IGLU: EFFICIENT GCN TRAINING VIA LAZY UPDATES

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph Convolution Networks (GCN) are used in numerous settings involving a large underlying graph as well as several layers. Standard SGD-based training scales poorly here since each descent step ends up updating node embeddings for a large portion of the graph. Recent methods attempt to remedy this by sub-sampling the graph which does reduce the compute load, but at the cost of biased gradients which may offer suboptimal performance. In this work we introduce a new method IGLU that caches intermediate computations at various GCN layers. This enables IGLU to perform lazy updates that do not require updating a large number of node embeddings during descent and offers much faster convergence without significantly biasing the gradients. Under standard assumptions such as objective smoothness, IGLU provably converges to a first-order saddle point. We validate IGLU extensively on a variety of benchmarks, where it offers up to 1.2% better accuracy despite requiring up to 88% less wall-clock time.

1 INTRODUCTION

The Graph Convolution Network (GCN) model has received much attention as an effective graph representation learning technique. It can exploit network topology while embedding data points enabling superior performance in several applications such as node classification on graphs (Kipf & Welling, 2017), recommendation systems (Ying et al., 2020) and program repair (Yasunaga & Liang, 2020). Their success notwithstanding, training GCNs at scale remains challenging especially on large and dense graphs and multiple convolution layers. This is mainly due to the aggregation operation that enables GCNs to adapt to graph topology – a node's output layer embedding in a GCN is influenced by embeddings of its neighbors in the previous layer which recursively depend on embeddings of their own neighbors in the still previous layer, and so on.

Even in GCNs with 2-3 layers, when processing loss terms corresponding to a small sample of nodes in a mini-batch, this causes back propagation to update a large multi-hop neighborhood, causing mini-batch SGD-based techniques to scale poorly especially in dense graphs. Efforts to overcome this problem try to limit the number of nodes that receive updates as a result of a back-propagation step Chiang et al. (2019); Hamilton et al. (2018); Zeng et al. (2020). This is done either by sub-sampling the neighborhood or clustering (note the distinction between nodes sampled to create a mini-batch and neighborhood sampling done to limit the neighborhood of the mini-batch that receives updates). Variance reduction techniques have also been studied Chen et al. (2018a) to reduce additional variance introduced by neighborhood sampling. However, these techniques often require large graphs to be heavily subsampled resulting in poor accuracy due to insufficient aggregation. They also do not guarantee unbiased gradients or rigorous convergence guarantees. See Section 2 for a more detailed discussion about state-of-the-art methods for GCN training.

Our Contributions: This paper presents IGLU, an efficient technique for training GCNs based on lazy updates. An analysis of the gradient structure in GCNs reveals the most expensive component of the back-propagation step initiated at a node to be (re-)computation of forward-pass embeddings for its vast multi-hop neighborhood. Based on this observation, IGLU performs back-propagation with significantly reduced complexity using intermediate computations that are cached at regular intervals. This completely avoids neighborhood sampling and is a stark departure from the state-of-the-art. IGLU is architecture-agnostic and can be readily implemented on a wide range of GCN architectures. Avoiding neighborhood sampling also allows IGLU to completely avoid variance artifacts and offer

provable convergence to a first-order stationary point under standard assumptions. In experiments, IGLU offered superior accuracies and accelerated convergence on a range of benchmark datasets.

2 RELATED WORKS

(Bruna et al., 2013; Defferrard et al., 2016; Kipf & Welling, 2017) introduced the GCN architecture for transductive learning on graphs. Later works extended to inductive settings and explored architectural variants such as the GIN (Xu et al., 2019). Much effort has focused on speeding-up GCN training.

Sampling Based Approaches: The simplest *neighborhood sampling* strategy adopted by GraphSAGE (Hamilton et al., 2018) iteratively sub-samples the multi-hop neighborhood of a node before proceeding to initiate back-propagation at that node. Only the sub-sampled neighbors participate in the back-propagation step limiting the amount of computation. *Layer sampling* strategies such as FastGCN (Chen et al., 2018b), LADIES (Zou et al., 2019) and ASGCN (Huang et al., 2018) decrease computation by sampling nodes at each GCN layer, using importance sampling to reduce variance and improve connectivity among sampled nodes. FastGCN used the same sampling distribution for all layers and struggled to maintain connectivity unless large batch-sizes were used. LADIES used a per-layer distribution conditioned on nodes sampled for the layer above allowing more efficient use of samples under a budget. ASGCN used a linear model to jointly infer node importance weights. Recent works such as Cluster-GCN (Chiang et al., 2019) and GraphSAINT (Zeng et al., 2020) propose *subgraph sampling* creating mini-batches out of subgraphs and restricting back-propagation to nodes within the subgraph. To avoid losing too many edges, large mini-batch sizes are used. Cluster-GCN performs graph clustering and chooses multiple clusters per mini-batch while reinserting cross-cluster edges whereas GraphSAINT samples large subgraphs directly using random walks.

Bias and Variance: Sampling techniques face bias as non-linear activations in the GCN architecture make it difficult to offer unbiased estimates of the loss function e.g. Zeng et al. (2020) offer un-biased estimates only if non-linearities are discarded. Sampling techniques also face increased variance for which variance-reduction techniques have been proposed such as VR-GCN (Chen et al., 2018a), MVS-GNN (Cong et al., 2020) and AS-GCN (Huang et al., 2018). VR-GCN samples nodes whose embeddings are to be updated at each layer and uses stale embeddings for the rest and offers variance elimination in the limit under suitable conditions. MVS-GNN handles variance due to mini-batch creation by performing importance weighted sampling to construct mini-batches. Bandit Sampler (Liu et al., 2020) formulates variance reduction as an adversarial bandit problem.

Other Approaches: Recent approaches decouple propagation from prediction as a pre-processing step e.g. PPRGo (Bojchevski et al., 2020), APPNP (Klicpera et al., 2018) and SIGN (Frasca et al., 2020). APPNP makes use of the relationship between the GCNs and PageRank to construct improved propagation schemes via personalized PageRank. PPRGo extends APPNP by approximating the dense propagation matrix via the push-flow algorithm. SIGN proposes inception style pre-computation of graph convolutional filters to speed up training and inference. GNNAutoScale (Fey et al., 2021) builds on VR-GCN and makes use of historical embeddings for scaling GNN training to large graphs.

IGLU in Context of Related Work: IGLU avoids neighborhood sampling entirely and instead speeds-up learning using stale computations. Intermediate computations are cached and lazily updated at regular intervals e.g. once per epoch. We note that IGLU's *caching* is distinct and much more aggressive (e.g. lasting an entire epoch) than the internal caching performed by popular frameworks such as TensorFlow and PyTorch (where caches last only a single iteration). Recomputing these values in bulk offers IGLU economies of scale. IGLU faces zero sampling variance but incurs bias due to the use of stale computations. Fortunately, this bias is provably bounded, and can be made arbitrarily small by adjusting the step length and refresh frequency of the stale computations.

3 IGLU: EFFICIENT GCN TRAINING VIA LAZY UPDATES

Problem Statement: Consider the problem of learning a GCN architecture on an undirected graph $\mathcal{G}(\mathcal{V}, \mathcal{E})$ with each of the N nodes endowed with an *initial* feature vector $\mathbf{x}_i^0 \in \mathbb{R}^{d_0}, i \in \mathcal{V}$. $X^0 \in \mathbb{R}^{n \times d_0}$ denotes the matrix of these initial features stacked together. $\mathcal{N}(i) \subset \mathcal{V}$ denotes the set of neighbors of node i. A denotes the (normalized) adjacency matrix of the graph. A multi-layer GCN architecture uses a parameterized function at each layer to construct a node's embedding for the next

layer using embeddings of that node as well as those of its neighbors. Specifically

$$\mathbf{x}_i^k = f(\mathbf{x}_j^{k-1}, j \in \{i\} \cup \mathcal{N}(i); E^k),$$

where E^k denotes the parameters of k-th layer. For example, a standard GCN layer is given by,

$$\mathbf{x}_i^k = \sigma\left(\sum_{j \in \mathcal{V}} A_{ij} (W^k)^\top \mathbf{x}_j^{k-1}\right),$$

where E^k is simply the matrix $W^k \in \mathbb{R}^{d_{k-1} \times d_k}$ and d_k is the embedding dimensionality after the kth layer. More involved architectures exist that incorporate operations such as layer normalization, batch normalization, etc. In this paper, E^k will always denote the collection of all parameters of the kth layer e.g. offset and scale parameters used in a layer norm operation. $X^k \in \mathbb{R}^{n \times d_k}$ will denote the matrix of kth layer embeddings stacked together, giving us the handy shorthand $X^k = f(X^{k-1}; E^k)$. Given a K-layer GCN and a multi-label/multi-class task with C labels/classes, a fully-connected layer $W^{K+1} \in \mathbb{R}^{d_K \times C}$ and appropriate activation function such as softmax is applied to get predictions which are then fed into the task loss. We note that IGLU does not require the task loss to decompose over the classes. The convergence proofs only require a smooth training objective function.

Neighborhood Explosion: To understand the reason behind neighborhood explosion and cost of mini-batch based SGD training, consider a univariate regression problem with a no-frills 2-layer GCN with unidimensional features and sigmoidal activation within the hidden layers i.e. $K = 2$ and $C = 1 = d_0 = \ldots = d_k$. This GCN is parameterized by $w^1, w^2, w^3 \in \mathbb{R}$ and offers the output $\hat{y}_i = w^3 \sigma\left(z_i^2\right)$ where $z_i^2 = \sum_{j \in \mathcal{V}} A_{ij} w^2 x_j^1 \in \mathbb{R}$. In turn, we have $x_j^1 = \sigma\left(z_i^1\right)$ where $z_i^1 = \sum_{j' \in \mathcal{V}} A_{jj'} w^1 x_{j'}^0 \in \mathbb{R}$ and $x_{j'}^0 \in \mathbb{R}$ are the initial features of the nodes. Given a task loss $\ell : \mathbb{R}^N \times \mathbb{R}^N$ e.g. least squares, denoting $\ell_i' = \ell'(\hat{y}_i, y_i)$ gives us

$$\begin{aligned}
\frac{\partial \ell(\hat{y}_i, y_i)}{\partial w^1} &= \ell_i' \cdot \frac{\partial \hat{y}_i}{\partial z_i^2} \cdot \frac{\partial z_i^2}{\partial w^1} = \ell_i' \cdot w^3 \sigma'(z_i^2) \cdot \sum_{j \in \mathcal{V}} A_{ij} w^2 \frac{\partial x_j^1}{\partial w^1} \\
&= \ell_i' \cdot w^3 \sigma'(z_i^2) \cdot \sum_{j \in \mathcal{V}} A_{ij} w^2 \sigma'(z_j^1) \cdot \sum_{j' \in \mathcal{V}} A_{jj'} x_{j'}^0.
\end{aligned}$$

The nesting of the summations is conspicuous and establishes the neighborhood explosion: when seeking gradients in a K-layer GCN on a graph with average degree m, up to an m^{K-1}-sized neighborhood of a node may be involved in the back-propagation update initiated at that node. Note that the above expression involves terms such as $\sigma'(z_i^2), \sigma'(z_j^1)$ and that the values of z_i^2, z_j^1 etc change whenever the model receives updates. Consequently, for a fresh mini-batch of nodes chosen for a particular iteration, terms such as $\sigma'(z_i^2), \sigma'(z_j^1)$ need to be computed afresh if the gradient is to be computed exactly. Performing these computations amounts to doing *forward pass operations* that frequently involve a large neighborhood of the nodes of the mini-batch. Sampling strategies try to limit this cost by directly restricting the neighborhood over which such forward passes are computed. However, this introduces both bias and variance into the gradient updates as discussed in Section 2. IGLU instead lazily updates various *incomplete gradient* (defined below) and node embedding terms that participate in the above expression. This completely eliminates sampling variance but introduce a bias due to the use of stale terms. However, this bias is not just provably bounded, but can be made arbitrarily small by adjusting the step length and frequency of refreshing these terms.

Lazy Updates for GCN Training: Consider an arbitrary GCN architecture with the following structure: for some parameterized layer functions we have $X^k = f(X^{k-1}; E^k)$ where E^k denotes the collection of all parameters of the kth layer e.g. weight matrices, offset and scale parameters used in layer norm operations, etc. $X^k \in \mathbb{R}^{n \times d_k}$ denotes the matrix of kth layer embeddings stacked together and $X^0 \in \mathbb{R}^{n \times d_0}$ are the initial features. For a K-layer GCN on a multi-label/multi-class task with C labels/classes, a fully-connected layer $W^{K+1} \in \mathbb{R}^{d_K \times C}$ is additionally included to offer $\hat{\mathbf{y}}_i = (W^{K+1})^\top \mathbf{x}_i^K \in \mathbb{R}^C$. Let us use the shorthand $\hat{\mathbf{Y}} \in \mathbb{R}^{N \times C}$ to denote the matrix where the predicted outputs $\hat{\mathbf{y}}_i$ over all the nodes are stacked. We assume a task loss function $\ell : \mathbb{R}^C \times \mathbb{R}^C \times \mathbb{R}_+$ that need not decompose over the classes. We will use the abbreviation $\ell_i := \ell(\hat{y}_i, y_i)$. For sake of simplicity, we assume that the loss function itself includes any activation such as softmax that needs to be applied over the predictions $\hat{\mathbf{y}}_i$. Let $\mathcal{L} = \sum_{i \in \mathcal{V}} \ell_i$ denote the total loss function.

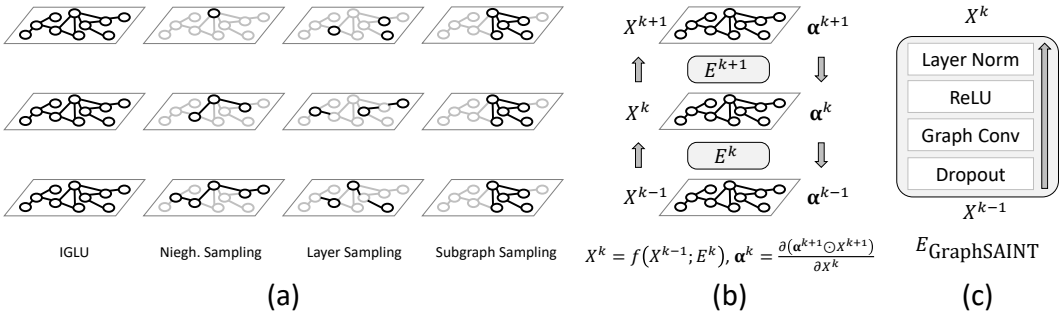


Figure 1: Fig 1(a) highlights the distinction between existing sampling-based approaches which introduce bias and variance in training due to the sampling operations, and IGLU which completely sidesteps these issues, instead being able to scalably execute GCN back-propagation steps on the full graph owing to its use of lazy updates which offer no sampling variance and provably bounded bias. Fig 1(b) summarizes the quantities useful in IGLU's updates. IGLU is architecture-agnostic and can be readily used with wide range of existing architectures. Fig 1(c) gives an example of a GCN layer architecture used by the GraphSAINT method.

