APPENDIX
THEORETICAL ANALYSIS OF SUGAR

In this appendix, we provide details of the following theoretical results used in Section III:

(1) **Lemma 1**: For a multi-layer GCN with fixed weights, the error of the activations of the SG estimator are bounded.

(2) **Lemma 2**: For a multi-layer GCN with fixed weights, the error of the gradients of the SG estimator are bounded.

(3) **Theorem 1**: With high probability gradient descent training with the approximated gradients by the SG estimator can converge to a local minimum.

The proof builds on [12], but with different assumptions. More precisely, while [12] assume that model weights change slowly during training, our theoretical analysis is based on the difference in the adjacency matrices produced by graph partitioning.

A. Notations

Let $[L] = \{1, ..., L\}$. The infinity norm of a matrix is defined as $\|A\|_\infty = \max_{i,j} |A_{i,j}|$. By Proposition B in [12], we know that:

1) $\|AB\|_\infty \leq col(A)\|A\|_\infty \|B\|_\infty$
2) $\|A \circ B\|_\infty \leq \|A\|_\infty \|B\|_\infty$
3) $\|A + B\|_\infty \leq \|A\|_\infty + \|B\|_\infty$

where $col(A)$ represents the number of columns of matrix A and $\circ$ is the element-wise product. We define η to be the maximum number of columns we can possibly encounter in the proof. We review some notations defined in the main text. Our proposed estimator is denoted by SG. The propagation rule of a l-th layer GCN with the exact estimator is given by:

$$Z^{(l+1)} = A^{norm} H^{(l)} W^{(l)}, H^{(l+1)} = \sigma(Z^{(l+1)}) \quad (14)$$

Similarly, the propagation rule of a l-th layer GCN with the SG estimator is given by:

$$Z_{SG}^{(l+1)} = A_{SG}^{norm} H_{SG}^{(l)} W^{(l)}, H_{SG}^{(l+1)} = \sigma(Z_{SG}^{(l+1)}) \quad (15)$$

where σ represents an activation function, A^{norm} denotes the normalized version of A, *i.e.*, $A^{norm} = \hat{D}^{-1/2}\hat{A}\hat{D}^{1/2}, \hat{A} = A + I_N, \hat{D}_{ii} = \sum_j \hat{A}_{ij}$ and I_N is an N-dimensional identity matrix.$H^{(l)}$ and $H_{SG}^{(l)}$ denote node representations in the l-th layer produced by the exact GCN and SG estimator, respectively. $W^{(l)}$ represents the weight matrix in layer l. Note that while we write Equation 15 in a compact matrix form, in real implementation, the training process is distributed across K devices.

Recall that A_{SG} is a block-diagonal matrix produced by the graph partitioning module that serves as an approximation of A. Before training, we run graph partitioning for M times to obtain a sample average, *i.e.*, $A_{SG}^{norm} = \frac{1}{M}\sum_{m=1}^{M} A_{SG,m}^{norm}$. Let $\epsilon = \|A_{SG}^{norm} - A^{norm}\|_\infty$ denote the error in approximating A^{norm} with A_{SG}^{norm}. For simplicity, we will omit the superscript $norm$ from now on.

The model parameters at training epoch t are denoted by W_t. For W at a given time point (*i.e.*, fixed model weights), we omit the subscript in the proof. Let W_* denote the optimal model weights. $\nabla\mathcal{L}(W) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial f(y_i, z_i^{(L)})}{\partial W}$ and $\nabla\mathcal{L}_{SG}(W) = \frac{1}{N}\sum_{i=1}^{N} \frac{\partial f(y_i, z_{SG,i}^{(L)})}{\partial W}$ represent the gradients of the exact GCN and SG estimator with respect to model weights W, respectively. $f(\cdot,\cdot)$ is the objective function (*e.g.*, cross entropy for node classification tasks).

B. Activations of Multi-layer GCN

1) Single-layer GCN: Proposition 2 states that for a single-layer GCN, (1) the outputs are bounded if the inputs are bounded, (2) if the difference between the input of the SG estimator and the exact GCN is small, then the output of the SG estimator is close to the output of the exact GCN.

