STOCHASTIC TRAINING OF GRAPH CONVOLUTIONAL NETWORKS

Anonymous authors
Paper under double-blind review

ABSTRACT

Graph convolutional networks (GCNs) are powerful deep neural networks for graph-structured data. However, GCN computes nodes’ representation recursively from their neighbors, making the receptive field size grow exponentially with the number of layers. Previous attempts on reducing the receptive field size by subsampling neighbors do not have any convergence guarantee, and their receptive field size per node is still in the order of hundreds. In this paper, we develop a preprocessing strategy and two control variate based algorithms to further reduce the receptive field size. Our algorithms are guaranteed to converge to GCN’s local optimum regardless of the neighbor sampling size. Empirical results show that our algorithms have a similar convergence speed per epoch with the exact algorithm even using only two neighbors per node. The time consumption of our algorithm on the Reddit dataset is only one fifth of previous neighbor sampling algorithms.

1 INTRODUCTION

Graph convolution networks (GCNs) (Kipf & Welling, 2017) generalize convolutional neural networks (CNNs) (LeCun et al., 1995) to graph structured data. The “graph convolution” operation applies same linear transformation to all the neighbors of a node, followed by mean pooling. By stacking multiple graph convolution layers, GCNs can learn nodes’ representation by utilizing information from distant neighbors. GCNs have been applied to semi-supervised node classification (Kipf & Welling, 2017), inductive node embedding (Hamilton et al., 2017a), link prediction (Kipf & Welling, 2016; Berg et al., 2017) and knowledge graphs (Schlichtkrull et al., 2017), outperforming multi-layer perceptron (MLP) models that do not use the graph structure and graph embedding approaches (Perozzi et al., 2014; Tang et al., 2015; Grover & Leskovec, 2016) that do not use node features.

However, the graph convolution operation makes it difficult to train GCN efficiently. A node’s representation at layer $L$ is computed recursively by all its neighbors of a node, followed by mean pooling. Therefore, the receptive field of a single node grows exponentially with respect to the number of layers, as illustrated in Fig. 1(a). Due to the large receptive field size, Kipf & Welling (2017) proposed training GCN by a batch algorithm, which computes the representation for all the neighbors of a node, followed by mean pooling. By stacking multiple graph convolution layers, GCNs can learn nodes’ representation by utilizing information from distant neighbors. GCNs have been applied to semi-supervised node classification (Kipf & Welling, 2017), inductive node embedding (Hamilton et al., 2017a), link prediction (Kipf & Welling, 2016; Berg et al., 2017) and knowledge graphs (Schlichtkrull et al., 2017), outperforming multi-layer perceptron (MLP) models that do not use the graph structure and graph embedding approaches (Perozzi et al., 2014; Tang et al., 2015; Grover & Leskovec, 2016) that do not use node features.

In this paper, we develop novel stochastic training algorithms for GCNs such that $D^{(l)}$ can be as low as two, so that the time complexity of training GCN is comparable with training MLPs. Our methods are built on two techniques. First, we propose a strategy which preprocesses the first graph convolution layer, so that we only need to consider all neighbors within $L-1$ hops instead of $L$ hops. This is significant because most GCNs only have $L = 2$ layers (Kipf & Welling, 2017; Hamilton et al., 2017a) made an initial attempt on developing stochastic algorithms to train GCNs, which is referred as neighbor sampling (NS) in this paper. Instead of considering all the neighbors, they randomly subsample $D^{(l)}$ neighbors at the $l$-th layer. Therefore, they reduce the receptive field size to $L-1$, as shown in Fig. 1(b). They found that for two layer GCNs, keeping $D^{(1)} = 10$ and $D^{(2)} = 25$ neighbors can achieve comparable performance with the original model. However, there is no theoretical guarantee on the predictive performance of the model learnt by NS comparing with the original algorithm. Moreover, the time complexity of NS is still $D^{(1)}D^{(2)} = 250$ times larger than training an MLP, which is unsatisfactory.

In this paper, we develop novel stochastic training algorithms for GCNs such that $D^{(l)}$ can be as low as two, so that the time complexity of training GCN is comparable with training MLPs. Our methods are built on two techniques. First, we propose a strategy which preprocesses the first graph convolution layer, so that we only need to consider all neighbors within $L-1$ hops instead of $L$ hops. This is significant because most GCNs only have $L = 2$ layers (Kipf & Welling, 2017; Hamilton et al., 2017a) made an initial attempt on developing stochastic algorithms to train GCNs, which is referred as neighbor sampling (NS) in this paper. Instead of considering all the neighbors, they randomly subsample $D^{(l)}$ neighbors at the $l$-th layer. Therefore, they reduce the receptive field size to $L-1$, as shown in Fig. 1(b). They found that for two layer GCNs, keeping $D^{(1)} = 10$ and $D^{(2)} = 25$ neighbors can achieve comparable performance with the original model. However, there is no theoretical guarantee on the predictive performance of the model learnt by NS comparing with the original algorithm. Moreover, the time complexity of NS is still $D^{(1)}D^{(2)} = 250$ times larger than training an MLP, which is unsatisfactory.

In this paper, we develop novel stochastic training algorithms for GCNs such that $D^{(l)}$ can be as low as two, so that the time complexity of training GCN is comparable with training MLPs. Our methods are built on two techniques. First, we propose a strategy which preprocesses the first graph convolution layer, so that we only need to consider all neighbors within $L-1$ hops instead of $L$ hops. This is significant because most GCNs only have $L = 2$ layers (Kipf & Welling, 2017; Hamilton et al., 2017a) made an initial attempt on developing stochastic algorithms to train GCNs, which is referred as neighbor sampling (NS) in this paper. Instead of considering all the neighbors, they randomly subsample $D^{(l)}$ neighbors at the $l$-th layer. Therefore, they reduce the receptive field size to $L-1$, as shown in Fig. 1(b). They found that for two layer GCNs, keeping $D^{(1)} = 10$ and $D^{(2)} = 25$ neighbors can achieve comparable performance with the original model. However, there is no theoretical guarantee on the predictive performance of the model learnt by NS comparing with the original algorithm. Moreover, the time complexity of NS is still $D^{(1)}D^{(2)} = 250$ times larger than training an MLP, which is unsatisfactory.

In this paper, we develop novel stochastic training algorithms for GCNs such that $D^{(l)}$ can be as low as two, so that the time complexity of training GCN is comparable with training MLPs. Our methods are built on two techniques. First, we propose a strategy which preprocesses the first graph convolution layer, so that we only need to consider all neighbors within $L-1$ hops instead of $L$ hops. This is significant because most GCNs only have $L = 2$ layers (Kipf & Welling, 2017; Hamilton et al., 2017a) made an initial attempt on developing stochastic algorithms to train GCNs, which is referred as neighbor sampling (NS) in this paper. Instead of considering all the neighbors, they randomly subsample $D^{(l)}$ neighbors at the $l$-th layer. Therefore, they reduce the receptive field size to $L-1$, as shown in Fig. 1(b). They found that for two layer GCNs, keeping $D^{(1)} = 10$ and $D^{(2)} = 25$ neighbors can achieve comparable performance with the original model. However, there is no theoretical guarantee on the predictive performance of the model learnt by NS comparing with the original algorithm. Moreover, the time complexity of NS is still $D^{(1)}D^{(2)} = 250$ times larger than training an MLP, which is unsatisfactory.
Neighbour sampling (NS) algorithm (Hamilton et al., 2017a). Second, we develop two control variate (CV) based stochastic training algorithms. We show that our CV-based algorithms have lower variance than NS, and for GCNs without dropout, our algorithm provably converges to a local optimum of the model regardless of $D^{(l)}$.

We empirically test on six graph datasets, and show that our techniques significantly reduce the bias and variance of the gradient from NS with the same receptive field size. Our algorithm with $D^{(l)} = 2$ achieves the same predictive performance with the exact algorithm in comparable number of epochs on all the datasets, while the training time is 5 times shorter on our largest dataset.

2 BACKGROUNDS

We now briefly review graph convolutional networks (GCNs) (Kipf & Welling, 2017) and the neighbor sampling (NS) algorithm (Hamilton et al., 2017a).

2.1 GRAPH CONVOLUTIONAL NETWORKS

The original GCN was presented in a semi-supervised node classification task (Kipf & Welling, 2017). We follow this setting throughout this paper. Generalization of GCN to other tasks can be found in Kipf & Welling (2016); Berg et al. (2017); Schlichtkrull et al. (2017) and Hamilton et al. (2017b). In the node classification task, we have an undirected graph found in Kipf & Welling (2016); Berg et al. (2017); Schlichtkrull et al. (2017) and Hamilton et al. (2017a). Second, we develop two control variate (CV) based stochastic training algorithms. We show that our CV-based algorithms have lower variance than NS, and for GCNs without dropout, our algorithm provably converges to a local optimum of the model regardless of $D^{(l)}$.