Motivation: We define the the *loss derivative* matrix $\mathbf{G} = [g_{ic}] = \mathbb{R}^{N \times C}$ as $g_{ic} := \frac{\partial \ell_i}{\partial \hat{y}_{ic}}$. As the proof of Lemma 1 (See Appendix C) shows, the loss derivative with respect to parameters E^k at any layer has the form $\frac{\partial \mathcal{L}}{\partial E^K} = \sum_{j=1}^{N} \sum_{p=1}^{d_k} \left(\sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \frac{\partial \hat{y}_{ic}}{\partial X_{jp}^k} \right) \frac{\partial X_{jp}^k}{\partial E^K}$. Note that the loss derivative is expressed in terms of partial derivatives $\frac{\partial X_{jp}^k}{\partial E^K}$ that can be computed for any node using only embeddings of its neighbors in the $(k-1)^{\text{th}}$ layer i.e. X^{k-1} thus avoiding any neighborhood explosion. This means that neighborhood explosion must be encountered while computing the terms encapsulated in the round brackets. Let us first formally recognize these terms as *incomplete gradients*. The notation $\left.\frac{\partial P}{\partial Q}\right|_R$ denotes the partial derivative of P w.r.t Q while keeping R fixed i.e. treated as a constant.

Definition 1. *For any layer $k \leq K$, define its incomplete task gradient to be $\boldsymbol{\alpha}^k = [\alpha_{jp}^k] \in \mathbb{R}^{N \times d_k}$,*

$$\alpha_{jp}^k := \left.\frac{\partial(\mathbf{G} \odot \hat{\mathbf{Y}})}{\partial X_{jp}^k}\right|_{\mathbf{G}} = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \frac{\partial \hat{y}_{ic}}{\partial X_{jp}^k}$$

The following lemma completely characterizes the loss gradients of the GCN but also shows that the incomplete gradient terms $\boldsymbol{\alpha}^k, k \in [K]$ can be efficiently computed using a recursive formulation that also does not involve any neighborhood explosion.

Lemma 1. *The following results hold whenever the task loss $\mathcal{L}$ is differentiable:*

1. *For the final fully-connected layer we have $\frac{\partial \mathcal{L}}{\partial W^{K+1}} = (X^K)^\top \mathbf{G}$ as well as for any $k \in [K]$ and any parameter E^k in the kth layer, $\frac{\partial \mathcal{L}}{\partial E^k} = \left.\frac{\partial(\boldsymbol{\alpha}^k \odot X^k)}{\partial E^k}\right|_{\boldsymbol{\alpha}^k} = \sum_{i \in \mathcal{V}} \sum_{p=1}^{d_k} \alpha_{ip}^k \cdot \frac{\partial X_{ip}^k}{\partial E^k}$.*

2. *For the final layer, we have $\boldsymbol{\alpha}^K = \mathbf{G}(W^{K+1})^\top$ as well as for any $k < K$, we have $\boldsymbol{\alpha}^k = \left.\frac{\partial(\boldsymbol{\alpha}^{k+1} \odot X^{k+1})}{\partial X^k}\right|_{\boldsymbol{\alpha}^{k+1}}$ i.e. $\alpha_{jp}^k = \sum_{i \in \mathcal{V}} \sum_{q=1}^{d_{k+1}} \alpha_{iq}^{k+1} \cdot \frac{\partial X_{iq}^{k+1}}{\partial X_{jp}^k}$.*

Lemma 1 establishes a recursive definition of the incomplete gradients using terms such as $\frac{\partial X_{iq}^{k+1}}{\partial X_{jp}^k}$ that only concern a single layer. Thus, computing $\boldsymbol{\alpha}^k$ for any $k \in [K]$ does not involve any neighborhood explosion since only the immediate neighbors of a node need be consulted. Lemma 1 also shows that if $\boldsymbol{\alpha}^k$ are computed and frozen, the loss derivatives $\frac{\partial \mathcal{L}}{\partial E^k}$ only involve additional computation of terms such as $\frac{\partial X_{ip}^k}{\partial E^k}$ which yet again involve a single layer and do not cause neighborhood explosion. This motivates lazy updates to $\boldsymbol{\alpha}^k, X^k$ values in order to accelerate back-propagation. However, performing lazy updates to both $\boldsymbol{\alpha}^k, X^k$ offers suboptimal performance. Hence IGLU

adopts two variants described in Algorithms 1 and 2. The *backprop* variant* keeps embeddings X^k stale for an entire epoch but performs eager updates to $\boldsymbol{\alpha}^k$. The *inverted* variant on the other hand keeps the incomplete gradients $\boldsymbol{\alpha}^k$ stale for an entire epoch but performs eager updates to X^k.

Algorithm 1 IGLU: backprop order

Input: GCN $\mathcal{G}$, initial features X^0, task loss $\mathcal{L}$
1: Initialize model parameters $E^k, k \in [K], W^{K+1}$
2: **while** not converged **do**
3: Do a forward pass to compute X^k for all $k \in [K]$ as well as $\hat{\mathbf{Y}}$
4: Compute $\mathbf{G}$ then $\frac{\partial \mathcal{L}}{\partial W^{K+1}}$ using Lemma 1 (1) and update $W^{K+1} \leftarrow W^{K+1} - \eta \cdot \frac{\partial \mathcal{L}}{\partial W^{K+1}}$
5: Compute $\boldsymbol{\alpha}^K$ using $\mathbf{G}, W^{K+1}$, Lemma 1 (2)
6: **for** $k = K \ldots 1$ **do**
7: Compute $\frac{\partial \mathcal{L}}{\partial E^k}$ using $\boldsymbol{\alpha}^k, X^k$, Lemma 1 (1)
8: Update $E^k \leftarrow E^k - \eta \cdot \frac{\partial \mathcal{L}}{\partial E^k}$
9: Update $\boldsymbol{\alpha}^k$ using $\boldsymbol{\alpha}^{k+1}$ using Lemma 1 (2)
10: **end for**
11: **end while**

Algorithm 2 IGLU: inverted order

Input: GCN $\mathcal{G}$, initial features X^0, task loss $\mathcal{L}$
1: Initialize model parameters $E^k, k \in [K], W^{K+1}$
2: Do an initial forward pass to compute $X^k, k \in [K]$
3: **while** not converged **do**
4: Compute $\hat{\mathbf{Y}}, \mathbf{G}$ and $\boldsymbol{\alpha}^k$ for all $k \in [K]$ using Lemma 1 (2)
5: **for** $k = 1 \ldots K$ **do**
6: Compute $\frac{\partial \mathcal{L}}{\partial E^k}$ using $\boldsymbol{\alpha}^k, X^k$, Lemma 1 (1)
7: Update $E^k \leftarrow E^k - \eta \cdot \frac{\partial \mathcal{L}}{\partial E^k}$
8: Update $X^k \leftarrow f(X^{k-1}; E^k)$
9: **end for**
10: Compute $\frac{\partial \mathcal{L}}{\partial W^{K+1}}$ using Lemma 1 (1) and use it to update $W^{K+1} \leftarrow W^{K+1} - \eta \cdot \frac{\partial \mathcal{L}}{\partial W^{K+1}}$
11: **end while**

SGD Implementation: Update steps in the algorithms (steps 4, 8 in Algorithm 1 and steps 7, 10 in Algorithm 2) are described as a single gradient step over the entire graph to simplify exposition – in practice these steps are implemented using *mini-batch* SGD. This is done as usual by sampling a mini-batch of nodes S and computing task gradients only w.r.t $\hat{\mathcal{L}}_S = \sum_{i \in S} \ell_i$ instead of $\mathcal{L}$.

Contribution: As noted in Section 2, IGLU uses caching in a manner fundamentally different from popular frameworks such as PyTorch or TensorFlow. Their caches are short-lived and always seek to compute exact gradients unlike IGLU that computes gradients much faster but with bounded bias. IGLU also extends other caching-based techniques such as VR-GCN which choose to cache node embeddings. In contrast, IGLU offers two variants and the variant that uses inverted order of updates (Algorithm 2) and caches incomplete gradients outperforms the backprop variant of IGLU (Algorithm 1) that caches node embeddings instead.

Theoretical Analysis: Note that, conditioned on the stale parameters (either $\boldsymbol{\alpha}^k$ or X^k depending on which variant is being executed), the gradients used by IGLU to perform model updates (steps 4, 8 in Algorithm 1 and steps 7, 10 in Algorithm 2) do not exhibit any sampling bias. However, the staleness itself introduces bias in the training process. By controlling the step length η and the frequency with which the stale parameters are updated, this bias can be provably controlled resulting in guaranteed convergence to a first-order stationary point. Due to lack of space, we postpone the detailed statement and proof of the convergence guarantee to Appendix C.

Theorem 2 (IGLU Convergence (Informal))**.** *Suppose the task loss function $\mathcal{L}$ has $\mathcal{O}(1)$-Lipschitz gradients and IGLU is executed with small enough step lengths η with model updates (steps 4, 8 in Algorithm 1 and steps 7, 10 in Algorithm 2) being carried out either in a full-batch or mini-batch SGD manner, then within T iterations, IGLU converges to a model iterate satisfying:*

1. $\|\nabla \mathcal{L}\|_2^2 \leq \mathcal{O}\left(1/T^{\frac{2}{3}}\right)$ *if update steps are carried out on the entire graph in a full-batch.*

2. $\|\nabla \mathcal{L}\|_2^2 \leq \mathcal{O}\left(1/\sqrt{T}\right)$ *if update steps are carried out using mini-batch SGD.*

This result holds under minimal assumptions of objective smoothness and boundedness that are standard Chen et al. (2018a); Cong et al. (2020), yet offers convergence rates comparable to those offered by standard mini-batch SGD. However, whereas works such as (Chen et al., 2018a) assume bounds on the sup-norm i.e. L_∞ norm of the gradients, Theorem 2 only requires an L_2 norm bound. Note that objective smoothness requires the architecture to use smooth activation functions. However, IGLU offers similar performance whether using non-smooth activations e.g. ReLU or smooth ones e.g. GELU (see Appendix B.7) as is also observed by other works (Hendrycks & Gimpel, 2020).

*The backprop variant is named so since it updates model parameters in the order back-propagation would have updated them i.e. W^{K+1} followed by $E^K, E^{K-1}, \ldots$ whereas the inverted variant performs updates in the reverse order i.e. starting from E^1, E^2 all the way to W^{K+1}.

4 EMPIRICAL EVALUATION

IGLU was compared to state-of-the-art (SOTA) baselines on several node classification benchmarks across multiple dimensions associated with GCN training such as accuracy and convergence rate. The inverted order of updates was used for IGLU that is justified empirically.

Datasets: The following five benchmark datasets were used: predicting the communities to which different posts belong in Reddit, classifying protein functions across various biological protein-protein interaction graphs in PPI-Large (Hamilton et al., 2018), categorizing types of images based on descriptions and common properties in Flickr (Zeng et al., 2020), predicting paper-paper associations in OGBN-Arxiv (Hu et al., 2021) and categorizing meaningful associations between proteins in OGBN-Proteins (Hu et al., 2021). Training, validation and test splits were used for each dataset in a manner consistent with the original release of the dataset as were metrics: specifically ROC AUC for OGBN-Proteins and Micro-F1 for all the other datasets. Dataset descriptions and statistics are presented in the Appendix. The graphs considered are varied in size (56K nodes in PPI-Large to 232K nodes in Reddit), density (average degree of 13 in OGBN-Arxiv to 597 in OGBN-Proteins) and number of edges (800K to 39 Million). They are fairly diverse in the nature of information they capture and the type of node classification tasks they solve, covering both multi-label and multi-class problems, thus enabling an extensive evaluation of IGLU.

Baselines: IGLU was compared to state-of-the-art algorithms, namely - GCN (Kipf & Welling, 2017), GraphSAGE (Hamilton et al., 2018), VR-GCN (Chen et al., 2018a), Cluster-GCN (Chiang et al., 2019) and GraphSAINT (Zeng et al., 2020) (using the Random Walk Sampler which was reported by the authors to have the best performance). The implementation of GCN provided by GraphSAGE authors was used since the version released by (Kipf & Welling, 2017) ran into run time error on all datasets. GraphSAGE and VRGCN alleviate the neighborhood explosion problem by sampling subsets of neighbors. ClusterGCN and GraphSAINT are subgraph sampling techniques using sampled subgraphs for constructing convolution layers. Thus, our baselines include neighbor sampling, layer-wise sampling, subgraph sampling and no-sampling methods. IGLU does not require any sampling and hence does not have additional sampling overheads. IGLU was implemented in TensorFlow and compared with the authors' TensorFlow implementations of the baselines. Due to lack of space, comparisons with the following additional baselines is provided in the Appendix B.2: LADIES (Zou et al., 2019), L2-GCN (You et al., 2020), AS-GCN (Huang et al., 2018), MVS-GNN (Cong et al., 2020), FastGCN (Chen et al., 2018b), SIGN (Frasca et al., 2020), PPRGo (Bojchevski et al., 2020) and Bandit Sampler (Liu et al., 2020).

Architecture: We note that all these baselines propose a specific network architecture along with their proposed sampling strategy. We find that these architectures augment the standard GCN architecture (Kipf & Welling, 2017) e.g. multiple intermediate non-linear layers, normalization layers and concatenation layers, all of which can help improve performance. As IGLU is architecture-agnostic, it can be readily used with any such architectures. For the experiments, the network architectures proposed by VR-GCN and GraphSAINT was used with IGLU owing to their consistent performance across datasets as demonstrated by (Chen et al., 2018a; Zeng et al., 2020). Both architectures were considered and results for the best performing architecture are reported for each dataset.

Detailed Setup. The inductive supervised learning setting was considered for all five datasets, motivated by the fact that this setting is more general than the transductive setting which assumes availability of the entire graph during training. This setting also generalizes to the real-world where graphs can grow over time and information for all the nodes may not available (Hamilton et al., 2018). Experiments on PPI-Large, Reddit and Flickr used 2 Layer GCNs for both IGLU and baselines whereas OGBN-Proteins and OGBN-Arxiv used 3 Layer GCNs for all methods as prescribed in the original benchmark (Hu et al., 2021). Further implementation details are provided in Appendix A.1.

Model Selection and Hyperparameter Tuning. Model selection was done for all methods based on their validation set performance. Each experiment was run five times and mean and standard deviation is reported on the test performance in Table 1 along with training times. Although the embedding dimension varies across datasets, they are same for all methods for any dataset. For IGLU, GraphSAGE and VR-GCN an exhaustive grid search was done over general hyperparameters such as batch size, learning rate and dropout rate (Srivastava et al., 2014). In addition method-specific hyperparameter sweeps were also carried out that are detailed in Appendix A.4 due to lack of space.

Table 1: **Accuracy of IGLU compared to SOTA algorithms.** The metric is ROC-AUC for Proteins and Micro-F1 for the others. IGLU is the only method with accuracy within 0.2% of the best accuracy on each dataset with significant speedups in training across datasets. On PPI-Large, IGLU is $\sim 8\times$ faster than VR-GCN, the most accurate baseline. * denotes speedup in initial convergence based on a high validation score of 0.955. – denotes no absolute gain.

