

APPENDIX

A PRELIMINARIES

Standard notations. For a positive integer n , we define $[n] := \{1, 2, \dots, n\}$. For two integers $a \leq b$, we define $[a, b] := \{a, a+1, \dots, b\}$, and $(a, b) := \{a+1, \dots, b-1\}$. Similarly we define $[a, b)$ and $(a, b]$. For a full rank square matrix A , we use A^{-1} to denote its true inverse. We define the big O notation such that $f(n) = O(g(n))$ means there exists $n_0 \in \mathbb{N}_+$ and $M \in \mathbb{R}$ such that $f(n) \leq M \cdot g(n)$ for all $n \geq n_0$.

Norms. For a matrix A , we use $\|A\|$ or $\|A\|_2$ to denote its spectral norm. We use $\|A\|_F$ to denote its Frobenius norm. We use A^\top to denote the transpose of A . For a matrix A and a vector x , we define $\|x\|_A := \sqrt{x^\top A x}$.

Functions. We use ϕ to denote the ReLU activation function, i.e. $\phi(z) = \max\{z, 0\}$. For a function $f: \mathbb{R} \rightarrow \mathbb{R}$, we use f' to denote the derivative of f .

Fast matrix multiplication. We define the notation $\mathcal{T}_{\text{mat}}(n, d, m)$ to denote the time of multiplying an $n \times d$ matrix with another $d \times m$ matrix. Let ω denote the exponent of matrix multiplication, i.e., $\mathcal{T}_{\text{mat}}(n, n, n) = n^\omega$. The first result shows $\omega < 3$ is [Strassen \(1969\)](#). The current best $\omega \approx 2.373$ due to [Williams \(2012\)](#); [Le Gall \(2014\)](#). The following fact is well-known in the fast matrix multiplication literature [Coppersmith \(1982\)](#); [Strassen \(1991\)](#); [Bürgisser et al. \(1997\)](#) : $\mathcal{T}_{\text{mat}}(a, b, c) = O(\mathcal{T}_{\text{mat}}(a, c, b)) = O(\mathcal{T}_{\text{mat}}(c, a, b))$ for any positive integers a, b, c .

Kronecker product and vectorization. Given two matrices $A \in \mathbb{R}^{n_1 \times d_1}$ and $B \in \mathbb{R}^{n_2 \times d_2}$. We use \otimes to denote the Kronecker product, i.e., for $C = A \otimes B \in \mathbb{R}^{n_1 n_2 \times d_1 d_2}$, the $(i_1 + (i_2 - 1) \cdot n_1, j_1 + (j_2 - 1) \cdot d_1)$ -th entry of C is $A_{i_1, j_1} B_{i_2, j_2}$, $\forall i_1 \in [n_1], i_2 \in [n_2], j_1 \in [d_1], j_2 \in [d_2]$. For any give matrix $H \in \mathbb{R}^{d_1 \times d_2}$, we use $h = \text{vec}(H) \in \mathbb{R}^{d_1 d_2}$ to denote the vector such that $h_{j_1 + (j_2 - 1) \cdot d_1} = H_{j_1, j_2}$, $\forall j_1 \in [d_1], j_2 \in [d_2]$.

B GNTK FORMULAS

In this section we first present the GNTK formulas for GNNs of [Du et al. \(2019a\)](#), then we show our approximate version of the GNTK formulas.

B.1 GNNs

A GNN has L AGGREGATE operations, each followed by a COMBINE operation, and each COMBINE operation has R fully-connected layers. The fully-connected layers have output dimension m , and use ReLU as non-linearity. In the end the GNN has a READOUT operation that corresponds to the pooling operation of normal neural networks.

Let $G = (U, E)$ be a graph with $|U| = N$ number of nodes. Each node $u \in U$ has a feature vector $h_u \in \mathbb{R}^d$.

We define the initial vector $h_u^{(0, R)} = h_u \in \mathbb{R}^d$, $\forall u \in U$.

