Supplementary materials

Contents

Contents 1
A Proof of theorems 1
A.1 Proof of theorem 1 1
A.2 Proof of theorem 2 2
B Algorithms of ECHO 2
B.1 VHS 2
B.2 Training with AES 3
C Additional Experimental Setup 3
C.1 Model details 3
C.2 Datasets 4
D Additional experimental results 4
D.1 Delving into the training process 4
D.2 Effectiveness of continuous partitioning. 4
D.3 Effectiveness of AES 4

A PROOF OF THEOREMS

A.1 Proof of theorem 1

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

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

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

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

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

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

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

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

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

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

Organizing the above formula gives us

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

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

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

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

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

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

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

A.2 Proof of theorem 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

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

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

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

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

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

Where (d) is due to Assumption 3.

Combining Eq.10,12,13 gives us

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

Reorganizing the above equation gives us

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

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

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

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

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

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

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

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

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

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

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

B ALGORITHMS OF ECHO

B.1 VHS

Algorithm 1 describes the detailed process for hyperparameter search. For each pair of p and s, we simulate node sets of mini-batches in one epoch by shuffling logical partitions (line 2-5). Then,

the training node label distribution in the simulated data is used to compute the corresponding ϵ and r (line 8-17). Finally, we check whether both conditions **M1** and **M2** hold (line 26). Intuitively, as the granularity of shuffle and sampling becomes finer with larger s and p, the introduced variance tends to decrease and is reflected by r and ϵ. Therefore, VHS starts with the option $s = 1, p = p_0$ for computation efficiency of sampling and expands them in the same proportion with a fixed step size during search (line 21-22).

Algorithm 1: VHS

Input: Input graph node set $\mathcal{V}$, grid search step size z,
Output: partition number p, selection number s

1 **Function** `Simulate(`p, s`)`:
2 $\quad \mathcal{P} \leftarrow \{\mathcal{P}_i = \mathcal{V}[i \times \frac{|\mathcal{V}|}{p} : (i+1) \times \frac{|\mathcal{V}|}{p}] | i \in [o, \cdots, p]\}$
3 $\quad \mathcal{P} \leftarrow$ `Shuffle(`$\mathcal{P}$`)`
4 $\quad \mathcal{B} \leftarrow \{\mathcal{B}_i = \mathcal{P}_{i\times s} \cup \cdots \cup \mathcal{P}_{i\times(s+1)-1} | i \in [0, \cdots, \frac{p}{s})\}$
5 $\quad$ **return** $\mathcal{B}$
6 **end**
7 **Function** `ComputeStandards(`$\mathcal{B}, D_{\mathcal{V}_T}$`)`:
8 $\quad \mathcal{B}_{\text{unskipped}} \leftarrow \emptyset$
9 $\quad$ **for** $\mathcal{B}_i \in \mathcal{B}$ **do**
10 $\quad\quad$ **if** $\mathcal{V}_T \cap \mathcal{B}_i \neq \emptyset$ **then**
11 $\quad\quad\quad \mathcal{B}_{\text{unskipped}} \leftarrow \mathcal{B}_{\text{unskipped}} \cup \mathcal{B}_i$
12 $\quad\quad\quad D_{\mathcal{B}_i} \leftarrow$ `LabelDistribution(`$\mathcal{B}_i$`)`
13 $\quad\quad$ **end**
14 $\quad$ **end**
15 $\quad r \leftarrow \frac{|\mathcal{B}_{\text{unskipped}}|}{|\mathcal{B}|}$
16 $\quad \epsilon \leftarrow \frac{1}{|\mathcal{B}|} \sum_{B\in\mathcal{B}} |D_{\mathcal{V}_T} - D_B|$
17 $\quad$ **return** r, ϵ
18 **end**
19 Random select a small p_0 without causing OOM
20 $D_{\mathcal{V}_T} \leftarrow$ `LabelDistribution(`$\mathcal{V}_T$`)`
21 **for** $l \in [1, 1z, 2z, 3z, \cdots]$ **do**
22 $\quad s \leftarrow 1 \times l, p \leftarrow p_0 \times l$
23 $\quad$ **if** $p > |\mathcal{V}|$ **then break**;
24 $\quad \mathcal{B} \leftarrow$ `Simulate(`p, s`)`
25 $\quad r, \epsilon \leftarrow$ `ComputeStandards(`$\mathcal{B}, D_{\mathcal{V}_T}$`)`
26 $\quad$ **if** $r \geq \mu$ **and** $\epsilon \leq \nu\sqrt{\frac{s}{pN_t}}$ **then return** p, s;
27 **end**