Finally, the loss is defined as

$$L = \frac{1}{|V_L|} \sum_{v \in V_L} f(y_v, Z_v^{(L)})$$

where $f(\cdot, \cdot)$ can be the square loss, cross entropy loss, etc., depending on the type of the label.

When $P = I$, GCN reduces to a multi-layer perceptron (MLP) model which does not use the graph structure. Comparing with MLP, GCN is able to utilize neighbor information for node classification. We define $n(v, L)$ as the set of all the L-neighbors of node $v$, i.e., the nodes that are reachable from $v$ within $L$ hops. It is easy to see from Fig. 1(a) that in an $L$-layer GCN, a node uses the information from all its $L$-neighbors. This makes GCN more powerful than MLP, but also complicates the stochastic training, which utilizes an approximated gradient

$$\nabla L \approx \frac{1}{|V_B|} \sum_{v \in V_B} \nabla f(y_v, Z_v^{(L)})$$

where $V_B \subset V_L$ is a minibatch of training data. The large receptive field size $|\cup_{v \in V_B} n(v, L)|$ per minibatch leads to high time complexity, space complexity and amount of IO. See Table 1 for the average number of 1- and 2-neighbors of our datasets.

2.2 ALTERNATIVE NOTATION

We introduce alternative notations to help compare different algorithms. Let $U^{(l)} = P \tilde{H}^{(l)}$, or $U^{(l)}_{v,v'} = \sum_{v'' \in n(v, 1)} P_{v,v''} \tilde{h}^{(l)}_{v''}$, we focus on studying how $U^{(l)}_{v,v'}$ is computed based on node $v$’s neighbors. To keep notations simple, we omit all the subscripts and tildes, and exchange the ID of nodes such as

$$L = \frac{1}{|V_L|} \sum_{v \in V_L} f(y_v, Z_v^{(L)})$$

where $f(\cdot, \cdot)$ can be the square loss, cross entropy loss, etc., depending on the type of the label.

When $P = I$, GCN reduces to a multi-layer perceptron (MLP) model which does not use the graph structure. Comparing with MLP, GCN is able to utilize neighbor information for node classification. We define $n(v, L)$ as the set of all the L-neighbors of node $v$, i.e., the nodes that are reachable from $v$ within $L$ hops. It is easy to see from Fig. 1(a) that in an $L$-layer GCN, a node uses the information from all its $L$-neighbors. This makes GCN more powerful than MLP, but also complicates the stochastic training, which utilizes an approximated gradient

$$\nabla L \approx \frac{1}{|V_B|} \sum_{v \in V_B} \nabla f(y_v, Z_v^{(L)})$$

where $V_B \subset V_L$ is a minibatch of training data. The large receptive field size $|\cup_{v \in V_B} n(v, L)|$ per minibatch leads to high time complexity, space complexity and amount of IO. See Table 1 for the average number of 1- and 2-neighbors of our datasets.

2.2 ALTERNATIVE NOTATION

We introduce alternative notations to help compare different algorithms. Let $U^{(l)} = P \tilde{H}^{(l)}$, or $U^{(l)}_{v,v'} = \sum_{v'' \in n(v, 1)} P_{v,v''} \tilde{h}^{(l)}_{v''}$, we focus on studying how $U^{(l)}_{v,v'}$ is computed based on node $v$’s neighbors. To keep notations simple, we omit all the subscripts and tildes, and exchange the ID of nodes such as

$$L = \frac{1}{|V_L|} \sum_{v \in V_L} f(y_v, Z_v^{(L)})$$

where $f(\cdot, \cdot)$ can be the square loss, cross entropy loss, etc., depending on the type of the label.
We first present a technique to preprocess the first graph convolution layer, by approximating used for testing and for computing stochastic gradient the number of L significant reduction of the receptive field size from the number of only have two graph convolution layers (Kipf & Welling, 2017; Hamilton et al., 2017a), this gives a first layer is merely a fully connected layer instead of a graph convolution one. Since most GCNs as the new input. In this way, the actual number of graph convolution layers is reduced by one — the This approximation does not change the expectation because E D D D

\[ E[D] \] where \( D = \left| \{u \mid \text{node } u \text{ has degree } 1 \} \right| \) is the number of neighbors. We get the propagation rule \( u = \sum_{v=1}^{D} p_{uv} h_v \), which is used interchangeably with the matrix form \( U(t) = PH(t) \).

2.3 Neighbor Sampling

To reduce the receptive field size, Hamilton et al. (2017a) propose a neighbor sampling (NS) algorithm. On the \( l \)-th layer, they randomly choose \( D(l) \) neighbors for each node, and develop an estimator \( u_{NS} \) of \( u \) based on Monte-Carlo approximation \( u \approx u_{NS} = \frac{D(l)}{D(l) \sigma} \sum_{v \in D(l)} p_{uv} h_v \), where \( D(l) \subset [D(l)] \) is a subset of \( D(l) \) neighbors. In this way, they reduce the receptive field size from \( |\cup_{v \in V_B} u(v, L)| \) to \( O(|V_B| \prod_{l=1}^{L} D(l)) \). Neighbor sampling can also be written in a matrix form as

\[ \hat{H}_{NS}(l) = \text{Dropout}_p(P_{NS}(l)), \quad Z_{NS}^{(l+1)} = \hat{H}_{NS}(l) W^{(l)}, \quad \hat{H}_{NS}^{(l+1)} = \sigma(Z_{NS}^{(l+1)}) \]  

where \( \hat{P}(l) \) is a sparser unbiased estimator of \( P \), i.e., \( E[\hat{P}(l)] = P \). The approximate prediction \( Z_{NS}^{(L)} \) used for testing and for computing stochastic gradient \( \frac{1}{|V_B|} \sum_{v \in V_B} \nabla f(y_v, Z^{(L)}_{CV,v}) \) during training.

The NS estimator \( u_{NS} \) is unbiased. However it has a large variance, which leads to biased prediction and gradients after the non-linearity in subsequent layers. Due to the biased gradients, training with NS does not converge to the local optimum of GCN. When \( D(l) \) is moderate, NS may has some regularization effect like dropout (Srivastava et al., 2014), where it drops neighbors instead of features. However, for the extreme case \( D(l) = 2 \), the neighbor dropout rate is too high to reach high predictive performance, as we will see in Sec. 5.4. Intuitively, making prediction solely depends on one neighbor is inferior to using all the neighbors. To keep comparable prediction performance with the original GCN, Hamilton et al. (2017a) use relatively large \( D(1) = 10 \) and \( D(2) = 25 \). Their receptive field size \( D(1) \times D(2) = 250 \) is still much larger than MLP, which is 1.

3 Preprocessing First Layer

We first present a technique to preprocess the first graph convolution layer, by approximating ADropout \( p \) (\( X \)) with \( \text{Dropout}_p(AX) \). The model becomes

\[ Z^{(l+1)} = \text{Dropout}_p(PH^{(l)}(l)) W^{(l)}, \quad H^{(l+1)} = \sigma(Z^{(l+1)}) \]  

This approximation does not change the expectation because \( E[\text{ADropout}_p(X)] = E[\text{Dropout}_p(AX)] \), and it does not affect the predictive performance, as we shall see in Sec. 5.1.

The advantage of this modification is that we can preprocess \( U(0) = PH(0) = PX \) and takes \( U(0) \) as the new input. In this way, the actual number of graph convolution layers is reduced by one — the first layer is merely a fully connected layer instead of a graph convolution one. Since most GCNs only have two graph convolution layers (Kipf & Welling, 2017; Hamilton et al., 2017a), this gives a significant reduction of the receptive field size from the number of \( L \)-neighbors \( |\cup_{v \in V_B} u(v, L)| \) to the number of \( L - 1 \)-neighbors \( |\cup_{v \in V_B} u(v, L - 1)| \). The numbers are reported in Table 1.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>( V )</th>
<th>( E )</th>
<th>Degree</th>
<th>Degree 2</th>
<th>Type</th>
</tr>
</thead>
<tbody>
<tr>
<td>Citeseer</td>
<td>3,327</td>
<td>12,431</td>
<td>4</td>
<td>15</td>
<td>Document network</td>
</tr>
<tr>
<td>Cora</td>
<td>2,708</td>
<td>13,264</td>
<td>5</td>
<td>37</td>
<td>Document network</td>
</tr>
<tr>
<td>PubMed</td>
<td>19,717</td>
<td>108,365</td>
<td>6</td>
<td>60</td>
<td>Document network</td>
</tr>
<tr>
<td>NELL</td>
<td>65,755</td>
<td>318,135</td>
<td>5</td>
<td>1,597</td>
<td>Knowledge graph</td>
</tr>
<tr>
<td>PPI</td>
<td>14,755</td>
<td>458,973</td>
<td>31</td>
<td>970</td>
<td>Protein-protein interaction</td>
</tr>
<tr>
<td>Reddit</td>
<td>232,965</td>
<td>23,446,803</td>
<td>101</td>
<td>10,858</td>
<td>Document network</td>
</tr>
</tbody>
</table>

Table 1: Number of vertexes, edges, average number of 1- and 2-neighbors per node for each dataset. Undirected edges are counted twice and self-loops are counted once. Reddit is already subsampled to have a max degree of 128 following Hamilton et al. (2017a).