Algorithm	PPI-Large	Reddit	Flickr	Proteins	Arxiv
GCN	0.614 ± 0.004	0.931 ± 0.001	0.493 ± 0.002	0.615 ± 0.004	0.657 ± 0.002
GraphSAGE	0.736 ± 0.006	0.954 ± 0.002	0.501 ± 0.013	0.759 ± 0.008	0.682 ± 0.002
VR-GCN	0.975 ± 0.007	0.964 ± 0.001	0.483 ± 0.002	0.752 ± 0.002	0.701 ± 0.006
Cluster-GCN	0.899 ± 0.004	0.962 ± 0.004	0.481 ± 0.005	‖	0.706 ± 0.004
GraphSAINT	0.956 ± 0.003	**0.966** ± 0.003	0.510 ± 0.001	0.764 ± 0.009	0.712 ± 0.006
IGLU	**0.987** ± 0.004	0.964 ± 0.001	**0.515** ± 0.001	**0.784** ± 0.004	**0.718** ± 0.001
Abs. Gain	**0.012**	**–**	**0.005**	**0.020**	**0.006**
% Speedup (1)	**88.12**	**8.1***	**44.74**	**11.05**	**13.94**

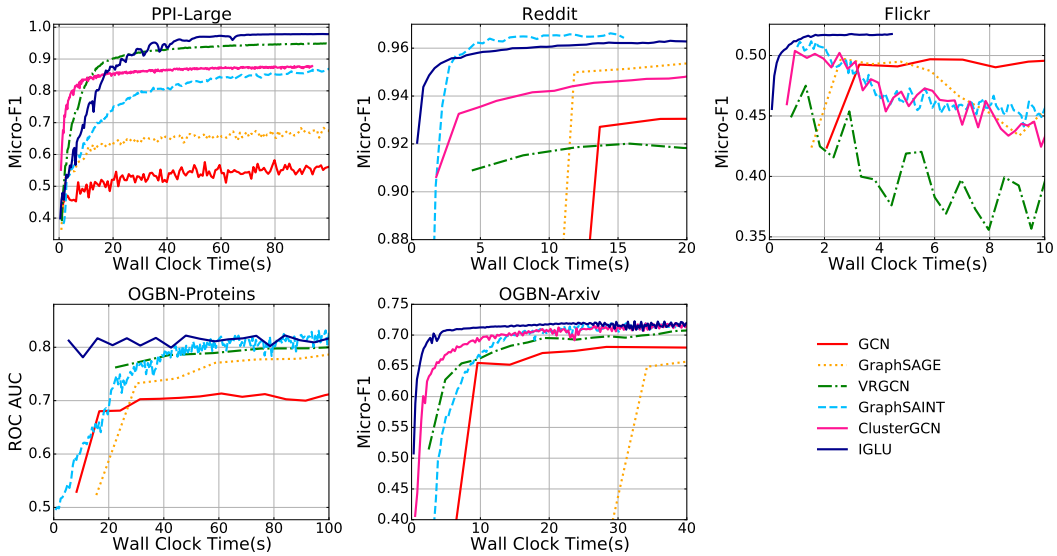


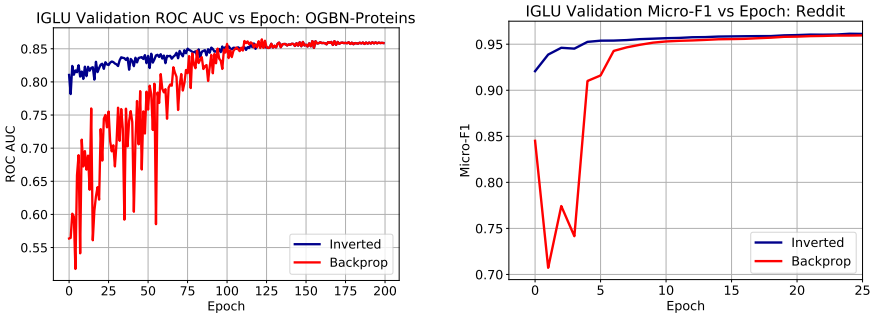
Figure 2: **Wall Clock Time vs Validation Accuracy on different datasets for various methods.** IGLU offers significant improvements in convergence rate over baselines across diverse datasets.

4.1 Results

For IGLU, the VR-GCN architecture performed better for Reddit and OGBN-Arxiv datasets while the GraphSAINT architecture performed better on the remaining three datasets. All baselines were extensively tuned on the 5 datasets and their reported performance was either replicated closely or else improved in some cases. Test accuracies are reported in Table 1 and convergence plots are shown in Figure 2. Additionally, Table 1 also reports the absolute accuracy gain of IGLU over the best baseline. The % speedup offered by IGLU is computed as follows: let the highest validation score obtained by the best baseline be v_1 and t_1 be the time taken to reach that score. Let the time taken by IGLU to reach v_1 be t_2. Then,

$$\% \text{ Speedup} := \frac{t_1 - t_2}{t_1} \times 100 \tag{1}$$

The wall clock training time in Figure 2 is strictly the optimization time for each method and excludes method-specific overheads such as pre-processing, sampling and sub-graph creation that other baselines incur. This is actually a disadvantage to IGLU since its overheads are much smaller. The various time and memory overheads incurred by all methods are summarized in Appendix A.2 and memory requirements for IGLU in Appendix A.3.



(a) (b)

Figure 3: Effect of backpropagation and inverted order of updates in IGLU on Reddit and OGBN-Proteins datasets. Clearly, inverted order of updates offer more stability and faster convergence.

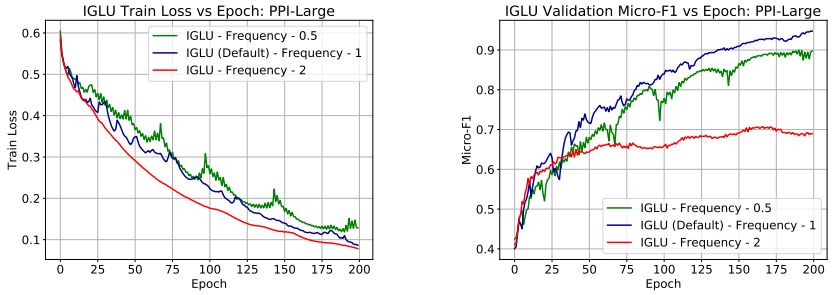
Performance on Test Set and Speedup Obtained: It is evident from Table 1 that IGLU significantly outperforms the baselines on PPI-Large, Flickr, OGBN-Proteins and OGBN-Arxiv and is competitive with best baselines on Reddit. On PPI-Large, IGLU improves accuracy upon the best baseline (VRGCN) by over **1.2**% while providing speedups of upto **88**%, i.e., IGLU is about $8\times$ faster to train than VR-GCN. On OGBN-Proteins, IGLU improves accuracy upon the best baseline (GraphSAINT) by over **2.6**% while providing speedup of **11**%. On Flickr, IGLU offers **0.98**% improvement in accuracy while simultaneously offering upto **45**% speedup over the previous state-of-the-art method GraphSAINT. Similarly on Arxiv, IGLU provides **0.84**% accuracy improvement over the best baseline GraphSAINT while offering nearly **14**% speedup. On Reddit, an **8.1**% speedup was observed in convergence to a high validation score of 0.955 while the final performance is within a standard deviation of the best baseline. The OGBN-Proteins and OGBN-Arxiv datasets are originally benchmarked in the transductive setting, with the entire graph information made available during training. However, we consider the more challenging inductive task yet IGLU outperforms the best transductive baseline by over **0.7**% for Proteins while matching the best transductive baseline for Arxiv (Hu et al., 2021). It is important to note that OGBN-Proteins is an atypical dataset because the graph is highly dense. Because of this, baselines such as ClusterGCN and GraphSAINT that drop a lot of edges while creating subgraphs show a deterioration in performance. ClusterGCN runs into a runtime error on this dataset (denoted by || in the table), while GraphSAINT requires a very large subgraph size to achieve reasonable performance.

Convergence Analysis: Figure 2 shows that IGLU converges faster to a higher validation score than other baselines. For PPI-Large, while Cluster-GCN and VR-GCN show promising convergence in the initial stages of training, they stagnate at a much lower validation score in the end whereas IGLU is able to improve consistently and converge to a much higher validation score. For Reddit, IGLU's final validation score is only marginally lower than GraphSAINT but IGLU offers rapid convergence in the early stages of training. For OGBN-Proteins, Flickr and OGBN-Arxiv, IGLU demonstrates a substantial improvement in both convergence and the final performance on the test set.

4.2 Ablation Studies

Effect of the Order of Updates. IGLU offers the flexibility of using either the backprop order of updates or the inverted order of updates as mentioned in Section 3. Ablations were performed on the Reddit and OGBN-Proteins datasets to analyse the effect of the different orders of updates. Figure 3 offers epoch-wise convergence for the same and shows that the inverted order of updates provides more stability to IGLU as well as improved convergence in the early stages of training. We also observe that both of the orders eventually converge to similar validation scores. These ablations highlight the fact that the inverted order of updates accelerates convergence and justifies our choice of the inverted order of updates for IGLU for all ablation experiments hereon.

Analysis of Degrees of Staleness. The effect of frequency of updates to the incomplete gradients ($\boldsymbol{\alpha}^k$) on the performance and convergence of IGLU was also analyzed. The ablation was conducted keeping all the other hyperparameters fixed. As the default setting all the $\boldsymbol{\alpha}^k$'s were updated only once every epoch and referred to as update with frequency 1. Apart from this scheme, two other



(a) (b)

Figure 4: Update frequency vs Accuracy. Experiments conducted on PPI-Large. As expected, refreshing the $\boldsymbol{\alpha}^k$'s too frequently or too infrequently can affect both performance and convergence.

update schemes are also considered: a) update the $\boldsymbol{\alpha}^k$'s every two epochs, and b) update the $\boldsymbol{\alpha}^k$'s twice within an epoch. We refer to these settings as update with frequency 0.5 and frequency 2 respectively. To summarize, the $\boldsymbol{\alpha}^k$'s are the most fresh with update frequency 2 and most stale update frequency 0.5. This ablation study was performed on the PPI-Large dataset and each variant was trained for 200 epochs. Table 2 summarizes the results doing model selection and Figure 4 plots the convergence of these variants. Figure 4b shows that on PPI-Large, the default update frequency 1 has the best convergence on the validation set, followed by update frequency 0.5. Both update frequency 1 and 0.5 massively outperform update frequency 2. Figure 4a shows that IGLU with update frequency 2 has the lowest training loss but poor validation performance, indicating overfitting to the training dataset. We observe from this ablation that updating embeddings prematurely can cause unstable training resulting in convergence to a suboptimal solution. However, not refreshing the $\boldsymbol{\alpha}^k$'s frequently enough can cause the model to take longer to converge to a good solution.

Table 2: Accuracy vs different update frequency on PPI-Large.

Update Frequency	Train Micro-F1	Validation Micro-F1	Test Micro-F1
0.5	0.947	0.899	0.916
1	0.970	0.947	0.961
2	0.960	0.708	0.726

Additional Ablations and Experiments: Due to lack of space, we provide the following additional ablations and experiments in the Appendix: a) Ablation on degrees of staleness for the backprop order of updates in Appendix B.8, b) Experiments demonstrating IGLU's scalability to deeper networks and larger datasets in Appendices B.3 and B.4 respectively, and c) Experiments demonstrating IGLU's applicability and architecture-agnostic nature in Appendices B.5 and B.6.

5 Discussion and Future Work

We introduced IGLU, a novel method for training GCN architectures that uses biased updates based on cached intermediate computations to speed up gradient descent updates significantly. This bias is shown to be provably bounded so overall convergence is still effective (see Theorem 2). We validated IGLU's performance on several datasets where it significantly outperformed SOTA methods in terms of accuracy and convergence speed. Ablation studies confirmed that IGLU is robust to its few hyperparameters to enable a near-optimal choice. Exploring other possible variants of IGLU, in particular reducing variance due to mini-batch SGD, sampling nodes to further speed-up updates, and exploring alternate staleness schedules are interesting future directions. On a theoretical side, it would be interesting to characterize properties of datasets and loss functions that influence the effect of lazy updates on convergence. Having such a property would allow practitioners to decide whether to execute IGLU with lazier updates or else reduce the amount of staleness. Exploring application- and architecture-specific variants of IGLU is also an interesting direction.

6 REPRODUCIBILITY STATEMENT

We have made efforts to ensure that the results reported in this paper and the submitted material are reproducible and the methodology is easy to follow.

Theoretical Clarity: We discuss the problem setup and preliminaries in detail in Section 3, provide a detailed algorithm for our proposed methodology in Section 4. The discussion seeks to promote a sound theoretical understanding of the problem and our solution. Detailed proofs for all lemmata and theorems are provided in the appendices due to lack of space.

Experimental Reproducibility: Section 5 of the main paper and Appendix B contain information needed to reproduce the empirical results, namely: Datasets, their pre-processing and their source; implementation details for our method and the baselines, specifics regarding the architectures, hyperparameter search spaces and the best hyperparameters corresponding to the results in the paper.

Code Release: We will release the code for IGLU in the discussion forum to the reviewers and AC's during the discussion period. Code for IGLU will also be released publicly upon acceptance of the paper in the spirit of reproducible research.

7 ETHICS STATEMENT

Our work presents a novel and effective mechanism to train GCN architectures on large graphs that outperforms the state of the art in terms of prediction accuracy and convergence speed. We do not consider any sensitive applications in this paper and our experiments focus primarily on publicly available benchmarks of scientific (e.g. PPI) and bibliographic (e.g. ArXiv) nature do not involve any user studies or human experimentation.

REFERENCES

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, Manjunath Kudlur, Josh Levenberg, Rajat Monga, Sherry Moore, Derek G. Murray, Benoit Steiner, Paul Tucker, Vijay Vasudevan, Pete Warden, Martin Wicke, Yuan Yu, and Xiaoqiang Zheng. Tensorflow: A system for large-scale machine learning. In *12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16)*, pp. 265–283, Savannah, GA, November 2016. USENIX Association. ISBN 978-1-931971-33-1. URL https://www.usenix.org/conference/osdi16/technical-sessions/presentation/abadi.

Aleksandar Bojchevski, Johannes Klicpera, Bryan Perozzi, Amol Kapoor, Martin Blais, Benedek Rózemberczki, Michal Lukasik, and Stephan Günnemann. Scaling Graph Neural Networks with Approximate PageRank. *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Jul 2020. doi: 10.1145/3394486.3403296. URL http://dx.doi.org/10.1145/3394486.3403296.

Joan Bruna, Wojciech Zaremba, Arthur Szlam, and Yann LeCun. Spectral Networks and Locally Connected Networks on Graphs, 2013. arXiv:1312.6203 [cs.LG].

Jianfei Chen, Jun Zhu, and Le Song. Stochastic Training of Graph Convolutional Networks with Variance Reduction. In Jennifer Dy and Andreas Krause (eds.), *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pp. 942–950. PMLR, 10–15 Jul 2018a. URL http://proceedings.mlr.press/v80/chen18p.html.

Jie Chen, Tengfei Ma, and Cao Xiao. FastGCN: Fast Learning with Graph Convolutional Networks via Importance Sampling. In *International Conference on Learning Representations*, 2018b.

Wei-Lin Chiang, Xuanqing Liu, Si Si, Yang Li, Samy Bengio, and Cho-Jui Hsieh. Cluster-GCN: An Efficient Algorithm for Training Deep and Large Graph Convolutional Networks. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '19, pp. 257–266, New York, NY, USA, 2019. Association for Computing Machinery.

ISBN 9781450362016. doi: 10.1145/3292500.3330925. URL https://doi.org/10.1145/3292500.3330925.

Cong et al. Minimal variance sampling with provable guarantees for fast training of graph neural networks. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, KDD '20, pp. 1393–1403, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450379984. doi: 10.1145/3394486.3403192. URL https://doi.org/10.1145/3394486.3403192.

Michaël Defferrard, Xavier Bresson, and Pierre Vandergheynst. Convolutional neural networks on graphs with fast localized spectral filtering. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pp. 3844–3852, 2016.

Matthias Fey, Jan E Lenssen, Frank Weichert, and Jure Leskovec. Gnnautoscale: Scalable and expressive graph neural networks via historical embeddings. *arXiv preprint arXiv:2106.05609*, 2021.

Fabrizio Frasca, Emanuele Rossi, Davide Eynard, Ben Chamberlain, Michael Bronstein, and Federico Monti. Sign: Scalable inception graph neural networks. *arXiv preprint arXiv:2004.11198*, 2020.

William L. Hamilton, Rex Ying, and Jure Leskovec. Inductive Representation Learning on Large Graphs, 2018.

Dan Hendrycks and Kevin Gimpel. Gaussian error linear units (gelus), 2020.