Proposition 2. *For a one-layer GCN, if the activation function $\sigma(\cdot)$ is ρ-Lipschitz and $\sigma(0) = 0$, for any input matrices A, A_{SG}, X, X_{SG} and any weight matrix W that satisfy:*

1) *All the matrices are bounded by β: $\|A\|_\infty \leq \beta$, $\|A_{SG}\|_\infty \leq \beta$, $\|X\|_\infty \leq \beta$, $\|X_{SG}\|_\infty \leq \beta$ and $\|W\|_\infty \leq \beta$,*
2) *The differences between inputs are bounded: $\|X_{SG} - X\|_\infty \leq \alpha\epsilon$, where $\epsilon = \|A_{SG} - A\|_\infty$.*

Then, there exist B and C that depend on ρ, η and β, s.t.,

1) *The outputs are bounded: $\|H\|_\infty \leq B$ and $\|H_{SG}\|_\infty \leq B$,*
2) *The differences between outputs of the SG estimator and the exact estimator are bounded: $\|Z_{SG} - Z\|_\infty \leq C(1+\alpha)\epsilon$ and $\|H_{SG} - H\|_\infty \leq C(1+\alpha)\epsilon$.*

Proof. We know that $\|Z\|_\infty = \|AXW\|_\infty \leq \eta^2\|A\|_\infty\|X\|_\infty\|W\|_\infty \leq \eta^2\beta^3$. By Lipschitz continuity of $\sigma(\cdot)$, $\|\sigma(Z) - \sigma(0)\|_\infty \leq \rho\eta^2\beta^3$ and we

have $\|\sigma(Z)\|_\infty \leq \rho\eta^2\beta^3$. Thus $\|H\|_\infty \leq D$, where $B = \max\{\eta^2\beta^3, \rho\eta^2\beta^3\}$. Similarly, $\|H_{SG}\|_\infty \leq B$.

We proceed to show that the differences between outputs are bounded below:

$$\begin{aligned}
&\|Z_{SG} - Z\|_\infty \\
&= \|A_{SG}X_{SG}W - AXW\|_\infty \\
&\leq \eta\|W\|_\infty\|A_{SG}X_{SG} - AX\|_\infty \\
&\leq \eta\beta(\|A_{SG}(X_{SG} - X)\|_\infty + \|X(A_{SG} - A)\|_\infty) \\
&\leq \eta\beta(\eta\beta\alpha\epsilon + \eta\beta\epsilon) \\
&= (1+\alpha)\eta^2\beta^2\epsilon
\end{aligned} \tag{16}$$

By Lipschitz continuity of $\sigma(\cdot)$, we have $\|H_{SG} - H_t\|_\infty \leq \rho(1+\alpha)\eta^2\beta^2\epsilon$. Choose $C = \max\{(1+\alpha)\eta^2\beta^2, \rho(1+\alpha)\eta^2\beta^2\}$, and the proof is complete.

2) Multi-layer GCN: The following lemma relates the approximation error in activations (*i.e.*, $\left\|H_{SG}^{(l)} - H^{(l)}\right\|_\infty$) with the approximation error in input adjacency matrices (*i.e.*, $\epsilon = \|A_{SG} - A\|_\infty$).

Lemma 1. *For a multi-layer GCN with fixed model weights, given a (fixed) graph dataset, assume that:*

1) *$\sigma(\cdot)$ is ρ-Lipschitz and $\sigma(0) = 0$,*
2) *The inputs are bounded by β: $\|A\|_\infty \leq \beta$, $\|A_{SG}\|_\infty \leq \beta$, $\|X\|_\infty \leq \beta$,*
3) *The model weights in each layer are bounded by β: $\left\|W^{(l)}\right\|_\infty \leq \beta, \forall l \in [L]$.*

Then, there exist B and C that depend on ρ, η and β, s.t.,

1) *$\left\|H^{(l)}\right\|_\infty \leq B, \left\|H_{SG}^{(l)}\right\|_\infty \leq B, \ \forall l \in [L-1]$,*
2) *$\left\|Z_{SG}^{(l)} - Z^{(l)}\right\|_\infty \leq C\epsilon, \ \forall l \in [L]$ and $\left\|H_{SG}^{(l)} - H^{(l)}\right\|_\infty \leq C\epsilon, \ \forall l \in [L-1]$.*

Proof. Applying Proposition 2 to each layer of the GCN proves that $H^{(l)}$ and $H_{SG}^{(l)}$ are bounded for each layer l.