AGGREGATE operation. There are in total L AGGREGATE operations. For any $l \in [L]$, the AGGREGATE operation aggregates the information from last level as follows:

$$h_u^{(l, 0)} := c_u \cdot \sum_{a \in \mathcal{N}(u) \cup \{u\}} h_a^{(l-1, R)}.$$

Note that the vectors $h_u^{(l, 0)} \in \mathbb{R}^m$ for all $l \in [2 : L]$, and the only special case is $h_u^{(1, 0)} \in \mathbb{R}^d$. $c_u \in \mathbb{R}$ is a scaling parameter, which controls weight of different nodes during neighborhood aggregation. In our experiments we choose c_u between values $\{1, \frac{1}{|\mathcal{N}(u)|+1}\}$ following [Du et al. \(2019a\)](#).

COMBINE operation. The COMBINE operation has R fully-connected layers with ReLU activation: $\forall r \in [R]$,

$$h_u^{(l,r)} := (c_\phi/m)^{1/2} \cdot \phi(W^{(l,r)} \cdot h_u^{(l,r-1)}) \in \mathbb{R}^m.$$

The parameters $W^{(l,r)} \in \mathbb{R}^{m \times m}$ for all $(l,r) \in [L] \times [R] \setminus \{(1,1)\}$, and the only special case is $W^{(1,1)} \in \mathbb{R}^{m \times d}$. $c_\phi \in \mathbb{R}$ is a scaling parameter, in our experiments we set c_ϕ to be 2, following the initialization scheme used in [Du et al. \(2019a\)](#); [He et al. \(2015\)](#).

READOUT operation. Using the simplest READOUT operation, the final output of the GNN on graph G is

$$f_{\text{gnn}}(G) := \sum_{u \in U} h_u^{(L,R)} \in \mathbb{R}^m.$$

Using the READOUT operation with jumping knowledge as in [Xu et al. \(2018b\)](#), the final output of the GNN on graph G is

$$f_{\text{gnn}}(G) := \sum_{u \in U} [h_u^{(0,R)}, h_u^{(1,R)}, \dots, h_u^{(L,R)}] \in \mathbb{R}^{m \times (L+1)}.$$

B.2 EXACT GNTK FORMULAS

We first present the GNTK formulas of [Du et al. \(2019a\)](#).

We consider a GNN with L AGGREGATE operations and L COMBINE operations, and each COMBINE operation has R fully-connected layers. We use $h(W, G) \in \mathbb{R}^m$ to denote the function corresponding to this GNN, where $W := \cup_{\ell \in [L], r \in [R]} \{W^{(\ell,r)}\}$ denotes all tunable parameters of this GNN.

Let $G = (U, E)$ and $H = (V, F)$ be two graphs with $|U| = N$ and $|V| = N'$. We use $A_G \in \mathbb{R}^{N \times N}$ and $A_H \in \mathbb{R}^{N' \times N'}$ to denote the adjacency matrix of G and H . We give the recursive formula to compute the kernel value $K_{\text{gntk}}(G, H) \in \mathbb{R}$ induced by this GNN, which is defined as

$$K_{\text{gntk}}(G, H) := \mathbb{E}_{W \sim \mathcal{N}(0, I)} \left[\lim_{m \rightarrow \infty} \left\langle \frac{\partial f_{\text{gnn}}(W, G)}{\partial W}, \frac{\partial f_{\text{gnn}}(W, H)}{\partial W} \right\rangle \right].$$

Recall that the GNN uses scaling factors c_u for each node $u \in G$. We define $C_G \in \mathbb{R}^{N \times N}$ to be the diagonal matrix such that $(C_G)_u = c_u$ for any $u \in U$. Similarly we define $C_H \in \mathbb{R}^{N' \times N'}$.

We will use intermediate matrices $\Sigma^{(\ell,r)}(G, H) \in \mathbb{R}^{N \times N'}$ and $K^{(\ell,r)}(G, H) \in \mathbb{R}^{N \times N'}$ for each $\ell \in [0 : L]$ and $r \in [0 : R]$.