Weihua Hu, Matthias Fey, Marinka Zitnik, Yuxiao Dong, Hongyu Ren, Bowen Liu, Michele Catasta, and Jure Leskovec. Open Graph Benchmark: Datasets for Machine Learning on Graphs, 2021.

Wenbing Huang, Tong Zhang, Yu Rong, and Junzhou Huang. Adaptive Sampling Towards Fast Graph Representation Learning, 2018.

Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Thomas N. Kipf and Max Welling. Semi-Supervised Classification with Graph Convolutional Networks, 2017.

Johannes Klicpera, Aleksandar Bojchevski, and Stephan Günnemann. Predict then propagate: Graph neural networks meet personalized pagerank. *arXiv preprint arXiv:1810.05997*, 2018.

Ziqi Liu, Zhengwei Wu, Zhiqiang Zhang, Jun Zhou, Shuang Yang, Le Song, and Yuan Qi. Bandit samplers for training graph neural networks. *arXiv preprint arXiv:2006.05806*, 2020.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, Alban Desmaison, Andreas Köpf, Edward Yang, Zach DeVito, Martin Raison, Alykhan Tejani, Sasank Chilamkurthy, Benoit Steiner, Lu Fang, Junjie Bai, and Soumith Chintala. Pytorch: An imperative style, high-performance deep learning library, 2019.

Nitish Srivastava, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. Dropout: A simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(56):1929–1958, 2014. URL http://jmlr.org/papers/v15/srivastava14a.html.

Petar Veličković, Guillem Cucurull, Arantxa Casanova, Adriana Romero, Pietro Lio, and Yoshua Bengio. Graph attention networks. *arXiv preprint arXiv:1710.10903*, 2017.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How Powerful are Graph Neural Networks? In *International Conference on Learning Representations*, 2019.

Michihiro Yasunaga and Percy Liang. Graph-based, Self-Supervised Program Repair from Diagnostic Feedback. In *International Conference on Machine Learning (ICML)*, 2020.

Rex Ying, Ruining He, Kaifeng Chen, Pong Eksombatchai, William L. Hamilton, and Jure Leskovec. Graph convolutional neural networks for web-scale recommender systems. In *24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD)*, 2020.

Yuning You, Tianlong Chen, Zhangyang Wang, and Yang Shen. L2-GCN: Layer-Wise and Learned Efficient Training of Graph Convolutional Networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2020.

Hanqing Zeng, Hongkuan Zhou, Ajitesh Srivastava, Rajgopal Kannan, and Viktor Prasanna. Graph-SAINT: Graph Sampling Based Inductive Learning Method. In *International Conference on Learning Representations*, 2020. URL `https://openreview.net/forum?id=BJe8pkHFwS`.

Difan Zou, Ziniu Hu, Yewen Wang, Song Jiang, Yizhou Sun, and Quanquan Gu. Layer-Dependent Importance Sampling for Training Deep and Large Graph Convolutional Networks, 2019.

APPENDIX

This appendix is segmented into three key parts.

1. **Section A** discusses additional implementation details. In particular, method-specific overheads are discussed in detail and detailed hyper-parameter settings for IGLU and the main baselines reported in Table 1 are provided. A key outcome of this analysis is that IGLU has the least overheads as compared to the other methods.
2. **Section B** reports dataset statistics, provides comparisons with additional baselines (not included in the main paper due to lack of space) and provides experiments on scaling to deeper models and larger graphs as mentioned in Section 4.2 in the main paper. It also provides experiments for the applicability of IGLU across settings and architectures, using smooth activation functions, continued ablation on degrees of staleness and optimization on the train set. The key outcome of this discussion is that IGLU scales well to deeper models and larger datasets, and continues to give performance boosts with state-of-the-art even when compared to these additional baselines. Experimental evidence also demonstrates IGLU's ability to generalize easily to a different setting and architecture, perform equivalently with smooth activation functions and achieve lower training losses faster compared to all the baselines across datasets.
3. **Section C** gives detailed proofs of Lemma 1 and the convergence rates offered by IGLU. It is shown that under standard assumptions such as objective smoothness, IGLU is able to offer both the standard rate of convergence common for SGD-style algorithms, as well as a *fast* rate if full-batch GD is performed.

A ADDITIONAL IMPLEMENTATION DETAILS

We recall from Section 4.1 that the wall-clock time reported in Figure 2 consists of strictly the optimization time for each method and excludes method-specific overheads. This was actually a disadvantage for IGLU since its overheads are relatively mild. This section demonstrates that other baselines incur much more significant overheads whereas IGLU does not suffer from these large overheads. When included, it further improves the speedups that IGLU provides over baseline methods.

A.1 HARDWARE

We implement IGLU in TensorFlow 1.15.2 and perform all experiments on an NVIDIA V100 GPU (32 GB Memory) and Intel Xeon CPU processor (2.6 GHz). We ensure that all baselines are experimented with the exact same hardware.

A.2 TIME AND MEMORY OVERHEADS FOR VARIOUS METHODS

We consider three main types of overheads that are incurred by different methods. This includes pre-processing overhead that is one-time, recurring overheads and additional memory overhead. We describe each of the overheads in the context of the respective methods below.

1. **GraphSAGE** - GraphSAGE recursively samples neighbors at each layer for every minibatch. This is done on-the-fly and contributes to a significant sampling overhead. Since this overhead is incurred for every minibatch, we categorize this under recurring overhead. We aggregate this overhead across all minibatches during training. GraphSAGE does not incur preprocessing or additional memory overheads.
2. **VRGCN** - Similar to GraphSAGE, VRGCN also recursively samples neighbors at each layer for every minibatch on-the-fly. We again aggregate this overhead across all minibatches during training. VRGCN also stores the stale/historical embeddings that are learnt for every node at every layer. This is an additional overhead of $\mathcal{O}(NKd)$, where K is the number of layers, N is the number of nodes in the graph and $d = \frac{1}{K}\sum_{k\in[K]} d_k$ is the average embedding dimensionality across layers.
3. **ClusterGCN** - ClusterGCN creates subgraphs and uses them as minibatches for training. For the creation of these subgraphs, ClusterGCN performs graph clustering using the highly optimized METIS tool[†]. This overhead is a one-time overhead since graph clustering is done before training and the same subgraphs are (re-)used during the whole training process. We categorize this under preprocessing overhead. ClusterGCN does not incur any recurring or additional memory overheads.
4. **GraphSAINT** - GraphSAINT, similar to ClusterGCN creates subgraphs to be used as mini-batches for training. We categorize this minibatch creation as the preprocessing overhead for GraphSAINT. However, unlike ClusterGCN, GraphSAINT also periodically creates new subgraphs on-the-fly. We categorize this overhead incurred in creating new subgraphs as recurring overhead. GraphSAINT does not incur any additional memory overheads.
5. **IGLU** - IGLU creates mini-batches only once using subsets of nodes with their full neighborhood information which is then reused throughout the training process. In addition to this, IGLU requires initial values of both the incomplete gradients $\boldsymbol{\alpha}^k$ and the X^k embeddings (Step 3 and first part of Step 4 in Algorithm 1 and Step 2 and Step 4 in Algorithm 2) before optimization can commence. We categorize these two overheads - mini-batch creation and initializations of $\boldsymbol{\alpha}^k$'s and X^k embeddings as IGLU's preprocessing overhead and note that **IGLU does not have any recurring overheads**. IGLU does incur an additional memory overhead since it needs to store the incomplete gradients $\boldsymbol{\alpha}^k$'s in the inverted variant and the embeddings X^k in the backprop variant. However, note that the memory occupied by X^k for a layer k is the same as that occupied by $\boldsymbol{\alpha}^k$ for that layer (see Definition 1). Thus, for both its variants, IGLU incurs an additional memory overhead of $\mathcal{O}(NKd)$, where K is the number of layers, N is the number of nodes in the graph and $d = \frac{1}{K}\sum_{k\in[K]} d_k$ is the average embedding dimensionality across layers.

Tables 3 and 4 report the overheads incurred by different methods on the OGBN-Proteins and Reddit datasets (the largest datasets in terms of edges and nodes respectively). ClusterGCN runs into a runtime error on the Proteins dataset **(denoted by ‖ in the table)**. N/A stands for Not Applicable in the tables. In the tables, specifically for IGLU, pre-processing time is the sum of initialization time required to pre-compute $\boldsymbol{\alpha}^k, X^k$ and mini-batch creation time. We also report the individual overhead for both initialization and mini-batch creation for IGLU. The total pre-processing time for IGLU is denoted by T, overheads incurred for initialization by I and overheads incurred for mini-batch creation by MB.

For IGLU, the minibatch creation code is currently implemented in Python, while GraphSAINT uses a highly optimized C++ implementation. Specifically for the Reddit dataset, the number of subgraphs that GraphSAINT samples initially is sufficient and it does not incur any recurring overhead. However, on Proteins, GraphSAINT samples 200 subgraphs every 18 epochs, once the initially sampled subgraphs are used, leading to a sizeable recurring overhead.

[†] `http://glaros.dtc.umn.edu/gkhome/metis/metis/download`

Table 3: **OGBN-Proteins: Overheads of different methods.** IGLU does not have any recurring overhead while VRGCN, GraphSAGE and GraphSAINT all suffer from heavy recurring overheads. ClusterGCN runs into runtime error on this dataset (**denoted by** ‖). GraphSAINT incurs an overhead that is $\sim 2\times$ the overhead incurred by IGLU, while GraphSAGE and VRGCN incur upto $\sim 4.7\times$ and $\sim 7.8\times$ the overhead incurred by IGLU respectively. For the last row, I denotes initialization time, MB denotes minibatch time and T denotes total preprocessing time. Please refer to Discussion on OGBN - Proteins at Section A.2 for more details.

Method	Preprocessing (One-time)	Recurring	Additional Memory
GraphSAGE	N/A	276.8s	N/A
VRGCN	N/A	465.0s	$\mathcal{O}(NKd)$
ClusterGCN	‖	‖	‖
GraphSAINT	22.1s	101.0s	N/A
IGLU	34.0s (I) + 25.0s (MB) = 59.0s (T)	N/A	$\mathcal{O}(NKd)$

Table 4: **Reddit: Overheads of different methods.** IGLU and GraphSAINT do not have any recurring overhead for this dataset while VRGCN and GraphSAGE incur heavy recurring overheads. ClusterGCN suffers from heavy preprocessing overhead incurred due to clustering. In this case, IGLU incurs an overhead that is marginally higher ($\sim 1.4\times$) than that of GraphSAINT, while VRGCN, GraphSAGE and ClusterGCN incur as much as $\sim 2.1\times$, $\sim 4.5\times$ and $\sim 5.8\times$ the overhead incurred by IGLU respectively. For the last line, I denotes initialization time, MB denotes minibatch time and T denotes total preprocessing time. Please refer to Discussion on Reddit at Section A.2 for more details.

Method	Preprocessing (One-time)	Recurring	Additional Memory
GraphSAGE	N/A	41.7s	N/A
VRGCN	N/A	19.2s	$\mathcal{O}(NKd)$
ClusterGCN	54.0s	N/A	N/A
GraphSAINT	6.7s	N/A	N/A
IGLU	3.5s (I) + 5.7s (MB)= 9.2s (T)	N/A	$\mathcal{O}(NKd)$

Discussion on OGBN - Proteins: Figure 5 summarizes the total overheads for all the methods. On the OGBN-Proteins dataset, IGLU incurs $\sim 2\times$ less overhead than GraphSAINT, the best baseline, while incurring as much as $\sim 7.8\times$ less overhead than VRGCN and $\sim 4.7\times$ less overhead than GraphSAGE. It is also important to note that these overheads are often quite significant as compared to the optimization time for the methods and can add to the overall experimentation time. For experiments on the OGBN-Proteins dataset, VRGCN's total overheads equal 46.25% of its optimization time, GraphSAINT's overheads equal 19.63% of its optimization time, GraphSAGE's overhead equal 9.52% of its optimization time. However IGLU's total overheads equal only **5.59%** of its optimization time which is the lowest out of all methods. Upon re-computing the speedup provided by IGLU using the formula defined in equation (1), but this time with overheads included, it was observed that IGLU offered an improved speedup of **16.88%** (the speedup was only 11.05% when considering only optimization time).

Discussion on Reddit: On the Reddit dataset, IGLU incurs upto $\sim 2.1\times$ less overhead than VRGCN and upto $\sim 4.5\times$ less overhead than GraphSAGE. However, IGLU incurs marginally higher overhead ($\sim 1.4\times$) than GraphSAINT. This can be attributed to the non-optimized minibatch creation routine currently used by IGLU compared to a highly optimized and parallelized implementation in C++ used by GraphSAINT. This is an immediate avenue for future work. Nevertheless, VRGCN's total overhead equals 15.41% of its optimization time, GraphSAINT's overhead equals 43.41% of its optimization time, GraphSAGE's overhead equals 31.25% of its optimization time, ClusterGCN's overhead equals 41.02% of its optimization time while IGLU's overhead equals 44.09% of its optimization time. Whereas the relative overheads incurred by IGLU and GraphSAINT in comparison to optimization time may seem high for this dataset, this is because the actual optimization times

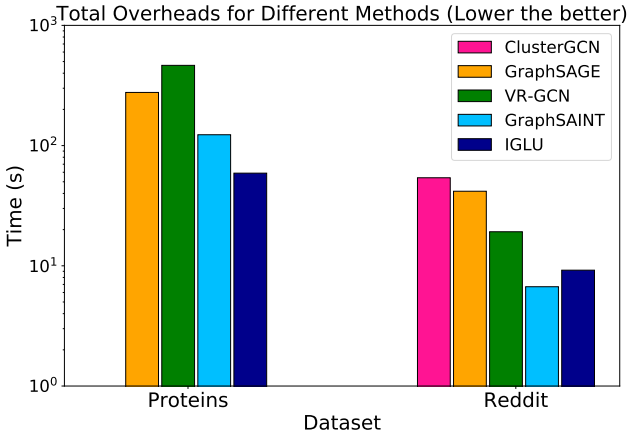


Figure 5: **Total Overheads in Wall Clock Time (Log Scale) for the different methods on OGBN-Proteins and Reddit dataset.** ClusterGCN runs into runtime error on the OGBN-Proteins dataset and hence has not been included in the plot. IGLU frequently offers least total overhead compared to the other methods and hence significantly lower overall experimentation time. Please refer to section A.2 for details.

Table 5: **URL's and commit numbers to run baseline codes**

Method	URL	Commit
GCN	github.com/williamleif/GraphSAGE	a0fdef
GraphSAGE	github.com/williamleif/GraphSAGE	a0fdef
VRGCN	github.com/thu-ml/stochastic_gcn	da7b78
ClusterGCN	github.com/google-research/google-research/tree/master/cluster_gcn	0c1bbe5
AS-GCN	github.com/huangwb/AS-GCN	5436ecd
L2-GCN	github.com/VITA-Group/L2-GCN	687fbae
MVS-GNN	github.com/CongWeilin/mvs_gcn	a29c2c5
LADIES	github.com/acbull/LADIES	c10b526
FastGCN	https://github.com/matenure/FastGCN	b8e6e64
SIGN	https://github.com/twitter-research/sign	42a230c
PPRGo	https://github.com/TUM-DAML/pprgo_pytorch	c92c32e
BanditSampler	https://github.com/xavierzw/gnn-bs	a2415a9

for these methods are rather small, being just 15.43 and 20.86 seconds for GraphSAINT and IGLU respectively in comparison to the other methods such as VRGCN, ClusterGCN and GraphSAGE whose optimization times are 124s, 131s and 133s respectively, almost an order of magnitude larger than that of IGLU.