For the first layer of GCN, by Proposition 2 and input conditions, we know that there exists $C^{(1)}$ that satisfies:

$$\left\|Z_{SG}^{(1)} - Z^{(1)}\right\|_\infty \leq C^{(1)}\epsilon, \quad \left\|H_{SG}^{(1)} - H^{(1)}\right\|_\infty \leq C^{(1)}\epsilon$$

Note that for the first layer, the node feature matrix of the SG estimator and exact GCN are identical, *i.e.*, $X_{SG} = X$; this yields $\alpha = 0$ in Equation 16.

Let $\hat{C}^{(1)} = C^{(1)}$. Next, we apply Proposition 2 to the second layer of GCN: there exists $C^{(2)}$ that satisfies: $\left\|Z_{SG}^{(2)} - Z^{(2)}\right\|_\infty \leq C^{(2)}(1 + \hat{C}^{(1)})\epsilon$, $\left\|H_{SG}^{(2)} - H^{(2)}\right\|_\infty \leq C^{(2)}(1 + \hat{C}^{(1)})\epsilon$.

Let $\hat{C}^{(2)} = C^{(2)}(1 + \hat{C}^{(1)})$. By applying Proposition 2 to the subsequent layer of GCN repetitively, we have $\hat{C}^{(l+1)} = C^{(l+1)}(1 + \hat{C}^{(l)}), \forall l \in [L-1]$. We choose $C = \max_l \hat{C}^{(l)}$ and complete the proof.

C. Gradients of Multi-layer GCN

Lemma 2 below provides a bound for the difference between gradients of the loss by the SG estimator and the exact GCN (*i.e.*, $\|\nabla\mathcal{L}_{SG}(W) - \nabla\mathcal{L}(W)\|_\infty$). Intuitively, the gradient difference is small if the approximation error in input adjacency matrices (*i.e.*, ϵ) is small.

Lemma 2. *For a multi-layer GCN with fixed model weights, given a (fixed) graph dataset, assume that:*

1) *$\frac{\partial f(y,z)}{\partial z}$ is ρ-Lipschitz and $\left\|\frac{\partial f(y,z)}{\partial z}\right\|_\infty \leq \beta$,*
2) *$\sigma(\cdot)$ is ρ-Lipschitz, $\sigma(0) = 0$ and $\|\sigma'(\cdot)\|_\infty \leq \beta$,*
3) *$\|A\|_\infty \leq \beta$, $\|A_{SG}\|_\infty \leq \beta$, $\|X\|_\infty \leq \beta$, $\left\|W^{(l)}\right\|_\infty \leq \beta, \forall l \in [L]$.*

Then, there exists C that depends on ρ, η and β, s.t., $\|\nabla\mathcal{L}_{SG}(W) - \nabla\mathcal{L}(W)\|_\infty \leq C\epsilon$.

Proof. We begin by proving the following statements:

If the above assumptions hold, then there exist C and D that depends on ρ, η and β, s.t.,

1) *The gradients with respect to the activations of each layer of the SG estimator are close to be unbiased:*

$$\left\|\frac{\partial f}{\partial Z_{SG}^{(l)}} - \frac{\partial f}{\partial Z^{(l)}}\right\|_\infty \leq C\epsilon, \quad \forall l \in [L] \tag{17}$$

2) *The gradients above are bounded:*

$$\left\|\frac{\partial f}{\partial Z_{SG}^{(l)}}\right\|_\infty \leq D\beta, \quad \left\|\frac{\partial f}{\partial Z^{(l)}}\right\|_\infty \leq D\beta, \quad \forall l \in [L] \tag{18}$$

We prove these statements by induction. First we show that Equations 17 and 18 hold true for the final layer of GCN (*i.e.*, $l = L$). By Assumption 1 and Lemma 1, we know that there exists $\hat{C}$ that satisfies:

$$\left\|\frac{\partial f}{\partial Z_{SG}^{(L)}} - \frac{\partial f}{\partial Z^{(L)}}\right\|_\infty \leq \rho\left\|Z_{SG}^{(L)} - Z^{(L)}\right\|_\infty \leq \rho\hat{C}\epsilon \tag{19}$$

