theta_{(t+1)})\right) \leq \nabla\mathcal{L}(\theta_0) - \nabla\mathcal{L}(\theta^*)$

$$\begin{aligned}&\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\left[\nabla\mathcal{L}(\boldsymbol{\theta}_t)\right]\right\|^2 \\ \leq&\frac{2}{\eta}((\nabla\mathcal{L}\theta_0) - \nabla\mathcal{L}(\theta^*)) - \frac{1}{T}\sum_{t=0}^{T-1}(1-L_f\eta)\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{P}}(\theta_t)\|^2\right] \\ &+2\kappa_1^2 + 4\kappa_2^2\end{aligned} \tag{8}$$

if we choose $\eta = \frac{1}{\sqrt{T}}$, where $0 < \eta < \frac{1}{L_f}$, then we have,

$$\frac{1}{T}\sum_{t=0}^{T-1}\mathbb{E}\left\|\left[\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right]\right\|^2 = O\left(\frac{1}{\sqrt{T}}\right) + O\left(\kappa_1^2 + 2\kappa_2^2\right) \tag{9}$$

B PROOF OF THEOREM 4.2

For convenience, we use $\mathcal{T} = \{1, \cdots, T\}$ to denote the total training epochs. Let $\mathcal{T}_s \subseteq \mathcal{T}$ to denote subgraph-wise sampling epochs and $\mathcal{T}_c \subseteq \mathcal{T}$ to denote correction epochs.

By the smoothness of $\mathcal{L}(\boldsymbol{\theta}_t)$, we have

$$\begin{aligned}\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right] \leq \mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t}\right)\right] + \mathbb{E}\left[\langle\nabla\mathcal{L}\left(\boldsymbol{\theta}_{t}\right), \boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_{t}\rangle\right] \\ +\frac{L_f}{2}\mathbb{E}\left[\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_{t}\|^2\right]\end{aligned} \tag{10}$$

For $t \in \mathcal{T}_c$, the update rule is:

$$\boldsymbol{\theta}_{t+1} = \boldsymbol{\theta}_{t} - \gamma\tilde{\nabla}\mathcal{L}_{\mathcal{B}}\left(\boldsymbol{\theta}_{t}\right) \tag{11}$$

By taking the norm on the both side, we have

$$\mathbb{E}\left[\|\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\|^2\right]=\gamma^2\mathbb{E}\left[\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2\right] \tag{12}$$

We can give the upper bound for the second term on the right of Eq.10

$$\begin{aligned}
&\mathbb{E}\left[\langle\nabla\mathcal{L}\left(\boldsymbol{\theta}_t\right),\boldsymbol{\theta}_{t+1}-\boldsymbol{\theta}_t\rangle\right]\\
=&-\gamma\mathbb{E}\left[\left\langle\nabla\mathcal{L}(\boldsymbol{\theta}_t),\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\rangle\right]\\
=&-\frac{\gamma}{2}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2+\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2-\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)-\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2\right]\\
\underset{(d)}{\leq}&-\frac{\gamma}{2}\left(\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]+\mathbb{E}\left[\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2\right]-\sigma_{\text{bias}}^2\right)
\end{aligned} \tag{13}$$

Where (d) is due to Assumption ??.

Combining Eq.10,12,13 gives us

$$\begin{aligned}
&\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right]\\
\leq&\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right]-\frac{\gamma}{2}\left(\mathbb{E}\left\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\right\|^2+\mathbb{E}\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2-\sigma_{\text{bias}}^2\right)\\
&+\frac{L_f}{2}\gamma^2\mathbb{E}\left[\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2\right]
\end{aligned} \tag{14}$$

Reorganizing the above equation gives us

$$\begin{aligned}
\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]\leq&\frac{2}{\gamma}\left(\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right]-\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right]\right)\\
&+(L_f\gamma-1)\mathbb{E}\left[\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2\right]+\sigma_{\text{bias}}^2
\end{aligned} \tag{15}$$

For subgraph-wise sampling epochs $t\in\mathcal{T}_m$, we have Eq.6. Summing over $t\in\{1,\cdots,T\}$ and combing Eq.6 and Eq.15, we have,

$$\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]=\sum_{t\in\mathcal{T}_c}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]+\sum_{t\in\mathcal{T}_m}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right] \tag{16}$$