B.2 Training with AES

Algorithm 2 describes the training process with AES. Suppose the total training needs T epochs including SS epochs and correction epochs. At correction epochs, a mini-batch is constructed by neighbor sampling (line 2) and the model is updated by $\tilde{\nabla}\mathcal{L}_{\mathcal{B}(\theta_t)}$ (line 4). At SS epochs, ECHO samples partitions to induce a mini-batch (line 8), and performs full-batch training on the sampled subgraphs. Then, the responding stochastic gradient $\nabla\mathcal{L}_{\mathcal{P}(\theta_t)}$ is used to update models (line 10). At the beginning of training, ECHO executes initial correction epochs (line 14-17) until the approximate loss drop is less than threshold (line 16). Then ECHO switches to execute

Algorithm 2: Training with AES

Input: Graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$, number of epochs T, loss drop threshold ξ, and learning rate η, γ
Output: Model with θ_T

1 **Function** `Correction()`:
2 $\quad$ Construct mini-batch $\mathcal{B}$ by neighbor sampling
3 $\quad$ Compute the approximate training loss $\mathcal{L}_{\mathcal{B}(\theta_t)}$ and stochastic gradient $\tilde{\nabla}\mathcal{L}_{\mathcal{B}(\theta_t)}$
4 $\quad \theta_{t+1} \leftarrow \theta_t - \gamma\tilde{\nabla}\mathcal{L}_{\mathcal{B}}(\theta_t)$
5 $\quad$ **return** $\mathcal{L}_{\mathcal{B}(\theta_t)}$
6 **end**
7 **Function** `SSTraining()`:
8 $\quad$ Construct mini-batch $\mathcal{P}$ by subgraph-wise sampling
9 $\quad$ Compute the approximate training loss $\mathcal{L}_{\mathcal{P}(\theta_t)}$ and stochastic gradient $\nabla\mathcal{L}_{\mathcal{P}(\theta_t)}$
10 $\quad \theta_{t+1} \leftarrow \theta_t - \eta\nabla\mathcal{L}_{\mathcal{P}}(\theta_t)$
11 $\quad$ **return** $\mathcal{L}_{\mathcal{P}(\theta_t)}$
12 **end**
13 $t \leftarrow 0$
14 **while** $t \leftarrow t + 1$ **do**
15 $\quad \mathcal{L}_{\mathcal{B}(\theta_t)} \leftarrow$ `Correction()`
16 $\quad$ **if** $\mathcal{L}_{\mathcal{B}}(\theta_{t-1}) - \mathcal{L}_{\mathcal{B}}(\theta_t) \leq \xi$ **then break**;
17 **end**
18 $t_c \leftarrow t$
19 **for** $t \leftarrow t + 1$ **to** T **do**
20 $\quad$ **if** $t = t_c$ **then** $\mathcal{L}_{\mathcal{B}(\theta_t)} \leftarrow$ `Correction()` ;
21 $\quad$ **else** $\mathcal{L}_{\mathcal{P}(\theta_t)} \leftarrow$ `SSTraining()` ;
22 $\quad$ **if** $t = t_c + 1$ **then**
23 $\quad\quad \tilde{\alpha_t} \leftarrow \left\lceil \frac{\mathcal{L}_{\mathcal{P}}(\theta_t)}{\mathcal{L}_{\mathcal{B}}(\theta_{t-1})} \right\rceil$
24 $\quad\quad t_c \leftarrow t - 1 + \tilde{\alpha_t}$
25 $\quad$ **end**
26 **end**

SS training with periodical correction (line 19-26). ECHO calculates the timing for next correction at every first SS epoch after correction (line 22-25).

C ADDITIONAL EXPERIMENTAL SETUP

C.1 Model details

For SAGE, the layer number is $L = 3$, hidden state dimension is $F = 256$; for GAT, the layer number is $L = 3$, hidden state dimension is $F = 128$, the head number is $H = 4$. For a specific dataset, we adopt the same backbone model implementation across sampling methods[1]. Details of model settings are reported in supplementary material. We summarize the detailed setting of all ECHO and baselines as follows:

- ECHO: For correction epochs, the sampling fan-out is [15,10,5], batch size is 1024. We use $\mu = 0.9$ in VHS condition **M1** to preserve the structural information, and $\xi = 0.1$ in AES's

[1]To ensure test accuracy, We remove BatchNorm for SAGE when training ogbn-arxiv with GraphSAINT and use 2 layer SAGE when training Flickr with GAS.