A.3 Memory Analysis for IGLU

While IGLU requires storing stale variables which can have additional memory costs, for most scenarios with real world graphs, saving these stale representations on modern GPUs are quite reasonable. We provide examples of additional memory usage required for two of the large datasets - Reddit and Proteins in Table 6 and we observe that IGLU requires only **150MB** and **260MB** of additional GPU memory. Even for a graph with 1 million nodes, the additional memory required would only be $\sim$ 2.86GB which easily fit on modern GPUs. For even larger graphs, CPU-GPU interfacing can be used. CPU and GPU interfacing for data movement is a common practice in training machine learning models and hence a potential method to mitigate the issue of limited memory availability in settings with large datasets. This has been explored by many works for dealing with large datasets in the context of GCNs, such as VR-GCN (Chen et al., 2018a) for storing historical activations in main memory (CPU). Such an interfacing is an immediate avenue of future work for IGLU.

Table 6: **Additional Memory Overheads incurred by IGLU on Large datasets**

Dataset	# of Train Nodes	Embedding Dimensions	Number of Layers	Memory Per Layer	Total GPU Memory
Reddit	155k	128	2	75MB	150MB
Proteins	90k	256	3	85MB	260MB
Million Sized Graph	1M	256	3	0.95GB	2.86GB

We observe that IGLU enjoys significant speedups and improvements in training cost across datasets as compared to the baselines as a result of using stale variables. Additionally, since IGLU requires training just a single layer at a time, there is score for further reduction in memory usage by using only the variables required for the current layer and by re-using the computation graphs across layers, and therefore making the IGLU even less memory expensive.

A.4 Hyperparameter Configurations for IGLU and baselines

Table 5 summarizes the source URLs and commit stamps using which baseline methods were obtained for experiments. The Adam optimizer Kingma & Ba (2014) was used to train IGLU and all the baselines until convergence for each of the datasets. A grid search was performed over the other hyper-parameters for each baseline which are summarized below:

1. **GraphSAGE**: Learning Rate - {0.01, 0.001}, Batch Size - {512, 1024, 2048}, Neighborhood Sampling Size - {25, 10, 5}, Aggregator - {Mean, Concat}
2. **VRGCN** : Batch Size - {512, 1000, 2048}, Degree - {1, 5, 10}, Method - {CV, CVD}, Dropout - {0, 0.2, 0.5}
3. **ClusterGCN** : Learning Rate - {0.01, 0.001, 0.005}, Lambda - {-1, 1, 1e-4}, Number of Clusters - {5, 50, 500, 1000, 1500, 5000}, Batch Size - {1, 5, 50, 100, 500}, Dropout - {0, 0.1, 0.3, 0.5}
4. **GraphSAINT-RW** : Aggregator - {Mean, Concat}, Normalization - {Norm, Bias}, Depth - {2, 3, 4}, Root Nodes - {1250, 2000, 3000, 4500, 6000}, Dropout - {0.0, 0.2, 0.5}
5. **IGLU** : Learning Rate - {0.01, 0.001} with learning rate decay schemes, Batch Size - {512, 2048, 4096, 10000}, Dropout - {0.0, 0.2, 0.5, 0.7}

B Dataset Statistics and Additional Experimental Results

B.1 Dataset Statistics

Table 7 provides details on the benchmark node classification datasets used in the experiments. The following five benchmark datasets were used to empirically demonstrate the effectiveness of IGLU: predicting the communities to which different posts belong in Reddit[‡] (Hamilton et al., 2018), classifying protein functions across various biological protein-protein interaction graphs in PPI-Large[§] (Hamilton et al., 2018), categorizing types of images based on descriptions and common properties in Flickr[¶] (Zeng et al., 2020), predicting paper-paper associations in OGBN-Arxiv[‖] (Hu et al., 2021) and categorizing meaningful associations between proteins in OGBN-Proteins[**] (Hu et al., 2021).

B.2 Comparison with additional Baselines

In addition to the baselines mentioned in Table 1, Table 8 compares IGLU to LADIES (Zou et al., 2019), L2-GCN (You et al., 2020), AS-GCN (Huang et al., 2018), MVS-GNN (Cong et al., 2020), FastGCN (Chen et al., 2018b), SIGN (Frasca et al., 2020), PPRGo (Bojchevski et al., 2020) and Bandit Sampler's (Liu et al., 2020) performance on the test set. However, a wall-clock time comparison with

[‡] http://snap.stanford.edu/graphsage/reddit.zip
[§] http://snap.stanford.edu/graphsage/ppi.zip
[¶] https://github.com/GraphSAINT/GraphSAINT - Google Drive Link
[‖] https://ogb.stanford.edu/docs/nodeprop/#ogbn-arxiv
[**] https://ogb.stanford.edu/docs/nodeprop/#ogbn-proteins

Table 7: **Datasets used in experiments along with their statistics.** MC refers to a multi-class problem, whereas ML refers to a multi-label problem.

Dataset	# Nodes	# Edges	Avg. Degree	# Features	# Classes	Train/Val/Test
PPI-Large	56944	818716	14	50	121 (ML)	0.79/0.11/0/10
Reddit	232965	11606919	60	602	41 (MC)	0.66/0.10/0.24
Flickr	89250	899756	10	500	7 (MC)	0.5/0.25/0.25
OGBN-Proteins	132534	39561252	597	8	112 (ML)	0.65/0.16/0.19
OGBN-Arxiv	169343	1166243	13	128	40 (MC)	0.54/0.18/0.28

these methods is not provided since the author implementations of LADIES, L2GCN, MVS-GNN and SIGN are in PyTorch (Paszke et al., 2019) which has been shown to be less efficient than Tensorflow (Chiang et al., 2019; Abadi et al., 2016) for GCN applications. Also, the official AS-GCN, FastGCN and Bandit Sampler implementations released by the authors were for 2 layer models only, whereas some datasets such as Proteins and Arxiv require 3 layer models for experimentation. Attempts to generalize the code to a 3 layer model ran into runtime errors, hence the missing results are denoted by ** in Table 8 and these methods are not considered for a timing analysis. MVS-GNN also runs into a runtime error on the Proteins dataset denoted by ‖. *IGLU continues to significantly outperform all additional baselines on all the datasets.*

Table 8: **Performance on Test Set for IGLU compared to additional algorithms.** The metric is ROC-AUC for Proteins and Micro-F1 for the others IGLU still retains the state-of-the-art results across all datasets even when compared to these new baselines. MVS-GNN ran into runtime error on the Proteins dataset **(denoted by ‖)**. AS-GCN, FastGCN and BanditSampler run into a runtime error on datasets that require more than two layers **(denoted by $**$)**. Please refer to section B.2 for details.

Algorithm	PPI-Large	Reddit	Flickr	Proteins	Arxiv
LADIES	0.548 ± 0.011	0.923 ± 0.008	0.488 ± 0.012	0.636 ± 0.011	0.667 ± 0.002
L2GCN	0.923 ± 0.008	0.938 ± 0.001	0.485 ± 0.001	0.531 ± 0.001	0.656 ± 0.004
ASGCN	0.687 ± 0.001	0.958 ± 0.001	0.504 ± 0.002	**	**
MVS-GNN	0.880 ± 0.001	0.950 ± 0.001	0.507 ± 0.002	‖	0.695 ± 0.003
FastGCN	0.513 ± 0.032	0.924 ± 0.001	0.504 ± 0.001	**	**
SIGN	0.970 ± 0.003	$\mathbf{0.966 \pm 0.003}$	0.510 ± 0.001	0.665 ± 0.008	0.649 ± 0.003
PPRGo	0.626 ± 0.002	0.946 ± 0.001	0.501 ± 0.001	0.659 ± 0.006	0.678 ± 0.003
BanditSampler	0.905 ± 0.003	0.957 ± 0.000	0.513 ± 0.001	**	**
IGLU	$\mathbf{0.987} \pm 0.004$	0.964 ± 0.001	$\mathbf{0.515} \pm 0.001$	$\mathbf{0.784} \pm 0.004$	$\mathbf{0.718} \pm 0.001$

Table 9: **Per epoch time (in seconds) for different methods as the number of layers increase on the OGBN-Proteins dataset.** ClusterGCN ran into a runtime error on this dataset as noted earlier. VRGCN ran into a runtime error for a 4 layer model (**denoted by ‖**). IGLU and GraphSAINT scale almost linearly with the number of layers. It should be noted that these times strictly include only optimization time. GraphSAINT has a much lower per-epoch time than IGLU because of the large sizes of subgraphs per batch ($\sim$ 10000 nodes), while IGLU uses minibatches of size of 512. This results in far less gradient updates within an epoch for GraphSAINT when compared with IGLU, resulting in a much smaller per-epoch time but requiring more epochs overall. Please refer to section B.3.1 for details.

	Number of Layers		
Method	**2**	**3**	**4**
GraphSAGE	2.6	14.5	163.1
VR-GCN	2.3	21.5	‖
GraphSAINT	0.45	0.61	0.76
IGLU	2.97	5.27	6.99

Table 10: **Test Performance (ROC-AUC) at Best Validation for different methods as the number of layers increase on the OGBN-Proteins Dataset.** Results are reported for a single run but trends were observed to remain consistent across repeated runs. VRGCN ran into runtime error for 4 layers (**denoted by** ||). IGLU offers steady increase in performance as the number of layers increase, as well as state-of-the-art performance throughout. GraphSAGE shows a decrease in performance on moving from 3 to 4 layers while GraphSAINT shows only a marginal increase in performance. Please refer to section B.3.2 for details.

	Number of Layers		
Method	**2**	**3**	**4**
GraphSAGE	0.755	0.759	0.742
VR-GCN	0.732	0.749	||
GraphSAINT	0.752	0.764	0.767
IGLU	**0.768**	**0.783**	**0.794**

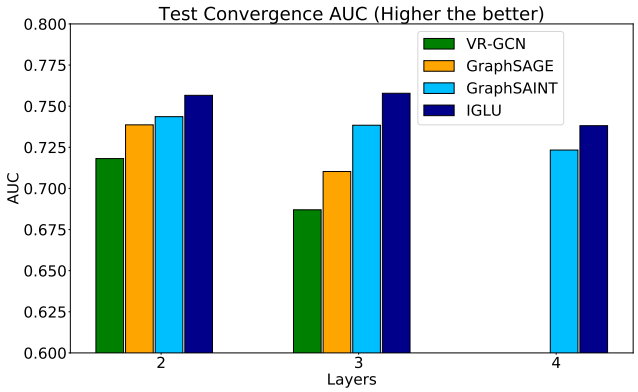


Figure 6: **Test Convergence AUC plots across different number of layers on the OGBN-Proteins dataset.** IGLU has consistently higher AUTC values compared to the other baselines, demonstrating increased stability, faster convergence and better generalization. GraphSAGE suffers from neighborhood explosion problem and the training became very slow as noted earlier. This results in a decrease in the AUTC while going from 3 to 4 layers. GraphSAGE's AUTC for 4 layers is only 0.313, and is thus not visible in the plot. VRGCN also suffers from the neighborhood explosion problem and runs into runtime errors for a 4 layer model. ClusterGCN runs into runtime error for the OGBN-Proteins for all of 2, 3 and 4 layers and is therefore not present in this analysis. Please refer to Section B.3.2 for details.

B.3 Timing Analysis for scaling to more layers

To compare the scalability and performance of different algorithms for deeper models, models with 2, 3 and 4 layers were trained for IGLU and the baseline methods. IGLU was observed to offer a per-epoch time that scaled roughly **linearly with the number of layers** as well as offer the **highest gain in test performance as the number of layers was increased**.

B.3.1 Per Epoch Training Time

Unlike neighbor sampling methods like VRGCN and GraphSAGE, IGLU does not suffer from the neighborhood explosion problem as the number of layers increases, since IGLU updates involve only a single layer at any given time. We note that GraphSAINT and ClusterGCN also do not suffer from the neighborhood explosion problem directly since they both operate by creating GCNs on subgraphs. However, these methods may be compelled to select large subgraphs or else suffer from poor convergence. To demonstrate IGLU's effectiveness in solving the neighborhood explosion problem, the per-epoch training times are summarized as a function of the number of layers in Table

9 on the Proteins dataset. A comparison with ClusterGCN could not be provided as since it ran into runtime errors on this dataset. In Table 9, while going from 2 to 4 layers, GraphSAINT was observed to require $\sim 1.6\times$ more time per epoch while IGLU required $\sim 2.3\times$ more time per epoch, with both methods scaling almost linearly with respect to number of layers as expected. However, GraphSAGE suffered a $\sim 62\times$ increase in time taken per epoch in this case, suffering from the neighborhood explosion problem. VRGCN ran into a run-time error for the 4 layer setting (denoted by || in Table 9). Nevertheless, even while going from 2 layers to 3 layers, VRGCN and GraphSAGE are clearly seen to suffer from the neighborhood explosion problem, resulting in an increase in training time per epoch of almost $\sim 9.4\times$ and $\sim 5.6\times$ respectively.

We note that the times for GraphSAINT in Table 9 are significantly smaller than those of IGLU even though earlier discussion reported IGLU as having the fastest convergence. However, there is no contradiction – GraphSAINT operates with very large subgraphs, with each subgraph having almost 10 % of the nodes of the entire training graph (~ 10000 nodes in a minibatch), while IGLU operates with minibatches of size 512, resulting in IGLU performing a lot more gradient updates within an epoch, as compared with GraphSAINT. Consequently, **IGLU also takes fewer epochs to converge to a better solution than GraphSAINT**, thus compensating for the differences in time taken per epoch.

B.3.2 Test Convergence AUC

This section explores the effect of increasing the number of layers on convergence rate, optimization time, and final accuracy, when scaling to larger number of layers. To jointly estimate the efficiency of a method in terms of wall clock time and test performance achieved, the area under the test convergence plots (AUTC) for various methods was computed. A method that converged rapidly and that too to better performance levels would have a higher AUTC than a method that converges to suboptimal values or else converges very slowly. To fairly time all methods, each method was offered time that was triple of the time it took the best method to reach its highest validation score. Defining the cut-off time this way ensures that methods that may not have rapid early convergence still get a fair chance to improve by having better final performance, while also simultaneously penalizing methods that may have rapid early convergence but poor final performance. We rescale the wall clock time to be between 0 and 1, where 0 refers to the start of training while 1 refers to the cut-off time.

Results: Figure 6 summarizes the AUTC values. IGLU consistently obtains higher AUC values than all methods for all number of layers demonstrating its stability, rapid convergence during early phases of training and ability to generalize better as compared to other baselines. GraphSAGE suffered from neighborhood explosion leading to increased training times and hence decreased AUC values as the number of layers increase. VR-GCN also suffered from the same issue, and additionally ran into a run-time error for 4 layer models.

Test Performance with Increasing Layers: Table 10 summarizes the final test performances for different methods, across different number of layers. The performance for some methods is inconsistent as the depth increases whereas *IGLU consistently outperforms all the baselines in this case as well* with gains in performance as we increase the number of layers hence *making it an attractive technique to train deeper GCN models*.

B.4 Scalability to Larger Datasets: OGBN - Products

To demonstrate the ability of IGLU to scale to very large datasets, we performed experiments on the OGBN-Products dataset (Hu et al., 2021), one of the largest datasets in the OGB collection, with its statistics summarized in Table 11.

Table 11: **Statistics for the OGB-Products datasets**. MC refers to a multi-class problem, whereas ML refers to a multi-label problem.

Dataset	# Nodes	# Edges	Avg. Degree	# Features	# Classes	Train/Val/Test
OGBN-Products	2,449,029	61,859,140	50.5	100	47 (MC)	0.08/0.02/0.90

To demonstrate scalability, we conducted experiments in the transductive setup since this setup involves using the full graph. In addition, this was the original setup in which the dataset was

Table 12: **Performance on the OGBN-Products Test Set for IGLU compared to baseline algorithms.** IGLU outperforms all the baseline methods on this significantly large dataset as well.