Let $C^{(L)} = \rho\hat{C}$ and $D^{(L)} = 1$. Next, suppose the statements hold for layer $l+1$, *i.e.*, there exist $C^{(l+1)}$ and $D^{(l+1)}$ that satisfy:

$$\begin{aligned} \left\| \frac{\partial f}{\partial Z_{SG}^{(l+1)}} - \frac{\partial f}{\partial Z^{(l+1)}} \right\|_\infty &\leq C^{(l+1)}\epsilon, \\ \left\| \frac{\partial f}{\partial Z_{SG}^{(l+1)}} \right\|_\infty &\leq D^{(l+1)}\beta, \\ \left\| \frac{\partial f}{\partial Z^{(l+1)}} \right\|_\infty &\leq D^{(l+1)}\beta \end{aligned} \tag{20}$$

We derive the gradients of the objective function with respect to activations in layer l by chain rule:

$$\begin{aligned} \left\| \frac{\partial f}{\partial Z^{(l)}} \right\|_\infty &= \left\| \sigma'(Z^{(l)}) \circ \frac{\partial f}{\partial H^{(l)}} \right\|_\infty \\ &= \left\| \sigma'(Z^{(l)}) \circ A^T \ \frac{\partial f}{\partial Z^{(l+1)}} \ {W^{(l)}}^T \right\|_\infty \\ &\leq \eta^2 \left\| \sigma'(Z^{(l)}) \right\|_\infty \|A\|_\infty \left\| \frac{\partial f}{\partial Z^{(l+1)}} \right\|_\infty \left\| W^{(l)} \right\|_\infty \\ &\leq \eta^2 \beta^4 D^{(l+1)} \end{aligned} \tag{21}$$

Thus, we know that $\left\| \frac{\partial f}{\partial Z^{(l)}} \right\|_\infty \leq D^{(l)}\beta$. Similarly, $\left\| \frac{\partial f}{\partial Z_{SG}^{(l)}} \right\|_\infty \leq D^{(l)}\beta$, where $D^{(l)} = \eta^2\beta^3 D^{(l+1)}$.

We proceed to derive the error of the gradients by the SG estimator in layer l:

$$\begin{aligned} &\left\| \frac{\partial f}{\partial Z_{SG}^{(l)}} - \frac{\partial f}{\partial Z^{(l)}} \right\|_\infty \\ &\leq \eta \left\| W^{(l)} \right\|_\infty \| \sigma'(Z_{SG}^{(l)}) \circ A_{SG}^T \ \frac{\partial f}{\partial Z_{SG}^{(l+1)}} \\ &\quad - \sigma'(Z^{(l)}) \circ A^T \ \frac{\partial f}{\partial Z^{(l+1)}} \|_\infty \\ &\leq \underbrace{\eta\beta \left\| (\sigma'(Z_{SG}^{(l)}) - \sigma'(Z^{(l)})) \circ A_{SG}^T \frac{\partial f}{\partial Z_{SG}^{(l+1)}} \right\|_\infty}_{(*)} \\ &\quad + \underbrace{\eta\beta \left\| \sigma'(Z^{(l)}) \circ A_{SG}^T \left(\frac{\partial f}{\partial Z_{SG}^{(l+1)}} - \frac{\partial f}{\partial Z^{(l+1)}} \right) \right\|_\infty}_{(**)} \\ &\quad + \underbrace{\eta\beta \left\| \sigma'(Z^{(l)}) \circ (A_{SG}^T - A^T) \frac{\partial f}{\partial Z^{(l+1)}} \right\|_\infty}_{(***)} \end{aligned} \tag{22}$$

By Assumption 2 and Lemma 1, we know that there exists $\hat{C}$ such that $\left\| \sigma'(Z_{SG}^{(l)}) - \sigma'(Z^{(l)}) \right\|_\infty \leq \rho\hat{C}\epsilon$. From Equation 20, we have:

(*) in Eq. (22)

$$\begin{aligned} &\leq \eta^2\beta \left\| \sigma'(Z_{SG}^{(l)}) - \sigma'(Z^{(l)}) \right\|_\infty \|A_{SG}\|_\infty \left\| \frac{\partial f}{\partial Z_{SG}^{(l+1)}} \right\|_\infty \\ &\leq \eta^2\beta \cdot \rho\hat{C}\epsilon \cdot \beta \cdot D^{(l+1)}\beta \\ &= (\eta^2\beta^3\rho\hat{C}D^{(l+1)})\epsilon \end{aligned}$$