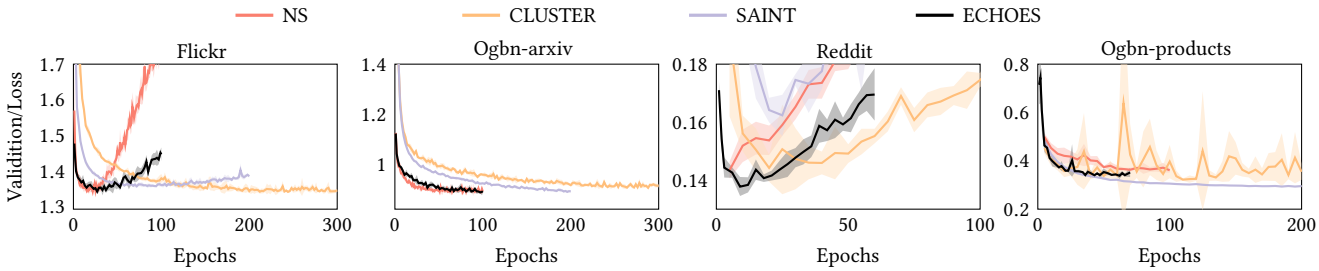


Figure 1: Validation loss over training epochs on GAT.

initial phrase. We set ν to 3 for large datasets and 1 for small ones to avoid OOM.

- NS: Neighbor sampling uses fan-out [15,10,5] and a batch size 1024.
- SAINT: To evaluate GraphSAINT, we use the default implementation for random walk sampler in PyG and refer to the hyperparameter settings in the official implementation[2] and OGB[3].
- CLUSTER: We use the default sampler for Cluster-GCN in PyG. We follow the hyperparameters suggested in its original paper for Reddit and report the best performance under several pairs of hyperparameters for other datasets.
- GAS: We use its official implementation[4] and refer to hyperparameter settings in it.

C.2 Datasets

Table 1 summarizes the statistical characteristics of datasets. Our data split follows the official ones for all four datasets.

Table 1: Dataset statistics

Datasets	#Nodes	#Edges	#Features	#Labels	degree
Flickr	89,250	899,756	500	7	10
Reddit	232,965	114,615,892	602	41	492
Arxiv	169,343	1,166,243	100	40	7
Products	2,449,029	61,859,140	128	47	25

D ADDITIONAL EXPERIMENTAL RESULTS

We provide additional experimental results not reported in paper due to the space limitation.

D.1 Delving into the training process

We plot the convergence on GAT in Figure 1 as a complement. Comparing with Figure 7 in paper, it is also revealed that the impact of subgraph-wise sampling on convergence varies on different models. The convergence of GAT trained using SAINT resembles that uses NS on Ogbn-products, whereas the convergence of SAGE is more affected by introduced variance. ECHO, thanks to the correction strategy AES, achieves convergence similar to NS for all tested datasets on both SAGE and GAT.

D.2 Effectiveness of continuous partitioning.

We perform subgraph-wise sampling training with continuous partitioning and Metis under the same p, s. Table 2 reports the corresponding shuffling error and prediction accuracy. The accuracy of neighbor sampling is tagged in parentheses for each dataset. Under specific p, s, continuous partitioning achieves comparable results as Metis. This demonstrates the ability of continuous partitioning to preserve structural information. Nevertheless, compared with neighbor sampling, both trainings suffer from performance drop, which confirms the residual error analysis in §3.1.

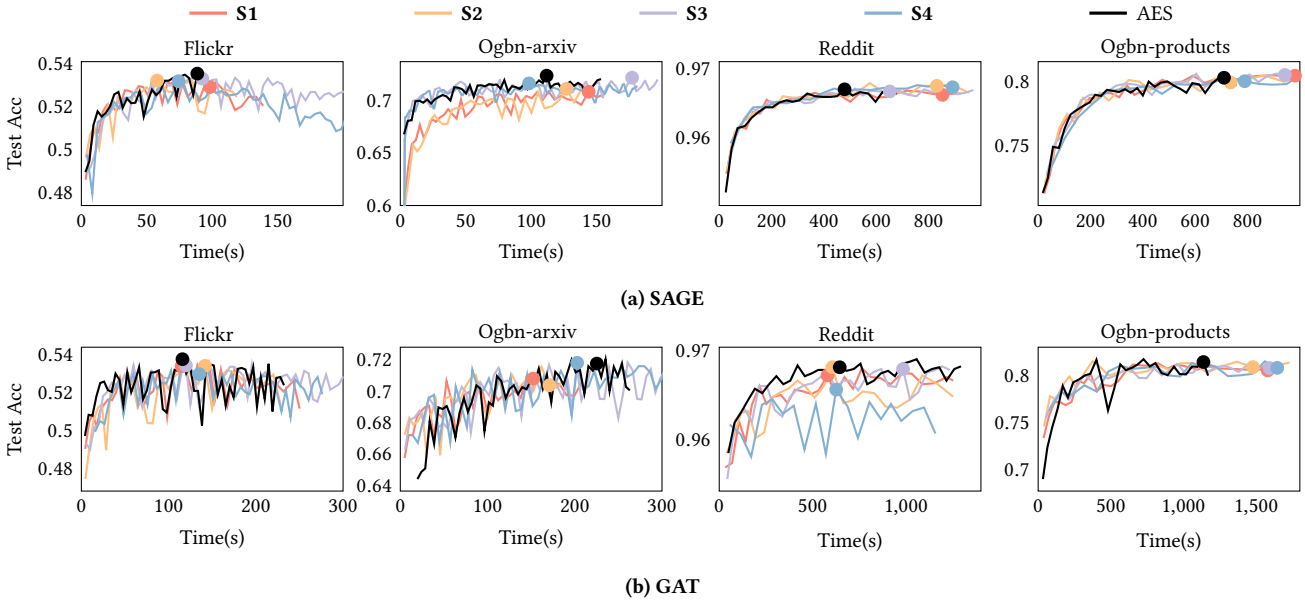
Table 2: Final test accuracy of SAGE under different p, s on Ogbn-arxiv and Ogbn-products with continuous partitioning (Cont.) and Metis (*a.k.a.* CLUSTER).