Algorithm	Test Micro-F1
GCN	0.760 ± 0.002
GraphSAGE	0.787 ± 0.004
ClusterGCN	0.790 ± 0.003
GraphSAINT	0.791 ± 0.002
SIGN (3,3,0)	0.771 ± 0.001
SIGN (5,3,0)	0.776 ± 0.001
IGLU	**0.793** ± 0.003

benchmarked, therefore allowing for a direct comparison with the baselines (results taken from Table 4 in OGB (Hu et al., 2021) and Table 6 in SIGN (Frasca et al., 2020)). We summarize the performance results in the Table 12, reporting Micro-F1 as the metric. We however do not provide timing comparisons with the baseline methods since IGLU is implemented in TensorFlow while the baselines in the original benchmark Hu et al. (2021) are implemented in PyTorch. This therefore renders a direct comparison of wall-clock time unsuitable. Please refer to Appendix B.2 for more details.

We observe that IGLU is able to scale to the OGBN-Products dataset with over 2.4 million nodes and outperforms all of the baseline methods.

B.5 Applicability of IGLU in the transductive setting: OGBN-Arxiv and OGBN-Proteins

In addition to the results in the inductive setting reported in the main paper, we perform additional experiments on the OGBN-Arxiv and OGBN-Proteins dataset in the transductive setting to demonstrate IGLU's applicability across inductive and transductive tasks and compare the performance of IGLU to that of transductive baseline methods in Table 13 (results taken from OGB (Hu et al., 2021)).

Table 13: **Comparison of IGLU's Test performance with the baseline methods in the transductive setting on the OGBN-Arxiv and OGBN-Proteins datasets.** The metric is Micro-F1 for OGBN-Arxiv and ROC-AUC for OGBN-Proteins.

Algorithm	OGBN-Arxiv	OGBN-Proteins
GCN	0.7174 ± 0.0029	0.7251± 0.0035
GraphSAGE	0.7149 ± 0.0027	0.7768 ± 0.0020
IGLU	**0.7193** ± 0.0018	**0.7840** ± 0.0061

We observe that even in the transductive setting, IGLU outperforms the baseline methods on both the OGBN-Arxiv and OGBN-Proteins datasets.

B.6 Architecture agnostic nature of IGLU

To demonstrate the applicability of IGLU to a wide-variety of architectures, we perform experiments on IGLU with Graph Attention Networks (GAT) (Veličković et al., 2017), GCN (Kipf & Welling, 2017) and GraphSAGE (Hamilton et al., 2018) based architectures and summarize the results for the same in Table 14 and 15 respectively as compared to these baseline methods. We use the Cora, Citeseer and Pubmed datasets as originally used in Veličković et al. (2017) for comparison with the GAT based architecture and the OGBN-Arxiv dataset for comparison with GCN and GraphSAGE based architectures (baseline results as originally reported in (Hu et al., 2021)).

We observe that the IGLU+GAT, IGLU+GCN and IGLU+GraphSAGE outperforms the baseline methods across datasets, thereby demonstrating the architecture agnostic nature of IGLU.

Table 14: **Comparison of IGLU +GAT's test performance with the baseline GAT on different datasets.**

Algorithm	Cora	Citeseer	Pubmed
GAT	0.823 ± 0.007	0.711 ± 0.006	0.786 ± 0.004
IGLU + GAT	$\mathbf{0.829} \pm 0.004$	$\mathbf{0.717} \pm 0.005$	$\mathbf{0.787} \pm 0.002$

Table 15: **Comparison of IGLU's test performance with GCN and GraphSAGE architectures with the baseline methods on the OGBN-Arxiv dataset.**

Algorithm	OGBN-Arxiv
GCN	0.7174 ± 0.0029
GraphSAGE	0.7149 ± 0.0027
IGLU + GCN	$\mathbf{0.7187} \pm 0.0014$
IGLU + GraphSAGE	$\mathbf{0.7155} \pm 0.0032$

B.7 Experiments Using a Smooth Activation Function: GELU

To understand the effect of using non-smooth vs smooth activation functions on IGLU, we perform experiments using the GELU (Hendrycks & Gimpel, 2020) activation function which is a smooth function and in-line with the objective smoothness assumptions made in Theorem 2.

$$\text{GELU}(x) = xP(X \leq x) = x\Phi(x) = x \cdot \frac{1}{2}[1 + \text{erf}(x/\sqrt{2})]$$

We compare the performance of IGLU using ReLU and GELU on all the 5 datasets in the main paper and summarize the results in Table 16.

Table 16: **Effect of Non-smooth vs Smooth activation functions: Test Performance of IGLU on different datasets using ReLU and GELU activation functions.** Metrics are the same for the datasets as reported in Table 1 of the main paper. Results reported are averaged over five different runs.

Dataset	ReLU	GELU
PPI-Large	0.987 ± 0.004	0.987 ± 0.000
Reddit	0.964 ± 0.001	0.962 ± 0.000
Flickr	0.515 ± 0.001	0.516 ± 0.001
Proteins	0.784 ± 0.004	0.782 ± 0.002
Arxiv	0.718 ± 0.001	0.720 ± 0.002

We observe that IGLU is able to enjoy very similar performance across both GELU and ReLU as the activation functions, thereby justifying the practicality of our smoothness assumptions.

B.8 Analysis on Degree of Staleness: Backprop Order of Updates

To understand the effect of more frequent updates in the backprop variant, we performed additional experiments using the backprop variant on the PPI-Large dataset and varied the frequency of updates, the results of which are reported in Table 17 for a single run. We fix the hyperparameters and train for 200 epochs.

Table 17: **Accuracy vs different update frequency on PPI-Large: Backprop Order**

Update Frequency	Train Micro-F1	Validation Micro-F1	Test Micro-F1
0.5	0.761	0.739	0.756
1	0.805	0.778	0.796
2	0.794	0.769	0.784

We observe that more frequent updates help stabilize training better. We also observe that update frequencies 1 and 2 perform competitively, and both significantly outperform update frequency 0.5.

However we note that with higher update frequency, we incur an additional computational cost since we need to re-compute embeddings more frequently. We believe that both improved stability and competitive performance can be attributed to the fresher embeddings, which are otherwise kept stale

within an epoch in this order of updates. The experiment with frequency 0.5 has a slower convergence and comparatively poor performance as expected.

B.9 CONVERGENCE - TRAIN LOSS VS WALL CLOCK TIME

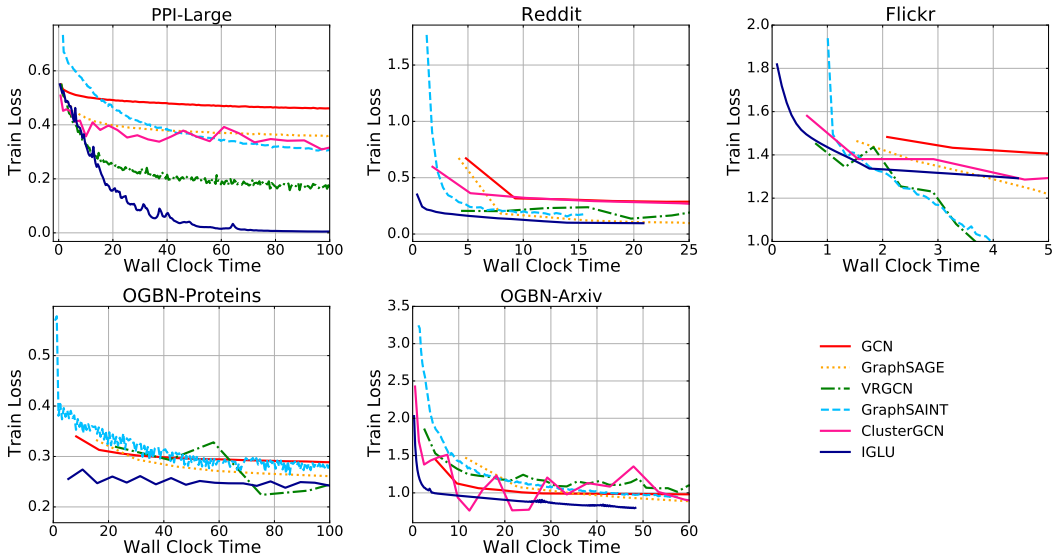


Figure 7: **Training Loss curves of different methods on the benchmark datasets against Wall clock time.**

Figure 7 provides train loss curves for all datasets and methods in Table 1. IGLU is able to achieve a lower training loss faster compared to all the baselines across datasets.

C THEORETICAL PROOFS

In this section we provide a formal restatement of Theorem 2 as well as its proof. We also provide the proof of Lemma 1 below.

C.1 PROOF FOR LEMMA 1

Proof of Lemma 1 We consider the two parts of Lemma 1 separately and consider two cases while proving each part. For each part, Case 1 pertains to the final layer and Case 2 considers intermediate layers.

Clarification about some Notation in the statements of Definition 1 and Lemma 1: The notation $\left.\frac{\partial(G\odot\hat{Y})}{\partial X^k_{jp}}\right|_G$ is meant to denote the partial derivative w.r.t X^k_{jp} but while keeping G fixed i.e. treated as a constant or being "conditioned upon" (indeed both G and $\hat{Y}$ depend on X^k_{jp} but the definition keeps G fixed while taking derivatives). Similarly, in Lemma 1 part 1, α^k is fixed (treated as a constant) in the derivative in the definition of $\partial\mathcal{L}/\partial E^k$ and in part 2, α^{k+1} is fixed (treated as a constant) in the derivative in the definition of α^k.

Recalling some Notation for sake of completeness: We recall that x^k_i is the k-th layer embedding of node i. y_i is the C-dimensional ground-truth label vector for node i and $\hat{y}_i$ is the C-dimensional predicted score vector for node i. K is total number of layers in the GCN. $\mathcal{L}_i$ denotes the loss on the i-th node and $\mathcal{L}$ denotes the total loss summed over all nodes. d_k is the embedding dimensionality after the k-th layer.

Proof of Lemma 1.1 We analyze two cases

Case 1 ($k = K$ i.e. final layer): Recall that the predictions for node i are obtained as $\hat{y}_i = (W^{K+1})^\top x_i^K$. Thus we have

$$\frac{\partial \mathcal{L}}{\partial W^{K+1}} = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} \frac{\partial \mathcal{L}_i}{\partial \hat{y}_{ic}} \frac{\partial \hat{y}_{ic}}{\partial W^{K+1}}.$$

Now, $\frac{\partial \mathcal{L}_i}{\partial \hat{y}_{ic}} = g_{ic}$ by definition. If we let w_c^{K+1} denote the c-th column of the $d_k \times C$ matrix W^{K+1}, then it is clear that $\hat{y}_{ic}$ depends only on w_c^{K+1} and x_i^K. Thus, we have

$$\frac{\partial \hat{y}_{ic}}{\partial W^{K+1}} = \frac{\partial \hat{y}_{ic}}{\partial w_c^{K+1}} \cdot e_c^\top = (x_i^K)^\top e_c^\top,$$

where e_c is the c-th canonical vector in C-dimensions with 1 at the c-th coordinate and 0 everywhere else. This gives us

$$\frac{\partial \mathcal{L}}{\partial W^{K+1}} = \sum_{i \in \mathcal{V}} (x_i^K)^\top \sum_{c \in [C]} g_{ic} \cdot e_c^\top = (X^K)^\top G$$

Case 2 ($k < K$ i.e. intermediate layers): We recall that X^k stacks all k-th layer embeddings as an $N \times d_k$ matrix and $X^k = f(X^{k-1}; E^k)$ where E^k denotes the parameters (weights, offsets, scales) of the k-th layer. Thus we have

$$\frac{\partial \mathcal{L}}{\partial E^K} = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} \frac{\partial \mathcal{L}}{\partial \hat{y}_{ic}} \frac{\partial \hat{y}_{ic}}{\partial E^K}.$$

As before, $\frac{\partial \mathcal{L}}{\partial \hat{y}_{ic}} = g_{ic}$ by definition and we have

$$\frac{\partial \hat{y}_{ic}}{\partial E^K} = \sum_{j \in \mathcal{V}} \sum_{p=1}^{d_k} \frac{\partial \hat{y}_{ic}}{\partial X_{jp}^k} \frac{\partial X_{jp}^k}{\partial E^K}$$

This gives us

$$\frac{\partial \mathcal{L}}{\partial E^K} = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \sum_{j \in \mathcal{V}} \sum_{p=1}^{d_k} \frac{\partial \hat{y}_{ic}}{\partial X_{jp}^k} \frac{\partial X_{jp}^k}{\partial E^K} = \sum_{j \in \mathcal{V}} \sum_{p=1}^{d_k} \alpha_{jp}^k \cdot \frac{\partial X_{jp}^k}{\partial E^K} = \left. \frac{\partial (\alpha^k \odot X^k)}{\partial E^k} \right|_{\alpha^k},$$

where we used the definition of α_{jp}^k in the second step and used a "conditional" notation to get the third step. We reiterate that $\left. \frac{\partial (\alpha^k \odot X^k)}{\partial E^k} \right|_{\alpha^k}$ implies that α^k is fixed (treated as a constant) while taking the derivative. This "conditioning" is critical since α^k also depends on E^k. This concludes the proof.

Proof of Lemma 1.2: We consider two cases yet again and use Definition 1 that tells us that

$$\alpha_{jp}^k = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \frac{\partial \hat{y}_{ic}}{\partial X_{jp}^k} = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \frac{\partial \hat{y}_{ic}}{\partial X_{jp}^k}$$

Case 1 ($k = K$): Since $\hat{y}_i = (W^{K+1})^\top x_i^K$, we know that $\hat{y}_{ic}$ depends only on x_i^K and w_c^{K+1} where as before, w_c^{K+1} is the c-th column of the matrix W^{K+1}. This gives us $\frac{\partial \hat{y}_{ic}}{\partial X_{jp}^K} = 0$ if $i \neq j$ and $\frac{\partial \hat{y}_{ic}}{\partial X_{jp}^K} = w_{pc}^K$ if $i = j$ where w_{pc}^K is the (p, c)-th entry of the matrix W^{K+1} (or in other words, the p-th coordinate of the vector w_c^{K+1}). This tells us that

$$\alpha_{jp}^K = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \frac{\partial \hat{y}_{ic}}{\partial X_{jp}^k} = \sum_{c \in [C]} g_{jc} \cdot w_{pc}^K,$$

which gives us $\alpha^K = G(W^{K+1})^\top$.

Case 2 ($k < K$): By Definition 1 we have

$$\alpha_{jp}^k = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \frac{\partial \hat{y}_{ic}}{\partial X_{jp}^k} = \sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \sum_{l \in \mathcal{V}} \sum_{q=1}^{d_{k+1}} \frac{\partial \hat{y}_{ic}}{\partial X_{lq}^{k+1}} \frac{\partial X_{lq}^{k+1}}{\partial X_{jp}^k}$$

Rearranging the terms gives us

$$\alpha_{jp}^{k} = \sum_{l \in \mathcal{V}} \sum_{q=1}^{d_{k+1}} \left(\sum_{i \in \mathcal{V}} \sum_{c \in [C]} g_{ic} \cdot \frac{\partial \hat{y}_{ic}}{\partial X_{lq}^{k+1}} \right) \cdot \frac{\partial X_{lq}^{k+1}}{\partial X_{jp}^{k}} = \sum_{l \in \mathcal{V}} \sum_{q=1}^{d_{k+1}} \alpha_{lq}^{k+1} \cdot \frac{\partial X_{lq}^{k+1}}{\partial X_{jp}^{k}},$$

where we simply used Definition 1 in the second step. However, the resulting term is simply $\left.\frac{\partial(\alpha^{k+1} \odot X^{k+1})}{\partial X_{jp}^{k}}\right|_{\alpha^{k+1}}$ which conditions on, or treats as a constant, the term α^{k+1} according to our notational convention. This finishes the proof of part 2.