Initially we define $\Sigma^{(0,R)}(G, H) \in \mathbb{R}^{N \times N'}$ as follows: $\forall u \in U, v \in V$,

$$[\Sigma^{(0,R)}(G, H)]_{u,v} := \langle h_u, h_v \rangle,$$

where $h_u, h_v \in \mathbb{R}^d$ are the input features of u and v . And we define $K^{(0,R)}(G, H) \in \mathbb{R}^{N \times N'}$ as follows: $\forall u \in U, v \in V$,

$$[K^{(0,R)}(G, H)]_{u,v} := \langle h_u, h_v \rangle.$$

Next we recursively define $\Sigma^{(\ell,r)}(G, H)$ and $K^{(\ell,r)}(G, H)$ for $\ell \in [L]$ and $r \in [R]$, where ℓ denotes the level of AGGREGATE and COMBINE operations, and r denotes the level of fully-connected layers inside a COMBINE operation. Then we define the final output after READOUT operation.

AGGREGATE operation. The AGGREGATE operation gives the following formula:

$$\begin{aligned} [\Sigma^{(\ell,0)}(G, H)]_{u,v} &:= c_u c_v \sum_{a \in \mathcal{N}(u) \cup \{u\}} \sum_{b \in \mathcal{N}(v) \cup \{v\}} [\Sigma^{(\ell-1,R)}(G, H)]_{a,b}, \\ [K^{(\ell,0)}(G, H)]_{u,v} &:= c_u c_v \sum_{a \in \mathcal{N}(u) \cup \{u\}} \sum_{b \in \mathcal{N}(v) \cup \{v\}} [K^{(\ell-1,R)}(G, H)]_{a,b}. \end{aligned}$$

Note that the above two equations are equivalent to the following two equations:

$$\begin{aligned} \Sigma^{(\ell,0)}(G, H) &= C_G A_G \cdot \Sigma^{(\ell-1,R)}(G, H) \cdot A_H C_H, \\ K^{(\ell,0)}(G, H) &= C_G A_G \cdot K^{(\ell-1,R)}(G, H) \cdot A_H C_H. \end{aligned}$$

COMBINE operation. The COMBINE operation has R fully-connected layers with ReLU activation $\phi(z) = \max\{0, z\}$. We use $\dot{\phi}(z) = \mathbb{1}[z \geq 0]$ to denote be the derivative of ϕ .

For each $r \in [R]$, for each $u \in U$ and $v \in V$, we define a covariance matrix

$$[A^{(\ell,r)}(G, H)]_{u,v} := \begin{pmatrix} [\Sigma^{(\ell,r-1)}(G, G)]_{u,u} & [\Sigma^{(\ell,r-1)}(G, H)]_{u,v} \\ [\Sigma^{(\ell,r-1)}(H, G)]_{u,v} & [\Sigma^{(\ell,r-1)}(H, H)]_{v,v} \end{pmatrix} \in \mathbb{R}^{2 \times 2}.$$

Then we recursively define $[\Sigma^{(\ell,r)}(G, H)]_{u,v}$ and $[K^{(\ell,r)}(G, H)]_{u,v}$ as follows:

$$[\Sigma^{(\ell,r)}(G, H)]_{u,v} := c_\phi \cdot \mathbb{E}_{(a,b) \sim \mathcal{N}(0, [A^{(\ell,r)}(G, H)]_{u,v})} [\phi(a)\phi(b)], \quad (3)$$

$$[\dot{\Sigma}^{(\ell,r)}(G, H)]_{u,v} := c_\phi \cdot \mathbb{E}_{(a,b) \sim \mathcal{N}(0, [A^{(\ell,r)}(G, H)]_{u,v})} [\dot{\phi}(a)\dot{\phi}(b)], \quad (4)$$

$$[K^{(\ell,r)}(G, H)]_{u,v} := [K^{(\ell,r-1)}(G, H)]_{u,v} \cdot [\dot{\Sigma}^{(\ell,r)}(G, H)]_{u,v} + [\Sigma^{(\ell,r)}(G, H)]_{u,v}.$$

These intermediate outputs will be used to calculate the final output of the corresponding GNTK.