4 Control Variate Based Stochastic Approximation

We now present two novel control variate based estimators that have smaller variance as well as stronger theoretical guarantees than NS.

\footnote{For an integer \( N \), we define \( \{0, \ldots, N\} \) and \( \{1, \ldots, N\} \).}
4.1 Control variate based estimator

We assume that the model does not have dropout for now and will address dropout in Sec. [4.2] The idea is that we can approximate \( u = \sum_{v=1}^{D} p_v h_v \) better if we know the latest historical activations \( h_v \) of the neighbors, where we expect \( h_v \) and \( h_v \) are similar if the model weights do not change too fast during the training. With the historical activations, we approximate

\[
    u = \sum_{v=1}^{D} p_v h_v = \sum_{v=1}^{D} p_v (h_v - \bar{h}_v) + \sum_{v=1}^{D} p_v \bar{h}_v \approx Dp_{v'} \Delta h_{v'} + \sum_{v=1}^{D} p_v \bar{h}_v := u_{CV},
\]

where \( v' \) is a random neighbor, and \( \Delta h_{v'} = h_{v'} - \bar{h}_{v'} \). For the ease of presentation, we assume that we only use the latest activation of one neighbor, while the implementation also include the node itself besides the random neighbor, so \( D^{(l)} = 2 \). Using historical activations is simple because we need not to be computed recursively using their neighbors’ activations, as shown in Fig. 1(c).

Unlike NS, we apply Monte-Carlo approximation on the sum of the historical activations, where \( u_{NS} \) are unbiased estimators of \( \sum_{v} p_v \Delta h_v \) instead of \( \sum_{v} p_v h_v \). Since we expect \( h_v \) and \( \bar{h}_v \) to be close, \( \Delta h_v \) will be small and \( u_{CV} \) should have a smaller variance than \( u_{NS} \). Particularly, if the model weight is kept fixed, \( h_v \) should be eventually equal with \( h_v \), so that \( u_{CV} = 0 + \sum_{v=1}^{D} p_v \bar{h}_v = \sum_{v=1}^{D} p_v h_v = u \), i.e., the estimator has zero variance. The term \( CV = u_{CV} - u_{NS} = -Dp_{v'} \Delta h_{v'} + \sum_{v=1}^{D} p_v \bar{h}_v \) is a control variate [Ripley 2009, Chapter 5], which has zero mean and large correlation with \( u_{NS} \), to reduce its variance. We refer this stochastic approximation algorithm as CV, and we will formally analyze the variance and prove the convergence of the training using CV for stochastic gradient in subsequent sections.

In matrix form, CV computes the approximate predictions as follows, where we explicitly write down the iteration number \( i \) and add the subscript \( CV \) to the approximate activations \( 2 \)

\[
    Z^{(l+1)}_{CV,i} \leftarrow \left(  \tilde{P}^{(l)}_{i}(H^{(l)}_{CV,i} - \bar{H}^{(l)}_{CV,i}) + P\bar{H}^{(l)}_{CV,i} \right) W^{(l)}_i, \tag{5}
\]

\[
    H^{(l+1)}_{CV,i} \leftarrow \sigma(Z^{(l+1)}_{CV,i}), \quad \bar{H}^{(l)}_{CV,i+1} \leftarrow m^{(l)}_{i} H^{(l)}_{CV,i} + (1 - m^{(l)}_{i}) \bar{H}^{(l)}_{CV,i}, \tag{6}
\]

where \( \bar{h}^{(l)}_{CV,i,v} \) stores the latest activation of node \( v \) on layer \( l \) computed before time \( i \). Formally, let \( m^{(l)}_{i} \in \mathbb{R}^{V \times V} \) be a diagonal matrix, and \( (m^{(l)}_{i})_{vv} = 1 \) if \( (\tilde{P}^{(l)}_{i})_{vv} > 0 \) for any \( v' \). After finishing one iteration we update history \( \bar{H} \) with the activations computed in that iteration as Eq. (6).

4.2 Control variate for dropout

With dropout, the activations \( H \) are no longer deterministic. They become random variables whose randomness come from different dropout configurations. Therefore, \( \Delta h_v = h_v - \bar{h}_v \) is not necessarily small even if \( h_v \) and \( h_v \) have the same distribution. We develop another stochastic approximation algorithm, control variate for dropout (CVD), that works well with dropout.

Our method is based on the weight scaling procedure [Srivastava et al. 2014] to approximately compute the mean \( \mu_v := \mathbb{E}[h_v] \). That is, along with the dropout model, we can run a copy of the model with no dropout to obtain the mean \( \mu_v \), as illustrated in Fig. 1(d). With the mean, we can obtain a better stochastic approximation by separating the mean and variance

\[
    u = \sum_{v=1}^{D} p_v ((h_v - \mu_v) + (\mu_v - \bar{\mu}_v) + \bar{\mu}_v) \approx \sqrt{D}p_{v'}(h_{v'} - \mu_{v'}) + Dp_{v'} \Delta \mu_{v'} + \sum_{v=1}^{D} p_v \bar{\mu}_v := u_{CVD},
\]

where \( \mu_v \) is the historical mean activation, obtained by storing \( \mu_v \) instead of \( h_v \), and \( \Delta \mu = \mu_v - \bar{\mu}_v \) an unbiased estimator of \( u \) because the term \( \sqrt{D}p_{v'}(h_{v'} - \mu_{v'}) \) has zero mean, and the Monte-Carlo approximation \( \sum_{v=1}^{D} p_v (\mu_v - \bar{\mu}_v) \approx Dp_{v'} \Delta \mu_{v'} \) does not change the mean. The approximation \( \sum_{v=1}^{D} p_v (h_v - \bar{\mu}_v) \approx \sqrt{D}p_{v'}(h_{v'} - \mu_{v'}) \) is made by assuming \( h_v \)’s to be independent Gaussians, which we will soon clarify. The pseudocodes for CV and CVD are in Appendix [E].

4.3 Variance analysis

NS, CV and CVD are all unbiased estimators of \( u = \sum_v p_v h_v \). We analyze their variance in a simple independent Gaussian case, where we assume that activations are Gaussian random variables

\[2\] We will omit the subscripts \( CV \) and \( i \) in subsequent sections when there is no confusion.
[Unbiased Gradient] The stochastic gradient we run forward propagation by epochs. In each epoch, we randomly partition the vertex set\( V \) into \( I \) mini-batches \( V_1, \ldots, V_I \) and in the \( i \)-th iteration, we run a forward pass to compute the prediction for nodes in \( V_i \). Note that in each epoch we scan all the nodes instead of just testing nodes, to ensure that the activation of each node is computed at least once per epoch. The following theorem reveals the connection of the exact predictions and gradients, and their approximate versions by CV.

**Theorem 1.** For a fixed \( W \) and any \( i > L \) we have: (1) (Exact Prediction) The activations computed by CV are exact, i.e., \( Z_{CV,i}^{(l)} = Z^{(l)} \) for each \( l \in [L] \) and \( H_{CV,i}^{(l)} = H^{(l)} \) for each \( l \in [L-1] \).

(2) (Unbiased Gradient) The stochastic gradient \( g_{CV,i}(W) := \frac{1}{|V_i|} \sum_{v \in V_i} \nabla W f_y (v, z_{CV,i,v}^{(L)}) \) is an unbiased estimator of GCN’s gradient, i.e., \( \mathbb{E}_{p_{V_i}, z_{V_i}} g_{CV,i}(W) = \nabla W H^{(L)} \big|_{\mathcal{V} \setminus V_1} \).
4.5 CONVERGENCE GUARANTEE

The following theorem guarantees that for a model without dropout, training using CV’s approximated gradients converges to a local optimum of GCN, regardless of the neighbor sampling size $D^{(i)}$. Therefore, we can choose arbitrarily small $D^{(i)}$ without worrying about the convergence.