C.2 Statement of Convergence Guarantee

The rate for full-batch updates, as derived below, is $\mathcal{O}\left(\frac{1}{T^{\frac{2}{3}}}\right)$. This *fast* rate offered by full-batch updates is asymptotically superior to the $\mathcal{O}\left(\frac{1}{\sqrt{T}}\right)$ rate offered by mini-batch SGD updates. This is due to the additional variance due to mini-batch construction that mini-batch SGD variants have to incur.

Theorem 3 (IGLU Convergence (Final))**.** *Suppose the task loss function $\mathcal{L}$ has H-smooth and an architecture that offers bounded gradients and Lipschitz gradients as quantified below, then if IGLU in its inverted variant (Algorithm 2) is executed with step length η and a staleness count of τ updates per layer as in steps 7, 10 in Algorithm 2, then within T iterations, we must have*

1. $\|\nabla \mathcal{L}\|_2^2 \leq \mathcal{O}\left(1/T^{\frac{2}{3}}\right)$ *if model update steps are carried out on the entire graph in a full-batch with step length* $\eta = \mathcal{O}\left(1/T^{\frac{1}{3}}\right)$ *and* $\tau = \mathcal{O}(1)$.

2. $\|\nabla \mathcal{L}\|_2^2 \leq \mathcal{O}\left(1/\sqrt{T}\right)$ *if model update steps are carried out using mini-batch SGD with step length* $\eta = \mathcal{O}\left(1/\sqrt{T}\right)$ *and* $\tau = \mathcal{O}\left(T^{\frac{1}{4}}\right)$.

It is curious that the above result predicts that when using mini-batch SGD, a non-trivial amount of staleness ($\tau = \mathcal{O}\left(T^{\frac{1}{4}}\right)$ as per the above result) may be optimal and premature refreshing of embeddings/incomplete gradients may be suboptimal as was also seen in experiments reported in Table 2 and Figure 4. Our overall proof strategy is the following

1. **Step 1**: Analyze how lazy updates in IGLU affect model gradients
2. **Step 2**: Bound the bias in model gradients in terms of staleness due to the lazy updates
3. **Step 3**: Using various properties such as smoothness and boundedness of gradients, obtain an upper bound for the bias in the gradients in terms of number of iterations since last update
4. **Step 4**: Use the above to establish the convergence guarantee

We will show the results for the variant of IGLU that uses the *inverted* order of updates as given in Algorithm 2. A similar proof technique will also work for the variant that uses the *backprop* order of updates. Also, to avoid clutter, we will from hereon assume that a normalized total loss function is used for training i.e. $\mathcal{L} = \frac{1}{N} \sum_{i \in \mathcal{V}} \ell_i$ where $N = |\mathcal{V}|$ is the number of nodes in the training graph.

C.3 Step 1: Partial Staleness and its Effect on Model Gradients

A peculiarity of the inverted order of updates is that the embeddings $X^k, k \in [K]$ are never stale in this variant. To see this, we use a simple inductive argument. The base case of X^0 is obvious – it is never stale since it is never meant to be updated. For the inductive case, notice how, the moment any parameter E^k is updated in step 7 of Algorithm 2 (whether by mini-batch SGD or by full-batch GD), immediately thereafter in step 8 of the algorithm, X^k is updated using the current value of X^{k-1} and E^k. Since by induction X^{k-1} never has a stale value, this shows that X^k is never stale either, completing the inductive argument.

This has an interesting consequence: by Lemma 1, we have $\frac{\partial \mathcal{L}}{\partial E^k} = \frac{1}{N} \left.\frac{\partial(\boldsymbol{\alpha}^k \odot X^k)}{\partial E^k}\right|_{\boldsymbol{\alpha}^k}$ (notice the additional $1/N$ term since we are using a normalized total loss function now). However, as $\left.\frac{\partial(\boldsymbol{\alpha}^k \odot X^k)}{\partial E^k}\right|_{\boldsymbol{\alpha}^k}$ is completely defined given $E^k, \boldsymbol{\alpha}^k$ and X^{k-1} and by the above argument, X^{k-1} never has a stale value. This shows that the only source of staleness in $\frac{\partial \mathcal{L}}{\partial E^k}$ is the staleness in values of the incomplete task gradient $\boldsymbol{\alpha}^k$. Similarly, it is easy to see that the only source of staleness in $\frac{\partial \mathcal{L}}{\partial W^{K+1}} = (X^K)^\top \mathbf{G}$ is the staleness in $\mathbf{G}$.

The above argument is easily mirrored for the backprop order of updates where an inductive argument similar to the one used to argue above that X^k values are never stale in the inverted update variant, would show that the incomplete task gradient $\boldsymbol{\alpha}^k$ values are never stale in the backprop variant and the only source of staleness in $\frac{\partial \mathcal{L}}{\partial E^k}$ would then be the staleness in the X^k values.

C.4 Step 2: Relating the Bias in Model Gradients to Staleness

The above argument allows us to bound the bias in model gradients as a result of lazy updates. To avoid clutter, we will present the arguments with respect to the E^k parameter. Similar arguments would hold for the W^{K+1} parameter as well. Let $\tilde{\boldsymbol{\alpha}}^k, \boldsymbol{\alpha}^k$ denote the stale and actual values of the incomplete task gradient relevant to E^k. Let $\frac{\partial \widetilde{\mathcal{L}}}{\partial E^k} = \left.\frac{\partial(\tilde{\boldsymbol{\alpha}}^k \odot X^k)}{\partial E^k}\right|_{\tilde{\boldsymbol{\alpha}}^k}$ be the stale gradients used by IGLU in its inverted variant to update E^k and similarly let $\frac{\partial \mathcal{L}}{\partial E^k} = \left.\frac{\partial(\boldsymbol{\alpha}^k \odot X^k)}{\partial E^k}\right|_{\boldsymbol{\alpha}^k}$ be the true gradient that could have been used had there been no staleness.

We will abuse notation and let the vectorized forms of these incomplete task gradients be denoted by the same symbols i.e. we stretch the matrix $\boldsymbol{\alpha}^k \in \mathbb{R}^{N \times d_k}$ into a long vector denoted also by $\boldsymbol{\alpha}^k \in \mathbb{R}^{N \cdot d_k}$. Let $\dim(E^k)$ denote the number of dimensions in the model parameter E^k (recall that E^k can be a stand-in for weight matrices, layer norm parameters etc used in layer k of the GCN). Similarly, we let $Z^k_{jp} \in \mathbb{R}^{\dim(E^k)}, j \in [N], p \in [d_k]$ denote the vectorized form of the gradient $\frac{\partial X^k_{ip}}{\partial E^k}$ and let $Z^k \in \mathbb{R}^{N \cdot d_k \times \dim(E^k)}$ denote the matrix with all these vectors Z^k_{jp} stacked up.

As per the above notation, it is easy to see that the vectorized form of the model gradient is given by

$$\frac{\partial \mathcal{L}}{\partial E^k} = \frac{(Z^k)^\top \boldsymbol{\alpha}^k}{N} \in \mathbb{R}^{\dim(E^k)}$$

The $1/N$ term appears since we are using a normalized total loss function. This also tells us that

$$\left\| \frac{\partial \widetilde{\mathcal{L}}}{\partial E^k} - \frac{\partial \mathcal{L}}{\partial E^k} \right\|_2 = \frac{\sqrt{(\tilde{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k)^\top (Z^k (Z^k)^\top)(\tilde{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k)}}{N} \leq \frac{\left\| \tilde{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k \right\|_2 \cdot \sigma_{\max}(Z^k)}{N}, \tag{2}$$

where $\sigma_{\max}(Z^k)$ is the largest singular value of the matrix Z^k.

C.5 Step 3: Smoothness and Bounded Gradients to Bound Gradient Bias

The above discussion shows how to bound the bias in gradients in terms of staleness in the incomplete task gradients. However, to utilize this relation, we assume that the loss function $\mathcal{L}$ is H-smooth which is a standard assumption in literature. We will also assume that the network offers bounded gradients. Specifically, for all values of model parameters $E^k, k \in [K], W^{K+1}$ we have $\|\mathbf{G}\|_2, \left\| \frac{\partial X^k_{ip}}{\partial E^k} \right\|_2, \left\| \frac{\partial \mathcal{L}}{\partial E^k} \right\|_2, \left\| \frac{\partial X^{k+1}}{\partial X^k} \right\|_2, \|\boldsymbol{\alpha}^k\|_2 \leq B$. For sake of simplicity, we will also assume the same bound on parameters e.g. $\left\|\mathbf{W}^K\right\|_2 \leq B$. Assuming bounded gradients and bounded parameters is also standard in literature. However, whereas works such as Chen et al. (2018a) assume bounds on the sup-norm i.e. L_∞ norm of the gradients, our proofs only require an L_2 norm bound.

We will now show that if the model parameters $\tilde{E}^k, k \in [K], \tilde{W}^{K+1}$ undergo gradient updates to their new values $E^k, k \in [K], W^{K+1}$ and the amount of *travel* is bounded by $r > 0$ i.e. $\left\| \tilde{E}^k - E^k \right\|_2 \leq r$, then we have $\left\| \tilde{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k \right\|_2 \leq I_k \cdot r$ where $\tilde{\boldsymbol{\alpha}}^k, \boldsymbol{\alpha}^k$ are the incomplete task gradients corresponding to

respectively old and new model parameter values and I_k depends on various quantities such as the smoothness parameter and the number of layers in the network.

Lemma 1 (part 2) tells us that for the final layer, we have $\boldsymbol{\alpha}^K = \mathbf{G}(W^{K+1})^\top$ as well as for any $k < K$, we have $\boldsymbol{\alpha}^k = \frac{\partial(\boldsymbol{\alpha}^{k+1} \odot X^{k+1})}{\partial X^k}\Big|_{\boldsymbol{\alpha}^{k+1}}$. We analyze this using an inductive argument.

1. Case 1: $k = K$ (Base Case): In this case we have
$$\begin{aligned}\tilde{\boldsymbol{\alpha}}^K - \boldsymbol{\alpha}^K &= \tilde{\mathbf{G}}(\tilde{W}^{K+1})^\top - \mathbf{G}(W^{K+1})^\top \\ &= (\tilde{\mathbf{G}} - \mathbf{G})(\tilde{W}^{K+1})^\top + \mathbf{G}(\tilde{W}^{K+1} - W^{K+1})^\top\end{aligned}$$
Now, the travel condition tells us that $\left\|\tilde{W}^{K+1} - W^{K+1}\right\|_2 \leq r$, boundedness tells us that $\|\mathbf{G}\|_2, \left\|\tilde{\mathbf{W}}^{K+1}\right\|_2 \leq B$. Also, since the loss function is H-smooth, the task gradients are H-Lipschitz which implies along with the travel condition that $\left\|\tilde{\mathbf{G}} - \mathbf{G}\right\|_2 \leq H \cdot r$. Put together this tells us that
$$\left\|\tilde{\boldsymbol{\alpha}}^K - \boldsymbol{\alpha}^K\right\|_2 \leq (H+1)B \cdot r,$$
telling us that $I_K \leq (H+1)B$.
2. Case 2: $k < K$ (Inductive Case): In this case, let $\tilde{X}^k$ and X^k denote the embeddings with respect to the old and new parameters respectively. Then we have
$$\begin{aligned}\tilde{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k &= \frac{\partial(\tilde{\boldsymbol{\alpha}}^{k+1} \odot \tilde{X}^{k+1})}{\partial \tilde{X}^k}\Bigg|_{\tilde{\boldsymbol{\alpha}}^{k+1}} - \frac{\partial(\boldsymbol{\alpha}^{k+1} \odot X^{k+1})}{\partial X^k}\Bigg|_{\boldsymbol{\alpha}^{k+1}} \\ &= \underbrace{\frac{\partial(\tilde{\boldsymbol{\alpha}}^{k+1} \odot \tilde{X}^{k+1})}{\partial \tilde{X}^k}\Bigg|_{\tilde{\boldsymbol{\alpha}}^{k+1}} - \frac{\partial(\boldsymbol{\alpha}^{k+1} \odot \tilde{X}^{k+1})}{\partial \tilde{X}^k}\Bigg|_{\boldsymbol{\alpha}^{k+1}}}_{(P)} \\ &\quad + \underbrace{\frac{\partial(\boldsymbol{\alpha}^{k+1} \odot \tilde{X}^{k+1})}{\partial \tilde{X}^k}\Bigg|_{\boldsymbol{\alpha}^{k+1}} - \frac{\partial(\boldsymbol{\alpha}^{k+1} \odot X^{k+1})}{\partial X^k}\Bigg|_{\boldsymbol{\alpha}^{k+1}}}_{(Q)}\end{aligned}$$
By induction we have $\left\|\tilde{\boldsymbol{\alpha}}^{k+1} - \boldsymbol{\alpha}^{k+1}\right\|_2 \leq I_{k+1} \cdot r$ and bounded gradients tells us $\left\|\frac{\partial \tilde{X}^{k+1}}{\partial \tilde{X}^k}\right\|_2 \leq B$ giving us $\|(P)\|_2 \leq I_{k+1}B \cdot r$. To analyze the term $\|(Q)\|_2$, recall that we have $X^{k+1} = f(X^k; E^{k+1})$ and since the overall task loss is H smooth, so must be the function f. Abusing notation to let H denote the Lipschitz constant of the network gives us $\left\|\tilde{X}^k - X^k\right\|_2 \leq H \cdot r$. Then we have
$$\begin{aligned}\frac{\partial(\tilde{X}^{k+1})}{\partial \tilde{X}^k} - \frac{\partial X^{k+1}}{\partial X^k} &= f'(\tilde{X}^k; \tilde{E}^{k+1}) - f'(X^k; E^{k+1}) \\ &= \underbrace{f'(\tilde{X}^k; \tilde{E}^{k+1}) - f'(X^k; \tilde{E}^{k+1})}_{(M)} + \underbrace{f'(X^k; \tilde{E}^{k+1}) - f'(X^k; E^{k+1})}_{(N)}\end{aligned}$$
Applying smoothness, we have $\|(M)\|_2 \leq H^2 \cdot r$ as well as $\|(N)\|_2 \leq H \cdot r$. Together with bounded gradients that gives us $\left\|\boldsymbol{\alpha}^{k+1}\right\|_2 \leq B$, we have $\|(Q)\|_2 \leq BH(H+1) \cdot r$. Together, we have
$$\left\|\tilde{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k\right\|_2 \leq B(H(H+1) + I_{k+1}) \cdot r,$$
telling us that $I_k \leq B(H(H+1) + I_{k+1})$.