READOUT operation. Finally we compute $\mathbf{K}_{\text{gntk}}(G, H) \in \mathbb{R}$ using the intermediate matrices. This step corresponds to the READOUT operation.

If we do not use jumping knowledge,

$$\mathbf{K}_{\text{gntk}}(G, H) = \sum_{u \in U, v \in V} [K^{(L,R)}(G, H)]_{u,v}.$$

If we use jumping knowledge,

$$\mathbf{K}_{\text{gntk}}(G, H) = \sum_{u \in U, v \in V} \sum_{l=0}^L [K^{(l,R)}(G, H)]_{u,v}.$$

B.3 APPROXIMATE GNTK FORMULAS

We follow the notations of previous section. Again we consider two graphs $G = (U, E)$ and $H = (V, F)$ with $|U| = N$ and $|V| = N'$.

Now the goal is to compute an approximate version of the kernel value $\tilde{K}(G, H) \in \mathbb{R}$ such that

$$\tilde{\mathbf{K}}_{\text{gntk}}(G, H) \approx \mathbf{K}_{\text{gntk}}(G, H).$$

We will use intermediate matrices $\tilde{\Sigma}^{(\ell,r)}(G, H) \in \mathbb{R}^{N \times N'}$ and $\tilde{K}^{(\ell,r)}(G, H) \in \mathbb{R}^{N \times N'}$ for each $\ell \in [0 : L]$ and $r \in [0 : R]$. In the approximate version we add two random Gaussian matrices $S_G \in \mathbb{R}^{b \times N}$ and $S_H \in \mathbb{R}^{b' \times N'}$ where $b \leq N$ and $b' \leq N'$ are two parameters.

Initially, $\forall u \in U, v \in V$, we define $[\tilde{\Sigma}^{(0,R)}(G, H)]_{u,v} := \langle h_u, h_v \rangle$ and $[\tilde{K}^{(0,R)}(G, H)]_{u,v} := \langle h_u, h_v \rangle$, same as in the exact case.

Next we recursively define $\tilde{\Sigma}^{(\ell,r)}(G, H)$ and $\tilde{K}^{(\ell,r)}(G, H)$ for $\ell \in [L]$ and $r \in [R]$.

AGGREGATE operation. In the approximate version, we add two sketching matrices $S_G \in \mathbb{R}^{b \times N}$ and $S_H \in \mathbb{R}^{b' \times N'}$:

$$\begin{aligned}\tilde{\Sigma}^{(\ell,0)}(G, H) &:= C_G A_G \cdot (S_G^\top S_G) \cdot \tilde{\Sigma}^{(\ell-1,R)}(G, H) \cdot (S_H^\top S_H) \cdot A_H C_H, \\ \tilde{K}^{(\ell,0)}(G, H) &:= C_G A_G \cdot (S_G^\top S_G) \cdot \tilde{K}^{(\ell-1,R)}(G, H) \cdot (S_H^\top S_H) \cdot A_H C_H,\end{aligned}$$

where we define $C_G \in \mathbb{R}^{N \times N}$ to be the diagonal matrix such that $(C_G)_u = c_u$ for any $u \in V$. Similarly we define $C_H \in \mathbb{R}^{N' \times N'}$.

COMBINE operation. The COMBINE operation has R fully-connected layers with ReLU activation. The recursive definitions of $[\tilde{A}^{(\ell,r)}(G, H)]_{u,v} \in \mathbb{R}^{2 \times 2}$, $[\tilde{\Sigma}^{(\ell,r)}(G, H)]_{u,v} \in \mathbb{R}$, $[\tilde{\Sigma}^{(\ell,r)}(G, H)]_{u,v} \in \mathbb{R}$ and $[\tilde{K}^{(\ell,r)}(G, H)]_{u,v} \in \mathbb{R}$ are the same as in the exact case, except now we are always working with the tilde version.