(**) in Eq. (22)

$$\begin{aligned} &\leq \eta^2\beta \left\| \sigma'(Z^{(l)}) \right\|_\infty \|A_{SG}\|_\infty \left\| \frac{\partial f}{\partial Z_{SG}^{(l+1)}} - \frac{\partial f}{\partial Z^{(l+1)}} \right\|_\infty \\ &\leq \eta^2\beta \cdot \beta \cdot C^{(l+1)}\epsilon \\ &= (\eta^2\beta^3 C^{(l+1)})\epsilon \end{aligned}$$

(***) in Eq. (22)

$$\begin{aligned} &\leq \eta^2\beta \left\| \sigma'(Z^{(l)}) \right\|_\infty \left\| A_{SG}^T - A^T \right\|_\infty \left\| \frac{\partial f}{\partial Z^{(l+1)}} \right\|_\infty \\ &\leq \eta^2\beta \cdot \beta \cdot \epsilon \cdot D^{(l+1)}\beta \\ &= (\eta^2\beta^3 D^{(l+1)})\epsilon \end{aligned} \tag{23}$$

Therefore, $\left\| \frac{\partial f}{\partial Z_{SG}^{(l)}} - \frac{\partial f}{\partial Z^{(l)}} \right\|_\infty \leq C^{(l)}\epsilon$, where $C^{(l)} = \eta^2\beta^3[(\rho\hat{C}+1)D^{(l+1)} + C^{(l+1)}]$. By induction, Equations 17 and 18 hold true.

Next, we show below that there exists C that depends on ρ, η and β, *s.t.*,

$$\left\| \frac{\partial f}{\partial W_{SG}^{(l)}} - \frac{\partial f}{\partial W^{(l)}} \right\|_\infty \leq C\epsilon, \quad \forall l \in [L] \tag{24}$$

By backpropagation rule we derive that $\frac{\partial f}{\partial W^{(l)}} = (AH^{(l)})^T \frac{\partial f}{\partial Z^{(l)}}$. By Lemma 1, we know that $H_{SG}^{(l)}$ is bounded by some $\hat{B}$ and $\left\| H_{SG}^{(l)} - H^{(l)} \right\|_\infty \leq \tilde{C}\epsilon$ hold for some $\tilde{C}$. From the previous proof, we know that there exists $\hat{C}$ and $\hat{D}$, *s.t.*, Equations 17 and 18

hold; thus, we have:

$$\begin{aligned}
&\left\|\frac{\partial f}{\partial W_{SG}^{(l)}}-\frac{\partial f}{\partial W^{(l)}}\right\|_\infty \\
&\le \left\|(A_{SG}H_{SG}^{(l)})^T\frac{\partial f}{\partial Z_{SG}^{(l+1)}}-(AH^{(l)})^T\frac{\partial f}{\partial Z^{(l+1)}}\right\|_\infty \\
&\le \left\|(A_{SG}H_{SG}^{(l)})^T(\frac{\partial f}{\partial Z_{SG}^{(l+1)}}-\frac{\partial f}{\partial Z^{(l+1)}})\right\|_\infty \\
&\quad+\left\|((A_{SG}H_{SG}^{(l)})^T-(AH^{(l)})^T)\frac{\partial f}{\partial Z^{(l+1)}}\right\|_\infty \\
&\le \eta^2\beta\cdot\hat{B}\cdot\hat{C}\epsilon+\eta\left\|A_{SG}H_{SG}^{(l)}-AH^{(l)}\right\|_\infty\hat{D}\beta \\
&\le \eta^2\beta\hat{B}\hat{C}\epsilon+\eta\beta\hat{D}\left\|A_{SG}(H_{SG}^{(l)}-H^{(l)})\right\|_\infty \\
&\quad+\eta\beta\hat{D}\left\|(A_{SG}-A)H^{(l)}\right\|_\infty \\
&\le \eta^2\beta\hat{B}\hat{C}\epsilon+\eta\beta\hat{D}\cdot\eta\beta\cdot\tilde{C}\epsilon+\eta\beta\hat{D}\cdot\eta\epsilon\cdot\hat{B} \\
&= \eta^2\beta(\hat{B}\hat{C}+\beta\tilde{C}\hat{D}+\hat{B}\hat{D})\epsilon
\end{aligned} \tag{25}$$

Therefore, Equation 24 holds, where $C=\eta^2\beta(\hat{B}\hat{C}+\beta\tilde{C}\hat{D}+\hat{B}\hat{D})$.