**Theorem 2.** Assume that (1) all the activations are $\rho$-Lipschitz, (2) the gradient of the cost function $\nabla z f(y, z)$ is $\rho$-Lipschitz and bounded, (3) $\|g_{CV}(W)\|_\infty$ and $\|g(W)\|_\infty = \|\nabla L(W)\|_\infty$ are bounded by $G > 0$ for all $P, V_B$ and $W$. (4) The loss $L(W)$ is $\rho$-smooth, i.e., $|L(W_1) - L(W_2)| \leq \frac{\rho}{2} \|W_2 - W_1\|^2 \forall W_1, W_2$, where $(A, B) = tr(A^T B)$ is the inner product of matrix $A$ and matrix $B$. We randomly run SGD for $R \leq N$ iterations, where $P_R(R = i) = \sum_{i=1}^{2^\gamma_i - \rho \gamma_i}$. Then, for the updates $W_{i+1} = W_i - \gamma_i g_{CV}(W_i)$ and step sizes $\gamma_i = \min\{\frac{1}{2}, \frac{1}{\sqrt{R}}\}$, there exists constants $K_1$ and $K_2$ which are irrelevant with $N$, s.t. $\forall N \geq L$,

$$E_{R \sim P_R} E_{P, V_B} \|\nabla L(W_R)\|^2 \leq \frac{2 \rho (L(W_1) - L(W^*)) + K_2}{N} + \frac{2 (L(W_1) - L(W^*)) + K_1}{\sqrt{N}}.$$

The proof can be found in Appendix B. Particularly, $\lim_{N \to \infty} E_{R \sim P_R} E_{P, V_B} \|\nabla L(W_R)\|^2 = 0$. Therefore, our algorithm converges to a local optimum.

4.6 TIME COMPLEXITY AND IMPLEMENTATION DETAILS

Finally, we discuss the time complexity of different algorithms. We decompose the time complexity as sparse time complexity for sparse-dense matrix multiplication such as $PH^{(i)}$, and dense time complexity for dense-dense matrix multiplication such as $U^{(i)}W^{(i)}$. Assume that the node feature is $K$-dimensional and the first hidden layer is $A$-dimensional, the batch GCN has $O(VEK)$ sparse and $O(VKA)$ dense time complexity per epoch. NS has $O(V \prod_{i=1}^{L} D^{(i)} K)$ sparse and $O(V \prod_{i=2}^{L} D^{(i)} K A)$ dense time complexity per epoch. The dense time complexity of CV is the same as NS. The sparse time complexity depends on the cost of computing the sum $\sum_v p_v \mu_v$.

There are $V \prod_{i=2}^{L} D^{(i)}$ such sums to compute on the first graph convolution layer, and overall cost is not larger than $O(VD \prod_{i=2}^{L} D^{(i)} K)$, if we subsample the graph such that the max degree is $D$, following Hamilton et al. (2017a). The sparse time complexity is $D/D^{(1)}$ times higher than NS.

Our implementation is similar as Kipf & Welling (2017). We store the node features in the main memory, without assuming that they fit in GPU memory as Hamilton et al. (2017a), which makes our implementation about 2 times slower than theirs. We keep the histories in GPU memory for efficiency since they are only $LH < K$ dimensional.

5 EXPERIMENTS

We examine the variance and convergence of our algorithms empirically on six datasets, including Citeseer, Cora, PubMed and NELL from Kipf & Welling (2017) and Reddit, PPI from Hamilton et al. (2017a), as summarized in Table 1. To measure the predictive performance, we report Micro-F1 for the multi-label PPI dataset, and accuracy for all the other multi-class datasets. We use the same model architectures with previous papers but slightly different hyperparameters (see Appendix D for the details). We repeat the convergence experiments 10 times on Citeseer, Cora, PubMed and NELL, and 5 times on Reddit and PPI. The experiments are done on a Titan X (Maxwell) GPU.

5.1 IMPACT OF PREPROCESSING

We first examine the approximation in Sec. 3 that switches the order of dropout and aggregating the neighbors. Let $M_0$ be the original model (Eq. 1) and $M_1$ be our approximated model (Eq. 3), we compare three settings: (1) $M_0$, $D^{(i)} = \infty$ is the exact algorithm without any neighbor sampling. (2) $M_1+PP$, $D^{(i)} = \infty$ changes the model from $M_0$ to $M_1$. Preprocessing does not affect the training for $D^{(i)} = \infty$. (3) $M_1+PP$, $D^{(i)} = 20$ uses NS with a relatively large number of neighbors. In Table 3 we can see that all the three settings performs similarly, i.e., our approximation does not affect the predictive performance. Therefore, we use $M_1+PP$, $D^{(i)} = 20$ as the exact baseline in following convergence experiments because it is the fastest among these three settings.
We first validate Theorem 2, which states that CV+PP converges to a local optimum of Exact, for models without dropout, regardless of $D^{(i)}$. We disable dropout and plot the training loss with respect to number of epochs as Fig. 2. We can see that CV+PP can always reach the same training loss with Exact, which matches the conclusion of Theorem 2. Meanwhile, NS and NS+PP have a higher training loss because their gradients are biased.

### 5.3 Convergence with Dropout

Next, we compare the predictive accuracy obtained by the model trained by different algorithms, with dropout turned on. We use different algorithms for training and the same Exact algorithm for testing, and report the validation accuracy at each training epoch. The result is shown in Fig. 3. We find that CVD+PP is the only algorithm that is able to reach comparable validation accuracy with Exact on all datasets. Furthermore, its convergence speed with respect to the number of epochs is comparable with Exact despite its $D^{(i)}$ is 10 times smaller. Note that CVD+PP performs much better than Exact on the PubMed dataset; we suspect it finds a better local optimum.

![Figure 3: Comparison of validation accuracy with respect to number of epochs. NS converges to 0.94 on the Reddit dataset and 0.6 on the PPI dataset.](image-url)
<table>
<thead>
<tr>
<th>Alg.</th>
<th>Valid. acc.</th>
<th>Epochs</th>
<th>Time (s)</th>
<th>Sparse GFLOP</th>
<th>Dense TFLOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>96.0</td>
<td>4.2</td>
<td>252</td>
<td>507</td>
<td>7.17</td>
</tr>
<tr>
<td>NS</td>
<td>94.4</td>
<td>102.0</td>
<td>577</td>
<td>76.5</td>
<td>21.4</td>
</tr>
<tr>
<td>NS+PP</td>
<td>96.0</td>
<td>35.0</td>
<td>195</td>
<td>2.53</td>
<td>7.36</td>
</tr>
<tr>
<td>CV+PP</td>
<td>96.0</td>
<td>7.8</td>
<td>56</td>
<td>40.6</td>
<td>1.64</td>
</tr>
<tr>
<td>CVD+PP</td>
<td>96.0</td>
<td>5.8</td>
<td><strong>50</strong></td>
<td>60.3</td>
<td>2.44</td>
</tr>
</tbody>
</table>

Table 4: Time complexity comparison of different algorithms on the Reddit dataset.

Meanwhile, simper algorithms CV+PP and NS+PP work acceptably on most of the datasets. CV+PP reaches a comparable accuracy with Exact for all datasets except PPI. NS+PP works slightly worse but the final validation accuracy is still within 2%. These algorithms can be adopted if there is no strong need for predictive performance. We however emphasize that exact algorithms must be used for making predictions, as we will show in Sec. 5.4. Finally, the algorithm NS without preprocessing works much worse than others, indicating the significance of our preprocessing strategy.

5.4 FURTHER ANALYSIS ON TIME COMPLEXITY, TESTING ACCURACY AND VARIANCE

Table 4 reports the average number of epochs, time, and total number of floating point operations to reach a given 96% validation accuracy on the largest Reddit dataset. Sparse and dense computations are defined in Sec. 4.6. We found that CVD+PP is about 5 times faster than Exact due to the significantly reduced receptive field size. Meanwhile, simply setting $D^{(l)} = 2$ for NS does not converge to the given accuracy.

We compare the quality of the predictions made by different algorithms, using the same model trained by Exact in Fig. 4. As Thm. 1 states, CV reaches the same testing accuracy as Exact, while NS and NS+PP perform much worse. Testing using exact algorithms (CV or Exact) corresponds to the weight scaling algorithm for dropout (Srivastava et al., 2014).

Finally, we compare the average bias and variance of the gradients per dimension for first layer weights relative to the weights’ magnitude in Fig. 5. For models without dropout, the gradient of CV+PP is almost unbiased. For models with dropout, the bias and variance of CV+PP and CVD+PP are usually smaller than NS and NS+PP, as we analyzed in Sec. 4.3.