READOUT operation. Finally we compute $\tilde{K}_{\text{gntk}}(G, H) \in \mathbb{R}$ using the intermediate matrices. It is also the same as in the exact case, except now we are always working with the tilde version.

C GENERALIZATION BOUND OF APPROXIMATE GNTK

In this section, we prove a generalization bound for the approximate version of GNTK. This generalizes the generalization bound of [Du et al. \(2019a\)](#).

Similar to [Du et al. \(2019a\)](#), we consider a GNN that has one AGGREGATE operation followed by one COMBINE operation with one fully-connected layer, and without jumping knowledge. We set the scaling parameters $c_u = (\|\sum_{v \in \mathcal{N}(u) \cup \{u\}} h_v\|_2)^{-1}$ and $c_\phi = 2$. We use $\{(G_i, y_i)\}_{i=1}^n$ to denote the n training data, where $G_i = (V_i, E_i)$ is a graph with $|V_i| = N_i$ nodes, and $y_i \in \mathbb{R}$ is its label.

We use $\tilde{K} \in \mathbb{R}^{n \times n}$ to denote approximate version of the kernel matrix of the simple GNN, where $\tilde{K}_{i,j} = \tilde{K}(G_i, G_j)$, as defined in Section 3.2. We assume \tilde{K} is invertible. For a testing graph G_{te} , the predicted label of kernel regression using approximate version of GNTK is

$$f_K(G_{\text{test}}) = [\tilde{K}(G_{\text{test}}, G_1), \tilde{K}(G_{\text{test}}, G_2), \dots, \tilde{K}(G_{\text{test}}, G_n)]^\top \tilde{K}^{-1} y.$$

Similar to [Du et al. \(2019a\)](#), we use the following standard generalization bound of kernel regression.

Theorem C.1 ([Bartlett & Mendelson \(2002\)](#)). *Given n training data $\{(G_i, y_i)\}_{i=1}^n$ drawn i.i.d. from the underlying distribution \mathcal{D} . Consider any loss function $\ell : \mathbb{R} \times \mathbb{R} \rightarrow [0, 1]$ that is 1-Lipschitz in the first argument such that $\ell(y, y) = 0$. With probability at least $1 - \delta$, the population loss of the GNTK predictor f_{gntk} can be upper bounded by*

$$L_{\mathcal{D}}(f_{\text{gntk}}) = \mathbb{E}_{(G, y) \sim \mathcal{D}} [\ell(f_{\text{gntk}}(G), y)] \lesssim (\|y\|_{\tilde{K}^{-1}}^2 \cdot \text{tr}[\tilde{K}])^{1/2} / n + \sqrt{\log(1/\delta)/n}.$$

Now it remains to bound $\|y\|_{\tilde{K}^{-1}}$ and $\text{tr}[\tilde{K}]$. We generalize the corresponding bounds of [Du et al. \(2019a\)](#) to the approximate GNTK. We have the following two lemmas:

Lemma C.2 (Informal version of Lemma C.8). *For each $i \in [n]$, if the labels $\{y_i\}_{i=1}^n$ satisfy*

$$y_i = \alpha_1 \sum_{u \in V} \langle \bar{h}_u, \beta_1 \rangle + \sum_{l=1}^T \alpha_{2l} \sum_{u \in V} \langle \bar{h}_u, \beta_{2l} \rangle^{2l}, \quad (5)$$

where $\bar{h}_u = c_u \sum_{v \in \mathcal{N}(u) \cup \{u\}} h_v$, $\alpha_1, \alpha_2, \alpha_4, \dots, \alpha_{2T} \in \mathbb{R}$, $\beta_1, \beta_2, \beta_4, \dots, \beta_{2T} \in \mathbb{R}^d$, then we have

$$\|y\|_{\tilde{K}^{-1}} \leq 4|\alpha_1| \cdot \|\beta_1\|_2 + \sum_{l=1}^T 4\sqrt{\pi(2l-1)} |\alpha_{2l}| \cdot \|\beta_{2l}\|_2^{2l}.$$