Finally, we have: $\|\nabla\mathcal{L}_{SG}(W)-\nabla\mathcal{L}(W)\|_\infty \le C\epsilon$, and the proof is complete.

D. Convergence Analysis

Theorem 1. *Assume that:*

1) *The loss function $\mathcal{L}(W)$ is ρ-smooth, i.e., $|\mathcal{L}(W_2)-\mathcal{L}(W_1)-\langle\mathcal{L}(W_1),W_2-W_1\rangle| \le \frac{\rho}{2}\|W_2-W_1\|_F^2, \forall W_1, W_2$, where $\langle A,B\rangle = tr(A^TB)$ denotes the inner product of matrix A and B,*
2) *The gradients of the loss $\nabla\mathcal{L}(W)$ and $\nabla\mathcal{L}_{SG}(W)$ are bounded by G for any choice of W,*
3) *The gradient of the objective function $\frac{\partial f(y,z)}{\partial z}$ is ρ-Lipschitz and bounded,*
4) *The activation function $\sigma(\cdot)$ is ρ-Lipschitz, $\sigma(0)=0$ and $\sigma'(\cdot)$ is bounded.*

Then there exists $C>0$, s.t., $\forall M,T$, for a sufficiently small δ, if we run graph partitioning for M times and run gradient descent for $R\le T$ epochs (where R is chosen uniformly from $[T]$, the model update rule is $W_{t+1}=W_t-\gamma\nabla\mathcal{L}_{SG}(W_t)$, step size $\gamma=\frac{1}{\rho\sqrt{T}}$), we have:

$$\begin{aligned}
&P(\mathbb{E}_R\|\nabla\mathcal{L}(W_R)\|_F^2\le\delta) \\
&\ge 1-2\exp\{-2M(\frac{\delta}{2C}-\frac{2\rho[\mathcal{L}(W_1)-\mathcal{L}(W_*)]+C-\delta}{2C(\sqrt{T}-1)})^2\}
\end{aligned}$$

Proof. Let $\delta_t=\nabla\mathcal{L}_{SG}(W_t)-\nabla\mathcal{L}(W_t)$ denote the differences between gradients at epoch t. By ρ-smoothness of $\mathcal{L}(W)$ we know that:

$$\begin{aligned}
&\mathcal{L}(W_{t+1}) \\
&\le \mathcal{L}(W_t)+\langle\nabla\mathcal{L}(W_t),W_{t+1}-W_t\rangle+\frac{\rho}{2}\gamma^2\|\nabla\mathcal{L}_{SG}(W_t)\|_F^2 \\
&= \mathcal{L}(W_t)-\gamma\langle\nabla\mathcal{L}(W_t),\nabla\mathcal{L}_{SG}(W_t)\rangle+\frac{\rho}{2}\gamma^2\|\nabla\mathcal{L}_{SG}(W_t)\|_F^2 \\
&= \mathcal{L}(W_t)-\gamma\langle\nabla\mathcal{L}(W_t),\delta_t\rangle-\gamma\|\nabla\mathcal{L}(W_t)\|_F^2 \\
&\quad+\frac{\rho}{2}\gamma^2[\|\delta_t\|_F^2+\|\nabla\mathcal{L}(W_t)\|_F^2+2\langle\delta_t,\nabla\mathcal{L}(W_t)\rangle] \\
&= \mathcal{L}(W_t)-(\gamma-\rho\gamma^2)\langle\nabla\mathcal{L}(W_t),\delta_t\rangle \\
&\quad-(\gamma-\frac{\rho}{2}\gamma^2)\|\nabla\mathcal{L}(W_t)\|_F^2+\frac{\rho}{2}\gamma^2\|\delta_t\|_F^2
\end{aligned} \tag{26}$$

By Lemma 2, we know that at a given time point t, there exists $\hat{C}$ *s.t.*, δ_t is bounded by $\hat{C}\epsilon$. Therefore,

$$\begin{aligned}
&|\langle\nabla\mathcal{L}(W_t),\delta_t\rangle|\le\eta\|\nabla\mathcal{L}(W_t)\|_\infty\|\delta_t\|_\infty\le\eta G\hat{C}\epsilon \\
&\|\delta_t\|_F^2\le\|\nabla\mathcal{L}_{SG}(W_t)\|_\infty^2+\|\nabla\mathcal{L}(W_t)\|_\infty^2\le 2G^2
\end{aligned} \tag{27}$$