The above tells us that in Algorithm 2, suppose the parameter update steps i.e. step 7 and step 10 are executed by effecting τ mini-batch SGD steps or else τ full-batch GD steps each time, with

step length η, then the amount of travel in any model parameter is bounded above by $\tau\eta B$ i.e. $\left\|\tilde{E}^k - E^k\right\|_2 \leq \tau\eta B$ and so on. Now, incomplete task gradients $\boldsymbol{\alpha}^k$ are updated only after model parameters for all layers have been updated once and the algorithm loops back to step 6. Thus, Lipschitzness of the gradients tells us that the staleness in the incomplete task gradients, for any $k \in [K]$, is upper bounded by

$$\left\|\tilde{\boldsymbol{\alpha}}^k - \boldsymbol{\alpha}^k\right\|_2 \leq \tau\eta \cdot IB,$$

where we take $I = \max_{k \leq K} I_k$ for sake of simplicity. Now, since $\left\|\frac{\partial X_{ip}^k}{\partial E^k}\right\|_2 \leq B$ as gradients are bounded, we have $\sigma_{\max}(Z^k) \leq N \cdot d_k B$. Then, combining with the result in Equation (2) gives us

$$\left\|\frac{\widetilde{\partial \mathcal{L}}}{\partial E^k} - \frac{\partial \mathcal{L}}{\partial E^k}\right\|_2 \leq \tau\eta \cdot IB^2 d_k$$

Taking in contributions of gradients of all layers and the final classifier layer gives us the bias in the total gradient using triangle inequality as

$$\left\|\widetilde{\nabla \mathcal{L}} - \nabla \mathcal{L}\right\|_2 \leq \sum_{k \in [K]} \left\|\frac{\widetilde{\partial \mathcal{L}}}{\partial E^k} - \frac{\partial \mathcal{L}}{\partial E^k}\right\|_2 + \left\|\frac{\widetilde{\partial \mathcal{L}}}{\partial W^{K+1}} - \frac{\partial \mathcal{L}}{\partial W^{K+1}}\right\|_2 \leq \tau\eta \cdot IB^2 d_{\max}(K+1),$$

where $d_{\max} = \max_{k \in K} \ d_k$ is the maximum embedding dimensionality of any layer.

C.6 Step 4 (i): Convergence Guarantee (mini-batch SGD)

Let us analyze convergence in the case when updates are made using mini-batch SGD in steps 7, 10 of Algorithm 2. The discussion above establishes an upper bound on the *absolute* bias in the gradients. However, our proofs later require a *relative* bound which we tackle now. Let us decide to set the step length to $\eta = \frac{1}{C\sqrt{T}}$ for some constant $C > 0$ that will be decided later and also set some value $\phi < 1$. Then two cases arise

1. **Case 1**: The relative gradient bias is too large i.e. $\tau\eta \cdot IB^2 d_{\max}(K+1) > \phi \cdot \|\nabla \mathcal{L}\|_2$. In this case we are actually done since we get

$$\|\nabla \mathcal{L}\|_2 \leq \frac{\tau \cdot IB^2 d_{\max}(K+1)}{\phi \cdot C\sqrt{T}},$$

   i.e. we are already at an approximate first-order stationary point.
2. **Case 2**: The relative gradient bias is small i.e. $\tau\eta \cdot IB^2 d_{\max}(K+1) \leq \phi \cdot \|\nabla \mathcal{L}\|_2$. In this case we satisfy the relative bias bound required by Lemma 4 (part 2) with $\delta = \phi$.

This shows that either Case 1 happens in which case we are done or else Case 2 keeps applying which means that Lemma 4 (part 2) keeps getting its prerequisites satisfied. If Case 1 does not happen for T steps, then Lemma 4 (part 2) assures us that we will arrive at a point where

$$\mathbb{E}\left[\|\nabla \mathcal{L}\|_2^2\right] \leq \frac{2C(\mathcal{L}^0 - \mathcal{L}^*) + H\sigma^2/C}{(1-\phi)\sqrt{T}}$$

where $\mathcal{L}^0, \mathcal{L}^*$ are respectively the initial and optimal values of the loss function which we recall is H-smooth and σ^2 is the variance due to mini-batch creation and we set $\eta = \frac{1}{C\sqrt{T}}$. Setting $C = \sqrt{\frac{H\sigma^2}{2(\mathcal{L}^0 - \mathcal{L}^*)}}$ tells us that within T steps, either we will achieve

$$\|\nabla \mathcal{L}\|_2^2 \leq \frac{2\tau^2 \cdot I^2 B^4 d_{\max}^2 (K+1)^2 (\mathcal{L}^0 - \mathcal{L}^*)}{\phi^2 H \sigma^2 T},$$

or else we will achieve

$$\mathbb{E}\left[\|\nabla \mathcal{L}\|_2^2\right] \leq \frac{2\sigma\sqrt{2(\mathcal{L}^0 - \mathcal{L}^*)H}}{(1-\phi)\sqrt{T}}$$

Setting $\tau = T^{\frac{1}{4}} \cdot \left(\frac{\sqrt{\sigma^3 H \sqrt{H}}}{(\mathcal{L}^0 - \mathcal{L}^*)^{\frac{1}{4}} I B^2 d_{\max}(K+1)} \right)$ balances the two quantities in terms of their dependence on T (module absolute constants such as ϕ). Note that as expected, as the quantities $I, B, d_{\max}, K$ increase, the above limit on τ goes down i.e. we are able to perform fewer and fewer updates to the model parameters before a refresh is required. This concludes the proof of Theorem 3 for the second case.

C.7 Step 4 (ii): Convergence Guarantee (full-batch GD)

In this case we similarly have either the relative gradient bias to be too large in which case we get

$$\|\nabla \mathcal{L}\|_2 \leq \frac{\tau \eta \cdot I B^2 d_{\max}(K+1)}{\phi},$$

or else we satisfy the relative bias bound required by Lemma 4 (part 1) with $\delta = \phi$. This shows that either Case 1 happens in which case we are done or else Case 2 keeps applying which means that Lemma 4 (part 1) keeps getting its prerequisites satisfied. If Case 1 does not happen for T steps, then Lemma 4 (part 1) assures us that we will arrive at a point where

$$\|\nabla \mathcal{L}\|_2^2 \leq \frac{2(\mathcal{L}^0 - \mathcal{L}^*)}{\eta(1-\phi)T}$$

In this case, for a given value of τ, setting $\eta = \left(\frac{2(\mathcal{L}^0 - \mathcal{L}^*)}{(1-\phi)\tau^2 I^2 B^4 d_{\max}^2 (K+1)^2 T} \right)^{\frac{1}{3}}$ gives us that within T iterations, we must achieve

$$\|\nabla \mathcal{L}\|_2^2 \leq \left(\frac{2\tau(\mathcal{L}^0 - \mathcal{L}^*) I B^2 d_{\max}(K+1)}{(1-\phi)} \right)^{\frac{2}{3}} \cdot \frac{1}{T^{\frac{2}{3}}}$$

In this case, it is prudent to set $\tau = \mathcal{O}(1)$ so as to not deteriorate the convergence rate. This concludes the proof of Theorem 3 for the first case.

C.8 Generic Convergence Results

Lemma 4 (First-order Stationarity with a Smooth Objective)**.** *Let $f : \boldsymbol{\Theta} \to \mathbb{R}$ be an H-smooth objective over model parameters $\theta \in \boldsymbol{\Theta}$ that is being optimized using a gradient oracle and the following update for step length η:*

$$\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \cdot \mathbf{g}^t$$

Let $\boldsymbol{\theta}^*$ be an optimal point i.e. $\boldsymbol{\theta}^* \in \arg\min_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} f(\boldsymbol{\theta})$. Then, the following results hold depending on the nature of the gradient oracle:

1. *If a non-stochastic gradient oracle with bounded bias is used i.e. for some $\delta \in (0,1)$, for all t, we have $\mathbf{g}^t = \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t$ where $\|\boldsymbol{\Delta}^t\|_2 \leq \delta \cdot \|\nabla f(\boldsymbol{\theta}^t)\|_2$ and if the step length satisfies $\eta \leq \frac{(1-\delta)}{2H(1+\delta^2)}$, then for any $T > 0$, for some $t \leq T$ we must have*

$$\left\|\nabla f(\boldsymbol{\theta}^t)\right\|_2^2 \leq \frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))}{\eta(1-\delta)T}$$

2. *If a stochastic gradient oracle is used with bounded bias i.e. for some $\delta \in (0,1)$, for all t, we have $\mathbb{E}[\,|\,\mathbf{g}^t \boldsymbol{\theta}^t] = \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t$ where $\|\boldsymbol{\Delta}^t\|_2 \leq \delta \cdot \|\nabla f(\boldsymbol{\theta}^t)\|_2$, as well as bounded variance i.e. for all t, we have $\mathbb{E}\left[\,|\,\|\mathbf{g}^t - \nabla f(\boldsymbol{\theta}^t) - \boldsymbol{\Delta}^t\|_2^2 \boldsymbol{\theta}^t\right] \leq \sigma^2$ and if the step length satisfies $\eta \leq \frac{(1-\delta)}{2H(1+\delta^2)}$, as well as $\eta = \frac{1}{C\sqrt{T}}$ for some $C > 0$, then for any $T > \frac{4H^2(1+\delta^2)^2}{C^2(1-\delta)^2}$, for some $t \leq T$ we must have*

$$\mathbb{E}\left[\left\|\nabla f(\boldsymbol{\theta}^t)\right\|_2^2\right] \leq \frac{2C(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*)) + H\sigma^2/C}{(1-\delta)\sqrt{T}}$$

Proof (of Lemma 4). To prove the first part, we notice that smoothness of the objective gives us

$$f(\boldsymbol{\theta}^{t+1}) \leq f(\boldsymbol{\theta}^t) + \left\langle \nabla f(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t \right\rangle + \frac{H}{2} \left\| \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t \right\|_2^2$$

Since we used the update $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \cdot \mathbf{g}^t$ and we have $\mathbf{g}^t = \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t$, the above gives us

$$\begin{aligned} f(\boldsymbol{\theta}^{t+1}) &\leq f(\boldsymbol{\theta}^t) - \eta \cdot \left\langle \nabla f(\boldsymbol{\theta}^t), \mathbf{g}^t \right\rangle + \frac{H\eta^2}{2} \left\| \mathbf{g}^t \right\|_2^2 \\ &= f(\boldsymbol{\theta}^t) - \eta \cdot \left\langle \nabla f(\boldsymbol{\theta}^t), \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t \right\rangle + \frac{H\eta^2}{2} \left(\left\| \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t \right\|_2^2 \right) \\ &= f(\boldsymbol{\theta}^t) - \eta \cdot \left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2 - \eta \cdot \left\langle \nabla f(\boldsymbol{\theta}^t), \boldsymbol{\Delta}^t \right\rangle + \frac{H\eta^2}{2} \left(\left\| \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t \right\|_2^2 \right) \end{aligned}$$

Now, the Cauchy-Schwartz inequality along with the bound on the bias gives us $-\eta \cdot \langle \nabla f(\boldsymbol{\theta}^t), \boldsymbol{\Delta}^t \rangle \leq \eta\delta \cdot \|\nabla f(\boldsymbol{\theta}^t)\|_2^2$ as well as $\|\nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t\|_2^2 \leq 2(1+\delta^2) \cdot \|\nabla f(\boldsymbol{\theta}^t)\|_2^2$. Using these gives us

$$\begin{aligned} f(\boldsymbol{\theta}^{t+1}) &\leq f(\boldsymbol{\theta}^t) - \eta(1 - \delta - \eta H(1+\delta^2)) \cdot \left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2 \\ &\leq f(\boldsymbol{\theta}^t) - \frac{\eta(1-\delta)}{2} \cdot \left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2, \end{aligned}$$

since we chose $\eta \leq \frac{(1-\delta)}{2H(1+\delta^2)}$. Reorganizing, taking a telescopic sum over all t, using $f(\boldsymbol{\theta}^{T+1}) \geq f(\boldsymbol{\theta}^*)$ and making an averaging argument tells us that since we set , for any $T > 0$, it must be the case that for some $t \leq T$, we have

$$\left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2 \leq \frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))}{\eta(1-\delta)T}$$

This proves the first part. For the second part, we yet again invoke smoothness to get

$$f(\boldsymbol{\theta}^{t+1}) \leq f(\boldsymbol{\theta}^t) + \left\langle \nabla f(\boldsymbol{\theta}^t), \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t \right\rangle + \frac{H}{2} \left\| \boldsymbol{\theta}^{t+1} - \boldsymbol{\theta}^t \right\|_2^2$$

Since we used the update $\boldsymbol{\theta}^{t+1} = \boldsymbol{\theta}^t - \eta \cdot \mathbf{g}^t$, the above gives us

$$\begin{aligned} f(\boldsymbol{\theta}^{t+1}) &\leq f(\boldsymbol{\theta}^t) - \eta \cdot \left\langle \nabla f(\boldsymbol{\theta}^t), \mathbf{g}^t \right\rangle + \frac{H\eta^2}{2} \left\| \mathbf{g}^t \right\|_2^2 \\ &= f(\boldsymbol{\theta}^t) - \eta \cdot \left\langle \nabla f(\boldsymbol{\theta}^t), \mathbf{g}^t \right\rangle + \frac{H\eta^2}{2} \left(\left\| \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t \right\|_2^2 + \left\| \mathbf{g}^t - \nabla f(\boldsymbol{\theta}^t) - \boldsymbol{\Delta}^t \right\|_2^2 \right) \\ &\quad - H\eta^2 \left\langle \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t, \mathbf{g}^t - \nabla f(\boldsymbol{\theta}^t) - \boldsymbol{\Delta}^t \right\rangle \end{aligned}$$

Taking conditional expectations on both sides gives us

$$\mathbb{E}\left[\mid f(\boldsymbol{\theta}^{t+1}) \boldsymbol{\theta}^t \right] \leq f(\boldsymbol{\theta}^t) - \eta \cdot \left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2 - \eta \cdot \left\langle \nabla f(\boldsymbol{\theta}^t), \boldsymbol{\Delta}^t \right\rangle + \frac{H\eta^2}{2} \left(\left\| \nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t \right\|_2^2 + \sigma^2 \right)$$

Now, the Cauchy-Schwartz inequality along with the bound on the bias gives us $-\eta \cdot \langle \nabla f(\boldsymbol{\theta}^t), \boldsymbol{\Delta}^t \rangle \leq \eta\delta \cdot \|\nabla f(\boldsymbol{\theta}^t)\|_2^2$ as well as $\|\nabla f(\boldsymbol{\theta}^t) + \boldsymbol{\Delta}^t\|_2^2 \leq 2(1+\delta^2) \cdot \|\nabla f(\boldsymbol{\theta}^t)\|_2^2$. Using these and applying a total expectation gives us

$$\begin{aligned} \mathbb{E}\left[f(\boldsymbol{\theta}^{t+1}) \right] &\leq \mathbb{E}\left[f(\boldsymbol{\theta}^t) \right] - \eta(1 - \delta - \eta H(1+\delta^2)) \cdot \mathbb{E}\left[\left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2 \right] + \frac{H\eta^2}{2} \cdot \sigma^2 \\ &\leq \mathbb{E}\left[f(\boldsymbol{\theta}^t) \right] - \frac{\eta(1-\delta)}{2} \cdot \mathbb{E}\left[\left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2 \right] + \frac{H\eta^2}{2} \cdot \sigma^2 \end{aligned}$$

where the second step follows since we set $\eta \leq \frac{(1-\delta)}{2H(1+\delta^2)}$. Reorganizing, taking a telescopic sum over all t, using $f(\boldsymbol{\theta}^{T+1}) \geq f(\boldsymbol{\theta}^*)$ and making an averaging argument tells us that for any $T > 0$, it must be the case that for some $t \leq T$, we have

$$\mathbb{E}\left[\left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2 \right] \leq \frac{2(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*))}{\eta(1-\delta)T} + \frac{H\eta\sigma^2}{1-\delta}$$

However, since we also took care to set $\eta = \frac{1}{C\sqrt{T}}$, we get

$$\mathbb{E}\left[\left\| \nabla f(\boldsymbol{\theta}^t) \right\|_2^2 \right] \leq \frac{2C(f(\boldsymbol{\theta}^0) - f(\boldsymbol{\theta}^*)) + H\sigma^2/C}{(1-\delta)\sqrt{T}}$$

which proves the second part and finishes the proof. □