Lemma C.3 (Informal version of Lemma C.9). *If for all graphs $G_i = (V_i, E_i)$ in the training set, $|V_i|$ is upper bounded by N , then $\text{tr}[K] \leq O(nN^2)$. Here n is the number of training samples.*

Combining Lemma C.8 and Lemma C.9 with Theorem C.1, we have the following main generalization theorem:

Theorem C.4 (Main generalization theorem). *Following the notations of Definition C.5, and under the assumptions of Assumption C.6, if we further have the conditions that*

$$4 \cdot \alpha_1 \|\beta_1\|_2 + \sum_{l=1}^T 4\sqrt{\pi}(2l-1) \cdot \alpha_{2l} \|\beta_{2l}\|_2 = o(n), \quad N = o(\sqrt{n}),$$

then the generalization error of the approximate GNTK can be upper bounded by

$$L_{\mathcal{D}}(f_{\text{gntk}}) = \mathbb{E}_{(G,y) \sim \mathcal{D}}[\ell(f_{\text{gntk}}(G), y)] \lesssim O(1/n^c),$$

for some constant $c \in (0, 1)$.

Next we show how to prove the two lemmas needed for the main generalization theorem.

C.1 NOTATIONS AND ASSUMPTIONS

We first list all the notations and assumptions we used when proving the main generalization theorem.

Definition C.5 (Approximate GNTK with n data). *Let $\{(G_i, y_i)\}_{i=1}^n$ be the training data and labels, and $G_i = (V_i, E_i)$ with $|V_i| = N_i$, and we assume $N_i = O(N)$, $\forall i \in [n]$. For each $i \in [n]$ and each $u \in V_i$, let $h_u \in \mathbb{R}_+^d$ be the feature vector for u , and we define $H_{G_i} := [h_{u_1}, h_{u_2}, \dots, h_{u_{N_i}}] \in \mathbb{R}_+^{d \times N_i}$. We also define $A_{G_i} \in \mathbb{R}^{N_i \times N_i}$ to be the adjacency matrix of G_i , and define $S_{G_i} \in \mathbb{R}^{b_i \times N_i}$ to be the sketching matrix used for G_i .*

Let $\tilde{K} \in \mathbb{R}^{n \times n}$ be the approximate GNTK of a GNN that has one AGGREGATE operation followed by one COMBINE operation with one fully-connected layer ($L = 1$ and $R = 1$) and without jumping knowledge. For each $l \in [L]$, $r \in [R]$, $i, j \in [n]$, let $\tilde{\Sigma}^{(l,r)}(G_i, G_j), \tilde{K}^{(l,r)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$ be defined as of Section 3.2.

The scaling parameters used by the GNN for G_i are $c_\phi = 2$ and $c_u = (\| [H_{G_i} S_{G_i}^\top S_{G_i} A_{G_i}]_{,u} \|_2)^{-1}$, for each $u \in V_i$. We use $C_{G_i} \in \mathbb{R}^{N_i \times N_i}$ to denote the diagonal matrix with $[C_{G_i}]_{u,u} = c_u$.*

We further define two vectors for each $i \in [n]$ and each $u \in V_i$:

$$\bar{h}_u := [H_{G_i} A_{G_i} C_{G_i}]_{*,u} \in \mathbb{R}^d, \quad (6)$$

$$\tilde{h}_u := [H_{G_i} S_{G_i}^\top S_{G_i} A_{G_i} C_{G_i}]_{*,u} \in \mathbb{R}^d. \quad (7)$$