Let $C=\max\{\eta G\hat{C},2G^2\}$. Equation 26 can be further derived as:

$$\begin{aligned}
\mathcal{L}(W_{t+1})\le\mathcal{L}(W_t)&+(\gamma-\rho\gamma^2)C\epsilon \\
&-(\gamma-\frac{\rho}{2}\gamma^2)\|\nabla\mathcal{L}(W_t)\|_F^2+\frac{\rho}{2}C\gamma^2
\end{aligned} \tag{28}$$

By summing up the above inequalities from $t=1$ to T and rearranging the terms, we have:

$$\begin{aligned}
(\gamma-\frac{\rho}{2}\gamma^2)\sum_t\|\nabla\mathcal{L}(W_t)\|_F^2&\le\mathcal{L}(W_1)-\mathcal{L}(W_*) \\
&+CT(\gamma-\rho\gamma^2)\epsilon+\frac{\rho}{2}CT\gamma^2
\end{aligned} \tag{29}$$

Dividing both sides of Equation 29 by $T(\gamma-\frac{\rho}{2}\gamma^2)$ and choosing $\gamma=\frac{1}{\rho\sqrt{T}}$ gives us:

$$\begin{aligned}
&\mathbb{E}_R\|\nabla\mathcal{L}(W_R)\|_F^2 \\
&\le 2\frac{\mathcal{L}(W_1)-\mathcal{L}(W_*)+CT(\gamma-\rho\gamma^2)\epsilon+\frac{\rho}{2}CT\gamma^2}{T\gamma(2-\rho\gamma)} \\
&\le \frac{2[\mathcal{L}(W_1)-\mathcal{L}(W_*)]}{T\gamma}+2C(1-\rho\gamma)\epsilon+\rho C\gamma \\
&\le \frac{2\rho[\mathcal{L}(W_1)-\mathcal{L}(W_*)]}{\sqrt{T}}+2C(1-\frac{1}{\sqrt{T}})\epsilon+\frac{C}{\sqrt{T}} \\
&\le \frac{2\rho[\mathcal{L}(W_1)-\mathcal{L}(W_*)]+C}{\sqrt{T}}+2C(1-\frac{1}{\sqrt{T}})\epsilon
\end{aligned} \tag{30}$$

Recall that ϵ denotes the infinity norm of the error in approximating A through M runs, *i.e.*, $\epsilon=$

$\|A_{SG} - A\|_\infty$. Applying Hoeffding's inequality [35] to the largest element of the matrix $|A_{SG} - A|$ (which are bounded by the intervals $[0, 1]$), we have:

$$P(\epsilon \geq \delta) \leq 2\exp(-2M\delta^2), \quad \forall \delta \geq 0 \tag{31}$$

Combining the two inequalities above, we have:

$$\begin{aligned} &P(\mathbb{E}_R\|\nabla\mathcal{L}(W_R)\|_F^2 \geq \delta) \\ &\leq P(\frac{2\rho[\mathcal{L}(W_1) - \mathcal{L}(W_*)] + C}{\sqrt{T}} + 2C(1 - \frac{1}{\sqrt{T}})\epsilon \geq \delta) \\ &\leq 2\exp\{-2M(\frac{\delta}{2C} - \frac{2\rho[\mathcal{L}(W_1) - \mathcal{L}(W_*)] + C - \delta}{2C(\sqrt{T} - 1)})^2\} \end{aligned} \tag{32}$$

Therefore, for a sufficiently small δ, we have the following inequality for $P(\mathbb{E}_R\|\nabla\mathcal{L}(W_R)\|_F^2 \leq \delta)$:

$$\begin{aligned} &P(\mathbb{E}_R\|\nabla\mathcal{L}(W_R)\|_F^2 \leq \delta) \\ &\quad \geq 1 - 2\exp\{-2M(\frac{\delta}{2C} - \frac{2\rho[\mathcal{L}(W_1) - \mathcal{L}(W_*)] + C - \delta}{2C(\sqrt{T} - 1)})^2\} \end{aligned} \tag{33}$$

Theorem 1 is proved.