6 CONCLUSIONS

The large receptive field size of GCN hinders its fast stochastic training. In this paper, we present a preprocessing strategy and two control variate based algorithms to reduce the receptive field size. Our algorithms can achieve comparable convergence speed with the exact algorithm even the neighbor sampling size $D^{(l)} = 2$, so that the per-epoch cost of training GCN is comparable with training MLPs. We also present strong theoretical guarantees, including exact prediction and convergence to GCN’s local optimum, for our control variate based algorithm.
REFERENCES


A Proof of Theorem 1

Proof. 1. We prove by induction. After the first epoch the activation $h_{CV_i}^{(0)}$ is at least computed once for each node $v$, so $\bar{H}^{(0)}_{CV_i} = H^{(0)}_{CV_i} = H^{(0)}$ for all $i > I$. Assume that we have $\bar{H}^{(l)}_{CV_i} = H^{(l)}_{CV_i} = H^{(l)}$ for all $i > (l + 1)I$. Then for all $i > (l + 1)I$

$$Z_{CV,i}^{(l+1)} = \left( \hat{P}_i^{(l)}(H^{(l)}_{CV,i} - \bar{H}^{(l)}_{CV,i}) + P \bar{H}^{(l)}_{CV,i} \right) W^{(l)} = P \bar{H}^{(l)}_{CV,i} W^{(l)} = PH^{(l)}W^{(l)} = Z^{(l+1)}.$$  

(7)

After one more epoch, all the activations $h_{CV_i,v}^{(l+1)}$ are computed at least once for each $v$, so $\bar{H}^{(l+1)}_{CV_i} = H^{(l+1)}_{CV_i} = H^{(l+1)}$ for all $i > (l + 2)I$. By induction, we know that after LI steps, we have $\bar{H}^{(L+1)}_{CV_i} = H^{(L+1)}_{CV_i} = H^{(L)}$. By Eq. (4) we have $\bar{Z}^{(L)}_{CV,i} = Z^{(L)}$.

2. We omit the time subscript $t$ and denote $f_{CV,v} := f(y_v, z^{(L)}_{CV,v})$. By back propagation, the approximated gradients by CV can be computed as follows

$$\nabla_{H^{(l)}_{CV,v}} f_{CV,v} = \hat{P}_i^{(l)} \nabla_{Z^{(l+1)}_{CV,v}} f_{CV,v} W^{(l)\top} \quad l = 1, \ldots, L - 1$$

$$\nabla_{Z^{(l)}_{CV,v}} f_{CV,v} = \sigma'(Z^{(l)}_{CV,v}) \circ \nabla_{H^{(l)}_{CV,v}} f_{CV,v} \quad l = 1, \ldots, L - 1$$

$$\nabla_{W^{(l)}_{CV,v}} f_{CV,v} = \left( \hat{P}_i^{(l)} H^{(l)}_{CV,v} \right)^{\top} \nabla_{Z^{(l+1)}_{CV,v}} f_{CV,v} \quad l = 0, \ldots, L - 1,$nabla_{W^{(l)}_{CV,v}} f_{CV,v} = \left( \hat{P}_i^{(l)} H^{(l)}_{CV,v} \right)^{\top} \nabla_{Z^{(l+1)}_{CV,v}} f_{CV,v} \quad l = 0, \ldots, L - 1,$

$$g_{CV}(W) = \frac{1}{|V_B|} \sum_{v \in V_B} \nabla_W f_{CV,v},$$

(8)

Application $E_\hat{P}_i = E_{\hat{P}_i^{(1)}, \ldots, \hat{P}_i^{(L)}}$ to both sides of Eq. (8) and utilizing

- $1'$s conclusion that after $L$ epochs, $Z^{(l)}_{CV} = Z^{(l)}$, so $\nabla_{Z^{(l)}_{CV,v}} f_{CV,v}$ is also deterministic.

- $E_{\hat{P}_i} [\nabla_{Z^{(l)}_{CV,v}} f_{CV,v}] = E_{\hat{P}_i^{(1)}, \ldots, \hat{P}_i^{(L)}} [\nabla_{Z^{(l)}_{CV,v}} f_{CV,v}]$.

- $E_{\hat{P}_i} [\nabla_{W^{(l)}_{CV,v}} f_{CV,v}] = E_{\hat{P}_i^{(1)}, \ldots, \hat{P}_i^{(L)}} [\nabla_{W^{(l)}_{CV,v}} f_{CV,v}]$.

we have

$$E_{\hat{P}_i^{(1)}, \ldots, \hat{P}_i^{(L)}} \nabla_{H^{(l)}_{CV,v}} f_{CV,v} = E_{\hat{P}_i^{(1)}, \ldots, \hat{P}_i^{(L)}} \nabla_{Z^{(l+1)}_{CV,v}} f_{CV,v} W^{(l)\top} \quad l = 1, \ldots, L - 1$$

$$E_{\hat{P}_i^{(1)}, \ldots, \hat{P}_i^{(L)}} \nabla_{Z^{(l)}_{CV,v}} f_{CV,v} = \sigma'(Z^{(l)}_{CV,v}) \circ E_{\hat{P}_i^{(1)}, \ldots, \hat{P}_i^{(L)}} \nabla_{H^{(l)}_{CV,v}} f_{CV,v} \quad l = 1, \ldots, L - 1$$

$$E_{\hat{P}_i} \nabla_{W^{(l)}_{CV,v}} f_{CV,v} = \left( \hat{P}_i^{(l)} H^{(l)}_{CV,v} \right)^{\top} E_{\hat{P}_i^{(1)}, \ldots, \hat{P}_i^{(L)}} \nabla_{Z^{(l+1)}_{CV,v}} f_{CV,v} \quad l = 0, \ldots, L - 1.$$g_{CV}(W) = \frac{1}{|V_B|} \sum_{v \in V_B} E_{\hat{P}_i} \nabla_W f_{CV,v},$$

(9)

Comparing Eq. (10) and Eq. (8), we get

$$E_{\hat{P}_i} \nabla_{W^{(l)}_{CV,v}} f_{CV,v} = \nabla_{W^{(l)}_{CV,v}} f_{CV,v}, \quad l = 0, \ldots, L - 1.$$
so

$$E_{\rho,V_B} g_{CV}(W) = E_{V_B} \frac{1}{|V_B|} \sum_{v \in V_B}^{} E_{\rho} \nabla_W f_{CV,v} = \frac{1}{V} \sum_{v \in V} \nabla_W f_v.$$ 

\[ \square \]

**B PROOF OF THEOREM 2**

We proof Theorem 2 in 3 steps:

1. **Lemma 1** For a sequence of weights $W^{(1)}, \ldots, W^{(N)}$ which are close to each other, CV’s approximate activations are close to the exact activations.

2. **Lemma 2** For a sequence of weights $W^{(1)}, \ldots, W^{(N)}$ which are close to each other, CV’s gradients are close to be unbiased.

3. **Theorem 2** An SGD algorithm generates the weights that changes slow enough for the gradient bias goes to zero, so the algorithm converges.

The following proposition is needed in our proof

**Proposition 1.** Let $\|A\|_\infty = \max_{ij} |A_{ij}|$, then

- $\|AB\|_\infty \leq \text{col}(A) \|A\|_\infty \|B\|_\infty$, where $\text{col}(A)$ is the number of columns of the matrix $A$.
- $\|A \circ B\|_\infty \leq \|A\|_\infty \|B\|_\infty$.
- $\|A + B\|_\infty \leq \|A\|_\infty + \|B\|_\infty$.

**Proof.**

$$\|AB\|_\infty = \max_{ij} \left| \sum_k A_{ik} B_{kj} \right| \leq \max_{ij} \left| \sum_k A_{ik} \|B\|_\infty \right| = \text{col}(A) \|A\|_\infty \|B\|_\infty .$$

$$\|A \circ B\|_\infty = \max_{ij} |A_{ij} B_{ij}| \leq \max_{ij} \|A\|_\infty \|B\|_\infty = \|A\|_\infty \|B\|_\infty .$$

$$\|A + B\|_\infty = \max_{ij} |A_{ij} + B_{ij}| \leq \max_{ij} \{|A_{ij}| + |B_{ij}| \} \leq \max_{ij} |A_{ij}| + \max_{ij} |B_{ij}| = \|A\|_\infty + \|B\|_\infty .$$

\[ \square \]

We define $C := \max\{\text{col}(P), \text{col}(H^{(0)}), \ldots, \text{col}(H^{(L)})\}$.