And let $T \in \mathbb{N}_+$ be a integer. For each $t \in \mathbb{N}_+$, we define two matrices $\bar{H}^{(t)}, \tilde{H}^{(t)} \in \mathbb{R}^{d \times n}$:

$$\bar{H}^{(t)} := \left[\sum_{u \in V_1} \Phi^{(t)}(\bar{h}_u), \sum_{u \in V_2} \Phi^{(t)}(\bar{h}_u), \dots, \sum_{u \in V_n} \Phi^{(t)}(\bar{h}_u) \right] \in \mathbb{R}^{d \times n}, \quad (8)$$

$$\tilde{H}^{(t)} := \left[\sum_{u \in V_1} \Phi^{(t)}(\tilde{h}_u), \sum_{u \in V_2} \Phi^{(t)}(\tilde{h}_u), \dots, \sum_{u \in V_n} \Phi^{(t)}(\tilde{h}_u) \right] \in \mathbb{R}^{d \times n}, \quad (9)$$

where we define $\Phi^{(t)}(\cdot)$ to be the feature map of the polynomial kernel of degree t such that $\langle x, y \rangle^t = \langle \Phi^{(t)}(x), \Phi^{(t)}(y) \rangle$ for any $x, y \in \mathbb{R}^d$.

Assumption C.6 (Assumptions to prove generalization bound). *We make the following assumptions about the input graphs, its feature vectors, and its labels.*

1. **Labels.** *For each $i \in [n]$, we assume the label $y_i \in \mathbb{R}$ satisfy*

$$y_i = \alpha_1 \sum_{u \in V_i} \langle \bar{h}_u, \beta_1 \rangle + \sum_{l=1}^T \alpha_{2l} \sum_{u \in V_i} \langle \bar{h}_u, \beta_{2l} \rangle^{2l},$$

where $\alpha_1, \alpha_2, \alpha_4, \dots, \alpha_{2T} \in \mathbb{R}$, $\beta_1, \beta_2, \beta_4, \dots, \beta_{2T} \in \mathbb{R}^d$.

2. **Feature vectors and graphs.** For each $t \in \{1\} \cup \{2l\}_{l=1}^T$, we assume we have

$$\|(\overline{H}^{(t)})^\top \overline{H}^{(t)}\|_F \leq \gamma_t \cdot \|(\overline{H}^{(t)})^\top \overline{H}^{(t)}\|_2,$$

where $\gamma_1, \gamma_2, \gamma_4, \dots, \gamma_{2T} \in \mathbb{R}$. We also let $\gamma = \max_{t \in \{1\} \cup \{2l\}_{l=1}^T} \{\gamma_t\}$. Note that $\gamma \geq 1$.

3. **Sketching sizes.** We assume the sketching sizes $\{b_i\}_{i=1}^n$ satisfy that $\forall i, j \in [n]$,

$$\begin{aligned} & \|A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}\|_2 \|A_{G_i} C_{G_i} \mathbf{1}_{[N_i]}\|_2 \cdot \|H_{G_j}^\top H_{G_i}\|_F \\ & \leq O(1) \cdot \frac{\min\{\sqrt{b_i}, \sqrt{b_j}\}}{\gamma T \log^3 N} \cdot \mathbf{1}_{[N_i]}^\top C_{G_i}^\top A_{G_i}^\top H_{G_i}^\top H_{G_j} A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}, \end{aligned}$$

where $\mathbf{1}_{[N_i]} \in \mathbb{R}^{N_i}$, $\mathbf{1}_{[N_j]} \in \mathbb{R}^{N_j}$ are the all one vectors of size N_i and N_j .

C.2 CLOSE-FORM FORMULA OF APPROXIMATE GNTK

Lemma C.7 (Close-form formula of approximate GNTK). *Following the notations of Definition C.5, we can decompose $\tilde{K} \in \mathbb{R}^{n \times n}$ into*

$$\tilde{K} = \tilde{K}_1 + \tilde{K}_2 \succeq \tilde{K}_1,$$

where $\tilde{K}_2 \in \mathbb{R}^{n \times n}$ is a PSD matrix, and $\tilde{K}_1 \in \mathbb{R}^{n \times n}$. \tilde{K}_1 satisfies the following:

$$\tilde{K}_1 = \frac{1}{4} (\tilde{H}^{(1)})^\top \cdot \tilde{H}^{(1)} + \frac{1}{2\pi} \sum_{l=1}^{\infty} \frac{(2l-3)!!}{(2l-2)!!(2l-1)} \cdot (\tilde{H}^{(2l)})^\top \cdot \tilde{H}^{(2l)},$$

and equivalently for each $i, j \in [n]$, $\tilde{K}_1(G_i, G_j) \in \mathbb{R}$ satisfies the following:

$$\tilde{K}_1(G_i, G_j) = \sum_{u \in V_i} \sum_{v \in V_j} \langle \tilde{h}_u, \tilde{h}_v \rangle \cdot \frac{1}{2\pi} (\pi - \arccos(\langle \tilde{h}_u, \tilde{h}_v \rangle)),$$