Dataset	p	s	$\sqrt{\frac{s}{pN_t}}$	Metis		Cont.	
				ϵ	Acc	ϵ	Acc
Arxiv	10	1	0.0010	0.0142	69.58	0.0005	69.50
(71.97)	50	5	0.0010	0.0090	71.10	0.0005	69.50
Products	1000	50	0.0005	0.0122	79.31	0.0004	79.21
(79.33)	10000	500	0.0005	0.0152	79.08	0.0004	79.11

D.3 Effectiveness of AES

Figure 2 and Figure 3 respectively plot the test accuracy and validation loss on all tested datasets for both models as complements to Figure 9 and Figure 10 in paper. Consistent with the statement in paper, AES achieves nearly the optimal trade-off between accuracy and time efficiency on all tested cases.

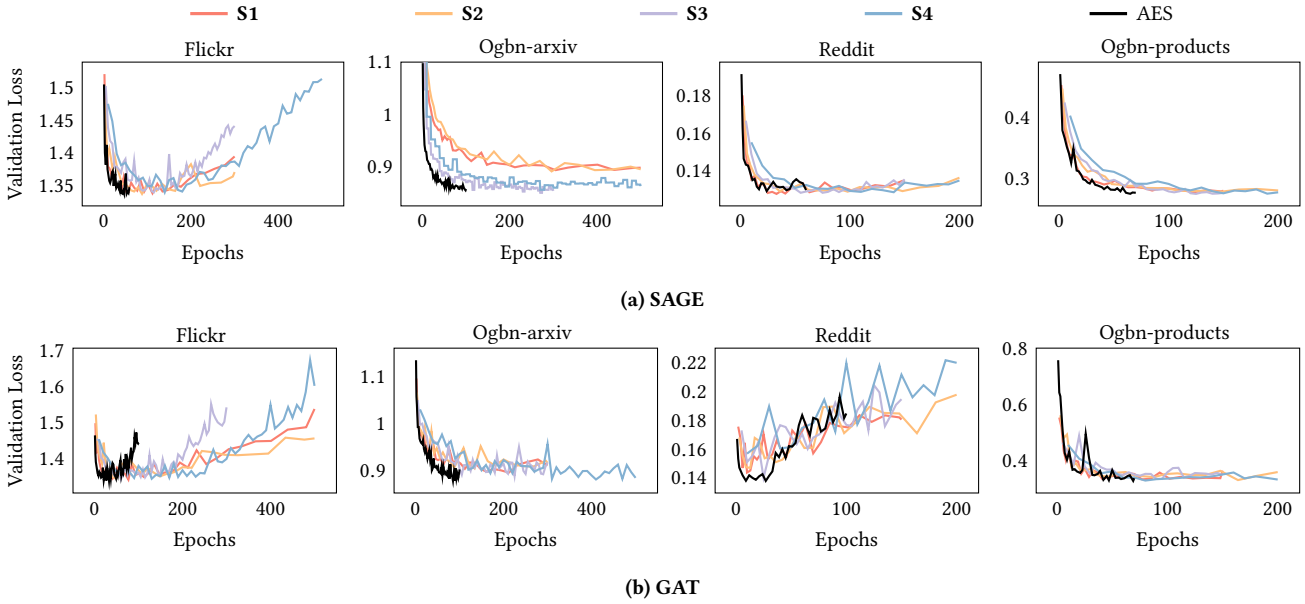
[2] https://github.com/GraphSAINT/GraphSAINT
[3] https://github.com/snap-stanford/ogb
[4] https://github.com/rusty1s/pyg_autoscale



(a) SAGE

(b) GAT

Figure 2: Test accuracy on with different correction strategies.



(a) SAGE

(b) GAT

Figure 3: Validation loss with different correction strategies.