Reorganizing the above equation gives us

$$\begin{aligned}
&\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]\\
\leq&\sum_{t\in\mathcal{T}_c}\left[\frac{2}{\gamma}\left(\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right]-\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right]\right)\right]\\
&+\sum_{t\in\mathcal{T}_s}\left[\frac{2}{\eta}\left(\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right]-\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right]\right)\right]\\
&+\sum_{t\in\mathcal{T}_c}(L_f\gamma-1)\mathbb{E}\left[\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2\right]\\
&+\sum_{t\in\mathcal{T}_s}(L_f\eta-1)\mathbb{E}\left[\left\|\nabla\mathcal{L}_{\mathcal{P}}^{\text{local}}\left(\boldsymbol{\theta}_t\right)\right\|^2\right]\\
&+\sum_{t\in\mathcal{T}_c}\sigma_{\text{bias}}^2+\sum_{t\in\mathcal{T}_s}\left(2\kappa_1^2+4\kappa_2^2\right)
\end{aligned} \tag{17}$$

We aim to select appropriate $\mathcal{T}_c,\mathcal{T}_s$ to satisfy the following inequality

$$\begin{aligned}
\sum_{t\in\mathcal{T}_c}\sigma_{\text{bias}}^2+\sum_{t\in\mathcal{T}_s}\left(2\kappa_1^2+4\kappa_2^2\right)\leq&\sum_{t\in\mathcal{T}_s}(1-L_f\eta)\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{P}}\left(\boldsymbol{\theta}_t\right)\|^2\right]\\
&+\sum_{t\in\mathcal{T}_c}(1-L_f\gamma)\mathbb{E}\left[\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2\right]
\end{aligned} \tag{18}$$

Let $G_s=\min_{t\in\mathcal{T}_s}\mathbb{E}\left[\|\nabla\mathcal{L}_{\mathcal{P}}(\boldsymbol{\theta}_t)\|^2\right]$, $G_c=\min_{t\in\mathcal{T}_c}\mathbb{E}\left[\left\|\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\boldsymbol{\theta}_t)\right\|^2\right]$. After rearranging Eq.18, we have

$$|\mathcal{T}_c|\geq\frac{T((L_f\eta-1)G_s+2\kappa_1^2+4\kappa_2^2)}{(1-L_f\gamma)G_c-(1-L_f\eta)G_s+2\kappa_1^2+4\kappa_2^2-\sigma_{\text{bias}^2}} \tag{19}$$

Suppose Eq.19 holds, we have

$$\begin{aligned}
\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]\leq&\sum_{t\in\mathcal{T}_c}\left[\frac{2}{\gamma}\left(\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right]-\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right]\right)\right]\\
&+\sum_{t\in\mathcal{T}_s}\left[\frac{2}{\eta}\left(\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right]-\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right]\right)\right]
\end{aligned} \tag{20}$$

Let $\eta=\gamma=\frac{1}{\sqrt{T}}$, we will have

$$\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]\leq\sum_{t\in\mathcal{T}}\left[\frac{2}{\sqrt{T}}\left(\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_t\right)\right]-\mathbb{E}\left[\mathcal{L}\left(\boldsymbol{\theta}_{t+1}\right)\right]\right)\right] \tag{21}$$

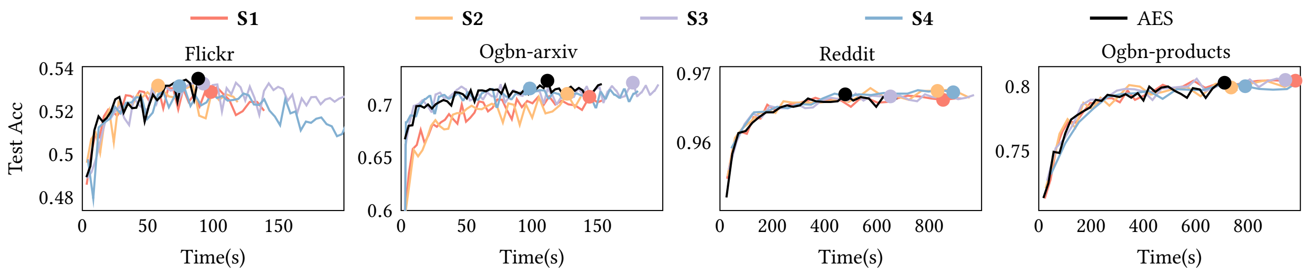
$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]\leq\frac{2}{\sqrt{T}}\left(\mathcal{L}\left(\boldsymbol{\theta}_0\right)-\mathcal{L}\left(\boldsymbol{\theta}^*\right)\right) \tag{22}$$

$$\frac{1}{T}\sum_{t=1}^{T}\mathbb{E}\left[\|\nabla\mathcal{L}(\boldsymbol{\theta}_t)\|^2\right]=O\left(\frac{1}{\sqrt{T}}\right) \tag{23}$$