**B.1 PROOF OF LEMMA 1**

**Proposition 2.** There are a series of $T$ inputs $X_1, \ldots, X_T$, $X_{CV,1}, \ldots, X_{CV,T}$ and weights $W_1, \ldots, W_T$ feed to an one-layer GCN with CV

$$Z_{CV,i} = \left( \tilde{P}_i (X_i - \bar{X}_i) + P \tilde{X}_i \right) W_i, \quad H_{CV,i} = \sigma(Z_{CV,i}), \quad \bar{H}_{CV,i+1} = s_i H_{CV,i} + (1-s_i) \tilde{H}_{CV,i},$$

and an one-layer exact GCN

$$Z_i = PX_i W_i, \quad H_i = \sigma(Z_i).$$

If

1. The activation $\sigma(\cdot)$ is $\rho$-Lipschitz;

2. $\|X_{CV,i} - X_{CV,j}\|_\infty < \epsilon$ and $\|X_{CV,i} - X_i\|_\infty < \epsilon$ for all $i, j \leq T$ and $\epsilon > 0$.

Then there exists some $K > 0$, s.t., $\|H_{CV,i} - H_{CV,j}\|_\infty < K\epsilon$ and $\|H_{CV,i} - H_i\|_\infty < K\epsilon$ for all $I < i, j \leq T$, where $I$ is the number of iterations per epoch.
\textbf{Proof.} Because for all $i > I$, the elements of $\bar{X}_{CV,i}$ are all taken from previous epochs, i.e., $X_{CV,1}, \ldots, X_{CV,i-1}$, we know that
\begin{equation}
\|X_{CV,i} - X_{CV,i}\|_{\infty} \leq \max_{j \leq i} \|X_{CV,j} - X_{CV,i}\| \leq \epsilon \quad (\forall i > I).
\end{equation}

By triangular inequality, we also know
\begin{align*}
\|\bar{X}_{CV,i} - X_{CV,i}\|_{\infty} &< 3\epsilon \quad (\forall i, j > I). \\
\|\bar{X}_{CV,i} - X_{i}\|_{\infty} &< 2\epsilon \quad (\forall i > I).
\end{align*}

Since $\|X_{CV,1}\|_{\infty}, \ldots, \|X_{CV,T}\|_{\infty}$ are bounded, $\|\bar{X}_{CV,i}\|_{\infty}$ is also bounded for $i > I$. Then,
\begin{align*}
\|H_{CV,i} - H_{CV,j}\|_{\infty} &\leq \rho \|Z_{CV,i} - Z_{CV,j}\|_{\infty} \\
&\leq \rho \left\| \left( \hat{P}_i(X_{CV,i} - X_{CV,j}) + P\bar{X}_{CV,i} \right) W_i - \left( \hat{P}_j(X_{CV,j} - X_{CV,j}) + P\bar{X}_{CV,j} \right) W_j \right\|_{\infty} \\
&\leq \rho \left\| \hat{P}_i(X_{CV,i} - X_{CV,j})W_i - \hat{P}_j(X_{CV,j} - X_{CV,j})W_j \right\|_{\infty} + \rho \|P\bar{X}_{CV,i}W_i - P\bar{X}_{CV,j}W_j\|_{\infty} \\
&\leq \rho \|C^2(\|\hat{P}_i - \hat{P}_j\|_{\infty} \|X_{CV,i} - X_{CV,j}\|_{\infty} \|W_i\|_{\infty}) \\
&+ \|\hat{P}_j\|_{\infty} \|X_{CV,i} - X_{CV,j}\|_{\infty} \|\bar{X}_{CV,j}\|_{\infty} \|W_i\|_{\infty} \\
&+ \|\hat{P}_j\|_{\infty} \|X_{CV,j} - \bar{X}_{CV,j}\|_{\infty} \|W_i - W_j\|_{\infty} \\
&+ \|P\|_{\infty} \|\bar{X}_{CV,i} - \bar{X}_{CV,j}\|_{\infty} \|W_i\|_{\infty} + \|P\|_{\infty} \|\bar{X}_{CV,j}\|_{\infty} \|W_i - W_j\|_{\infty} \right\|_{\infty} \\
&\leq \rho \|C^2(\|\hat{P}_i - \hat{P}_j\|_{\infty} \|W_i\|_{\infty}) + 2 \|\hat{P}_j\|_{\infty} \|W_i\|_{\infty} + \|\hat{P}_j\|_{\infty} \|W_i - W_j\|_{\infty} \right\|_{\infty} \\
&= K_1 \epsilon,
\end{align*}

and
\begin{align*}
\|H_{CV,i} - H_i\|_{\infty} &\leq \rho \|Z_{CV,i} - Z_i\|_{\infty} \\
&\leq \rho \left\| \left( \hat{P}_i(X_{CV,i} - X_i) + P(\bar{X}_{CV,i} - X_i) \right) \right\|_{\infty} W_i \\
&\leq \rho \|C(\|\hat{P}_i\|_{\infty} \epsilon + 2 \|P\|_{\infty} \epsilon) \|W_i\|_{\infty} \\
&\leq K_2 \epsilon.
\end{align*}

The following lemma bounds CV’s approximation error of activations

\textbf{Lemma 1.} Given a sequence of model weights $W_1, \ldots, W_T$. If $\|W_i - W_j\|_{\infty} < \epsilon, \forall i, j$, and all the activations are $\rho$-Lipschitz, there exists $K > 0$, s.t.,
\begin{itemize}
  \item $\|H^l_i - H^l_{CV,i}\|_{\infty} < K \epsilon, \forall i > LI, l = 1, \ldots, L - 1,$
  \item $\|Z_i - Z_{CV,i}\|_{\infty} < K \epsilon, \forall i > LI, l = 1, \ldots, L$.
\end{itemize}

\textbf{Proof.} We prove by induction. Because $H^0 = X$ is constant, $\bar{H}_{CV,i}^0 = H_i^0$ after $I$ iterations. So $H_{CV,i}^1 = \sigma(\hat{P}_i(H_{CV,i}^0 - \bar{H}_{CV,i}^0) + PH_{CV,i}^0) W_i = \sigma(PXW_i^0) = H_i^1$, and
\begin{align*}
\|H_{CV,i}^1 - H_{CV,j}^1\|_{\infty} = \|\sigma(PXW_i^0) - \sigma(PXW_j^0)\|_{\infty} \leq \rho C \|P\|_{\infty} \|X\|_{\infty} \epsilon.
\end{align*}

Repeatedly apply Proposition B.1 for $L - 1$ times, we get the intended results. \qed
B.2 Proof of Lemma 2

The following lemma bounds the bias of CV’s approximate gradient.

Lemma 2. Given a sequence of model weights $W_1, \ldots, W_T$, if

1. $||W_i - W_j||_\infty < \epsilon, \forall i, j$,

2. all the activations are $\rho$-Lipschitz,

3. the gradient of the cost function $\nabla_z f(y, z)$ is $\rho$-Lipschitz and bounded.

then there exists $K > 0$, s.t.,

$$\left\| \mathbb{E}_{P} g_{CV}(W_i) - g(W_i) \right\|_\infty < K \epsilon, \forall i > LI.$$

Proof. By Lipschitz continuity of $\nabla_z f(y, z)$ and Lemma 1, there exists $K > 0$, s.t.,

$$\left\| \nabla_{Z_{CV}} f_{CV,v} - \nabla Z(0) f_i \right\|_\infty < \rho K \epsilon.$$

Assume that $\left\| \mathbb{E}_{P} \nabla_{Z_{CV}}(l+1) f_{CV,v} - \nabla Z(0) f_i \right\|_\infty < K_1 \epsilon$, we now prove that there exists $K > 0$, s.t.,

$$\left\| \mathbb{E}_{P} \nabla_{Z_{CV}} f_{CV,v} - \nabla Z(0) f_i \right\|_\infty < K \epsilon.$$

By Eq. 9, Eq. 10 and Lemma 1, we have

$$\left\| \mathbb{E}_{P} \nabla_{Z_{CV}} f_{CV,v} - \nabla Z(0) f_i \right\|_\infty < K \epsilon.$$

and

$$\left\| \mathbb{E}_{P} \nabla_{Z_{CV}} f_{CV,v} - \nabla Z(0) f_i \right\|_\infty < K \epsilon.$$