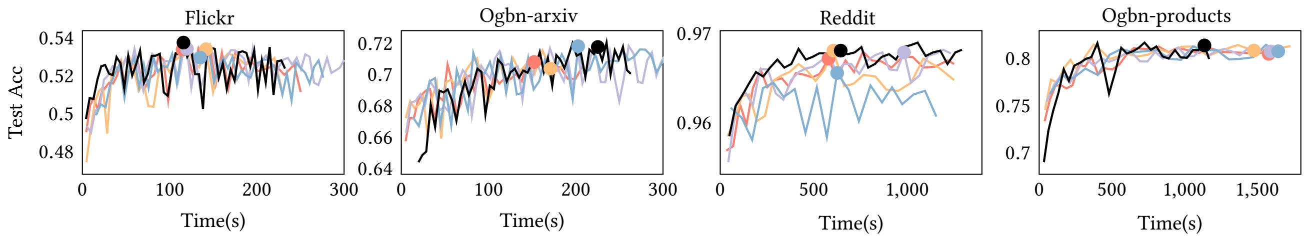
C ADDITIONAL EXPERIMENTAL RESULTS

C.1 Effectiveness of AES

Figure 1 and Figure 2 respectively plot the test accuracy and validation loss on all tested datasets for both models as complements to Figure 9 and Figure 10 in the paper. Consistent with the statement in the paper, AES achieves nearly the optimal trade-off between accuracy and time efficiency in all tested cases. Also, AES shows the fastest convergence among tested correction strategies.

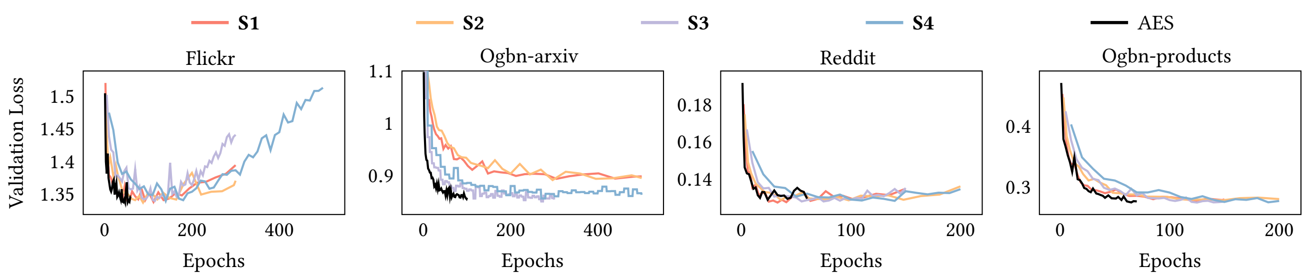


(a) SAGE

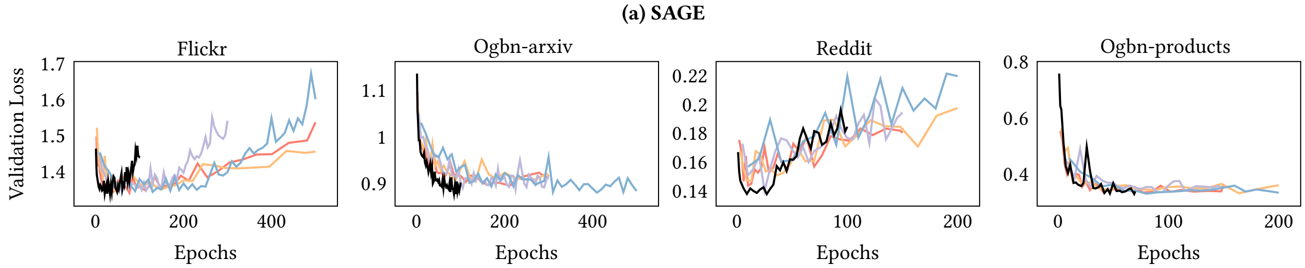


(b) GAT

Figure 1: Test accuracy over time with different correction strategies.



(a) SAGE



(b) GAT

Figure 2: Validation loss over training epochs with different correction strategies.