Again by Eq. 9, Eq. 10 and Lemma 1,

$$\left\| \mathbb{E}_{P} \nabla_{W} f_{CV,v} - \nabla W(0) f_v \right\|_\infty < K \epsilon$$
Finally,
\[
\left\| \mathbb{E}_{\hat{P}, \mathcal{V}_b} g_{CV}(W_i) - g(W_i) \right\|_{\infty} \\
= \left\| \mathbb{E}_{\mathcal{V}_b} \left( \frac{1}{|\mathcal{V}_b|} \sum_{v \in \mathcal{V}_b} \mathbb{E}_{\hat{P}} [\nabla_{W(v)} f_{CV,v}] - \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \nabla_{W(v)} f_v \right) \right\|_{\infty} \\
= \left\| \frac{1}{|\mathcal{V}|} \sum_{v \in \mathcal{V}} \left( \mathbb{E}_{\hat{P}} [\nabla_{W(v)} f_{CV,v}] - \nabla_{W(v)} f_v \right) \right\|_{\infty} \leq K_3 \epsilon.
\]

\[\square\]

\[\text{B.3 Proof of Theorem 2}\]

\textbf{Proof}. This proof is a modification of [Ghadimi & Lan (2013)], but using biased stochastic gradients instead. We assume the algorithm is already warmed-up for \(L_I\) steps with the initial weights \(W_0\), so that Lemma 2 holds for step \(i > 0\). Denote \(\delta_i = g_{CV}(W_i) - \nabla \mathcal{L}(W_i)\). By smoothness we have

\[
\mathcal{L}(W_{i+1}) \leq \mathcal{L}(W_i) + \langle \nabla \mathcal{L}(W_i), W_{i+1} - W_i \rangle + \frac{\rho}{2} \gamma_i^2 \|g_{CV}(W_i)\|^2 \\
= \mathcal{L}(W_i) - \gamma_i \langle \nabla \mathcal{L}(W_i), g_{CV}(W_i) \rangle + \frac{\rho}{2} \gamma_i^2 \|g_{CV}(W_i)\|^2 \\
= \mathcal{L}(W_i) - \gamma_i \langle \nabla \mathcal{L}(W_i), \delta_i \rangle - \gamma_i \|\nabla \mathcal{L}(W_i)\|^2 + \frac{\rho}{2} \gamma_i^2 \left[ \|\delta_i\|^2 + \|\nabla \mathcal{L}(W_i)\|^2 + 2 \langle \delta_i, \nabla \mathcal{L}(W_i) \rangle \right] \\
= \mathcal{L}(W_i) - (\gamma_i - \rho \gamma_i^2) \langle \nabla \mathcal{L}(W_i), \delta_i \rangle - (\gamma_i - \frac{\rho \gamma_i^2}{2}) \|\nabla \mathcal{L}(W_i)\|^2 + \frac{\rho}{2} \gamma_i^2 \|\delta_i\|^2. \tag{14}
\]

Consider the sequence of \(L + 1\) weights \(W_{i-L}, \ldots, W_i\).

\[
\max_{i-L \leq j \leq i} \|W_j - W_k\|_{\infty} \leq \sum_{j=i-L}^{i-1} \|W_j - W_{j+1}\|_{\infty} \\
= \sum_{j=i-L}^{i-1} \gamma_j \|g_{CV}(W_j)\|_{\infty} \leq \sum_{j=i-L}^{i-1} \gamma_j \leq LIG \gamma_i - L_I.
\]

By Lemma 2 there exists \(K > 0\), s.t.

\[
\mathbb{E}_{\hat{P}, \mathcal{V}_b} \|\delta\|_{\infty} = \mathbb{E}_{\hat{P}, \mathcal{V}_b} \|g_{CV}(W_i) - \nabla \mathcal{L}(W_i)\|_{\infty} \leq K LIG \gamma_i - L_I, \quad \forall i > 0.
\]

Assume that \(W\) is \(D\)-dimensional,

\[
\mathbb{E}_{\hat{P}, \mathcal{V}_b} \langle \nabla \mathcal{L}(W_i), \delta_i \rangle \geq -\mathbb{E}_{\hat{P}, \mathcal{V}_b} D \|\nabla \mathcal{L}(W_i)\|_{\infty} \|\delta_i\|_{\infty} \geq -K LID G^2 \gamma_i - L_I = K_1 \gamma_i - L_I, \\
\mathbb{E}_{\hat{P}, \mathcal{V}_b} \|\delta_i\|^2 \leq D \left( \mathbb{E}_{\hat{P}, \mathcal{V}_b} \|\delta_i\|_{\infty} \right)^2 \leq DK^2 L^2 B^2 G^2 \gamma_i - L_I = K_2 \gamma_i - L_I,
\]

where \(K_1 = K LID G^2\) and \(K_2 = DK^2 L^2 B^2 G^2\). Taking \(\mathbb{E}_{\hat{P}, \mathcal{V}_b}\) to both sides of Eq. (14) we have

\[
\mathcal{L}(W_{i+1}) \leq \mathcal{L}(W_i) + (\gamma_i - \rho \gamma_i^2) K_1 \gamma_i - L_I - (\gamma_i - \frac{\rho \gamma_i^2}{2}) \mathbb{E}_{\hat{P}, \mathcal{V}_b} \|\nabla \mathcal{L}(W_i)\|^2 + \frac{\rho}{2} \gamma_i^2 K_2 \gamma_i - L_I.
\]

Summing up the above inequalities and re-arranging the terms, we obtain,

\[
\sum_{i=1}^{N} (\gamma_i - \frac{\rho \gamma_i^2}{2}) \mathbb{E}_{\hat{P}, \mathcal{V}_b} \|\nabla \mathcal{L}(W_i)\|^2 \\
\leq \mathcal{L}(W_1) - \mathcal{L}(W^*) + K_1 \sum_{i=1}^{N} (\gamma_i - \rho \gamma_i^2) \gamma_i - L_I + \frac{\rho K_2}{2} \sum_{i=1}^{N} \gamma_i^2 \gamma_i - L_I.
\]
Dividing both sides by \( \sum_{i=1}^{N} (\gamma_i - \frac{\rho \gamma_i^2}{2}) \),

\[
\mathbb{E}_{R \sim P_R} \mathbb{E}_{\hat{p}, V_B} \| \nabla \mathcal{L}(W_R) \|^2 \\
\leq 2 \frac{\mathcal{L}(W_1) - \mathcal{L}(W^*) + K_1 \sum_{i=1}^{N} (\gamma_i - \rho \gamma_i^2) \gamma_i - L_T + \frac{\rho K_2}{2} \sum_{i=1}^{N} \gamma_i^2 \gamma_i - L_T}{\sum_{i=1}^{N} \gamma_i (2 - \rho \gamma_i)}.
\]

Taking \( \gamma_i = \gamma := \min \{ \frac{1}{\rho}, \frac{1}{\sqrt{N}} \} \), for all \( i = 1, \ldots, N \), we have

\[
\mathbb{E}_{R \sim P_R} \mathbb{E}_{\hat{p}, V_B} \| \nabla \mathcal{L}(W_R) \|^2 \\
\leq 2 \frac{\mathcal{L}(W_1) - \mathcal{L}(W^*) + K_1 N (\gamma - \rho \gamma^2) \gamma + \frac{\rho K_2}{2} N \gamma^3}{N \gamma (2 - \rho \gamma)} \\
\leq 2 \frac{\mathcal{L}(W_1) - \mathcal{L}(W^*) + K_1 (\gamma - \rho \gamma^2) \gamma + \frac{\rho K_2}{2} N \gamma^3}{N \gamma} \\
\leq 2 \frac{\mathcal{L}(W_1) - \mathcal{L}(W^*) + K_1 \gamma (1 - \rho \gamma) + \rho K_2 \gamma^2}{N \gamma} \\
\leq \frac{2 \rho (\mathcal{L}(W_1) - \mathcal{L}(W^*)) + K_2}{N} + \frac{2 (\mathcal{L}(W_1) - \mathcal{L}(W^*)) + K_1}{\sqrt{N}}.
\]

Particularly, when \( N \to \infty \), we have \( \mathbb{E}_{R \sim P_R} \mathbb{E}_{\hat{p}, V_B} \| \nabla \mathcal{L}(W_R) \|^2 = 0 \), which implies that the gradient is asymptotically unbiased.

\[ \square \]

C Derivation of the Variance

\[
\text{Var}[u] = \mathbb{E} \left[ \sum_v p_v (h_v - \mu_v)^2 \right] \\
= \sum_v p_v^2 \mathbb{E} [ (h_v - \mu_v)^2 ] \\
= \sum_v p_v^2 s_v^2 \\
= s^2.
\]

\[
\text{Var}[u_{NS}] = \mathbb{E} [ D p_v h_v - \mu ]^2 \\
= \mathbb{E} [ D p_v^2 (\mu_v^2 + s_v^2) + \mu^2 - D p_v \mu_v \mu_p ] \\
= D s^2 + (D \sum_v p_v^2 \mu_v^2 - \mu^2) \\
= D s^2 + \frac{1}{2} \sum_{v,v'} (p_v \mu_v - p_{v'} \mu_{v'})^2.
\]
In this section, we describe the details of our model architectures. We use the Adam optimizer [Kingma & Ba, 2014] with learning rate 0.01.

- Citeseer, Cora, PubMed and NELL: We use the same architecture as Kipf & Welling (2017): two graph convolution layers with one linear layer per graph convolution layer. We use 32 hidden units, 50% dropout rate and $5 \times 10^{-4}$ L2 weight decay for Citeseer, Cora and PubMed and 64 hidden units, 10% dropout rate and $10^{-5}$ L2 weight decay for NELL.

- PPI and Reddit: We use the mean pooling architecture proposed by Hamilton et al. (2017a). We use two linear layers per graph convolution layer. We set weight decay to be zero, dropout rate to be 0.2%, and adopt layer normalization (Ba et al., 2016) after each linear layer. We use 512 hidden units for PPI and 128 hidden units for Reddit. We find that our architecture can reach 97.8% testing micro-F1 on the PPI dataset, which is significantly higher than 59.8% reported by Hamilton et al. (2017a). We find the improvement is from wider hidden layer, dropout and layer normalization.
Algorithm 1 Training with the CV algorithm

for each minibatch $V_B \subset V$ do
  Randomly sample propagation matrices $\tilde{P}^{(0)}, \ldots, \tilde{P}^{(L-1)}$
  Compute the receptive fields $m^{(0)}, \ldots, m^{(L-1)}$
  (Forward propagation)
  for each layer $l \leftarrow 0$ to $L - 1$ do
    $Z^{(l+1)} \leftarrow \left(\tilde{P}^{(l)}(H^{(l)} - \bar{H}^{(l)} + P\bar{H}^{(l)})\right)W^{(l)}$
    $H^{(l+1)} \leftarrow \sigma(Z^{(l+1)})$
  end for
  Compute the loss $L = \frac{1}{|V_B|} \sum_{v \in V_B} f(y_v, Z^{(L)}_v)$
  (Backward propagation)
  $W \leftarrow W - \gamma_i \nabla_W L$
  (Update historical activations)
  for each layer $l \leftarrow 0$ to $L - 1$ do
    $\bar{H}^{(l)} \leftarrow m^{(l)}H^{(l)} + (1 - m^{(l)})\bar{H}^{(l)}$
  end for
end for

Algorithm 2 Training with the CVD algorithm

for each minibatch $V_B \subset V$ do
  Randomly sample propagation matrices $\tilde{P}^{(0)}, \ldots, \tilde{P}^{(L-1)}$
  Compute the receptive fields $m^{(0)}, \ldots, m^{(L-1)}$
  (Forward propagation)
  for each layer $l \leftarrow 0$ to $L - 1$ do
    $U \leftarrow \left(\tilde{P}^{(l)}(H^{(l)} - \mu^{(l)}) + \tilde{P}^{(l)}(\mu^{(l)} - \bar{\mu}^{(l)}) + P\bar{H}^{(l)}\right)$
    $H^{(l+1)} \leftarrow \sigma(\text{Dropout}_p(U)W^{(l)})$
    $\mu^{(l+1)} \leftarrow \sigma(UW^{(l)})$
  end for
  Compute the loss $L = \frac{1}{|V_B|} \sum_{v \in V_B} f(y_v, H^{(L)}_v)$
  (Backward propagation)
  $W \leftarrow W - \gamma_i \nabla_W L$
  (Update historical activations)
  for each layer $l \leftarrow 0$ to $L - 1$ do
    $\bar{\mu}^{(l)} \leftarrow m^{(l)}\mu^{(l)} + (1 - m^{(l)})\bar{\mu}^{(l)}$
  end for
end for

E PSEUDOCODE

E.1 TRAINING WITH THE CV ESTIMATOR

Alg. 1 depicts the training algorithm using the CV estimator in Sec. 4.1. We perform forward propagation according to Eq. (4), compute the stochastic gradient, and then update the historical activations $\bar{H}^{(l)}$ according to Eq. (5). We omit the subscripts $CV$ and the iteration number $i$ for concise. Let $W = (W^{(0)}, \ldots, W^{(L-1)})$ be all the trainable parameters, the gradient $\nabla_W L$ is computed automatically by frameworks such as TensorFlow. The diagonal matrix $m^{(l)}$ denotes the receptive field at layer $l$, i.e., the nodes that need to be computed in order to approximate $\hat{z}^{(l)}_v$ for $v$ in the minibatch $V_B$. We only need to compute and update the activations $H^{(l)}, Z^{(l)}, \bar{H}^{(l)}$ for nodes in $m^{(l)}$. 
Figure 6: Comparison of validation accuracy with respect to number of epochs for 3-layer GCNs.

Table 5: Time to reach 0.95 testing accuracy.

<table>
<thead>
<tr>
<th>Alg.</th>
<th>Valid. acc.</th>
<th>Epochs</th>
<th>Time (s)</th>
<th>Sparse GFLOP</th>
<th>Dense TFLOP</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exact</td>
<td>0.940</td>
<td>3.0</td>
<td>199</td>
<td>306</td>
<td>11.7</td>
</tr>
<tr>
<td>NS</td>
<td>0.940</td>
<td>24.0</td>
<td>148</td>
<td>33.6</td>
<td>9.79</td>
</tr>
<tr>
<td>NS+PP</td>
<td>0.940</td>
<td>12.0</td>
<td>68</td>
<td>253</td>
<td>4.89</td>
</tr>
<tr>
<td>CV+PP</td>
<td>0.940</td>
<td>5.0</td>
<td>32</td>
<td>8.06</td>
<td>2.04</td>
</tr>
<tr>
<td>CVD+PP</td>
<td>0.940</td>
<td>5.0</td>
<td>36</td>
<td>16.1</td>
<td>4.08</td>
</tr>
</tbody>
</table>

E.2 Training with the CVD estimator

Training with the CVD estimator is similar with the CV estimator, except it runs two versions of the network, with and without dropout, to compute the samples $H$ and their mean $\mu$ of the activation.

The matrix $P_{v,v'} = \bar{P}_{v,v'}/\sqrt{\text{n}(v,1)}$, where $|\text{n}(v,1)|$ is the degree of node $v$.

F Experiment for 3-layer GCNs

We test 3-layer GCNs on the Reddit dataset. The settings are the same with 2-layer GCNs in Sec. 5.3. To ensure the exact algorithm can run in a reasonable amount of time, we subsample the graph so that the maximum degree is 10. The convergence result is shown as Fig. 6, which is similar with the two-layer models. The time consumption to reach 0.94 testing accuracy is shown in Table 5.

G Justification of the independent Gaussian assumption

We justify the independent Gaussian assumption in Sec. 4.3 by showing that for a 2-layer GCN with the first layer pre-processed, the neighbor’s activations are independent. Without loss of generality, suppose that we want to compute $z^{(2)}_1$, and the neighbors of node 1 are $1, \ldots, D$. By Eq. (1), $h_v^{(1)} = \sigma\left(\phi_v \circ u_v^{(0)}W^{(0)}\right)$ is a random variable with respect to $\phi_v$, where $\phi_v \sim \text{Bernoulli}(p)$ is the dropout mask and $u_v^{(0)} = (PH^{(0)})_v$. The independent Gaussian assumption states that

1. $h_v^{(1)}$ is a Gaussian random variable with diagonal covariance;
2. $h_v^{(1)}$ and $h_{v'}^{(1)}$ are independent, for $v \neq v'$.

Assumption 1 is not GCN-specific and is discussed in [Wang & Manning (2013)], we now prove assumption 2 by the following lemma.

Lemma 3. If $a$ and $b$ are independent random variables, then their transformations $f_1(a)$ and $f_2(b)$ are independent.

Because for any event $A$ and $B$, $P(f_1(a) \in f_1(A), f_2(b) \in f_2(B)) = P(a \in A, b \in B) = P(a \in A)P(b \in B) = P(f_1(a) \in f_1(A))P(f_2(b) \in f_2(B))$, where $f_1(A) = \{f_1(a) | a \in A\}$ and $f_2(B) = \{f_2(b) | b \in B\}$.

Let $h_v^{(1)} = f_1(\phi_v) := \sigma\left(\phi_v \circ u_v^{(0)}W^{(0)}\right)$ and $h_{v'}^{(1)} = f_1(\phi_{v'}) := \sigma\left(\phi_{v'} \circ u_{v'}^{(0)}W^{(0)}\right)$, because $\phi_v$ and $\phi_{v'}$ are independent Bernoulli random variables, $h_v^{(1)}$ and $h_{v'}^{(1)}$ are independent.