For each $i, j \in [n]$, $\tilde{K}_2(G_i, G_j) \in \mathbb{R}$ satisfies the following:

$$\tilde{K}_2(G_i, G_j) = \sum_{u \in V_i} \sum_{v \in V_j} \langle \tilde{h}_u, \tilde{h}_v \rangle \cdot \frac{1}{2\pi} \left(\pi - \arccos(\langle \tilde{h}_u, \tilde{h}_v \rangle) + \sqrt{1 - \langle \tilde{h}_u, \tilde{h}_v \rangle^2} \right).$$

Proof. For $i, j \in [n]$, consider the two graphs $G_i = (V_i, E_i)$ and $G_j = (V_j, E_j)$ with $|V_i| = N_i$ and $|V_j| = N_j$, we first compute the approximate GNTK formulas that corresponds to the simple GNN (with $L = 1$ and $R = 1$) by following the recursive formula of Section 3.2.

Compute $l = 0$, $r = 1$ variables. We first compute the initial variables $\tilde{\Sigma}^{(0,1)}(G_i, G_j), \tilde{K}^{(0,1)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$, for any $u \in V_i$ and $v \in V_j$, we have

$$[\tilde{\Sigma}^{(0,1)}(G_i, G_j)]_{u,v} = \langle h_u, h_v \rangle, \quad [\tilde{K}^{(0,1)}(G_i, G_j)]_{u,v} = \langle h_u, h_v \rangle, \quad (10)$$

which follows from Section 3.2.

Compute $l = 1$, $r = 0$ variables. We compute $\tilde{\Sigma}^{(1,0)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$ as follows:

$$\begin{aligned} [\tilde{\Sigma}^{(1,0)}(G_i, G_j)]_{u,v} &= [C_{G_i} A_{G_i} \cdot (S_{G_i}^\top S_{G_i}) \cdot \tilde{\Sigma}^{(0,1)}(G_i, G_j) \cdot (S_{G_j}^\top S_{G_j}) \cdot A_{G_j} C_{G_j}]_{u,v} \\ &= [C_{G_i} A_{G_i} \cdot (S_{G_i}^\top S_{G_i}) \cdot H_{G_i}^\top H_{G_j} \cdot (S_{G_j}^\top S_{G_j}) \cdot A_{G_j} C_{G_j}]_{u,v} \\ &= \langle \tilde{h}_u, \tilde{h}_v \rangle, \end{aligned} \quad (11)$$

where the first step follows from Section 3.2, the second step follows from Eq.(10) and the definition $H_{G_i} := [h_{u_1}, h_{u_2}, \dots, h_{u_{N_i}}] \in \mathbb{R}^{d \times N_i}$ in lemma statement, and the third step follows from the definition of $\tilde{h}_u, \tilde{h}_v \in \mathbb{R}^d$ in Eq. (7). Note that we have $\|\tilde{h}_u\|_2 = 1$ since C_{G_i} is a diagonal matrix with $[C_{G_i}]_{u,u} = \|[H_{G_i} S_{G_i}^\top S_{G_i} A_{G_i} C_{G_i}]_{*,u}\|_2^{-1}$.

Then we compute $\tilde{K}^{(1,0)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$:

$$\begin{aligned} [\tilde{K}^{(1,0)}(G_i, G_j)]_{u,v} &= [C_{G_i} A_{G_i} \cdot (S_{G_i}^\top S_{G_i}) \cdot \tilde{K}^{(0,1)}(G_i, G_j) \cdot (S_{G_j}^\top S_{G_j}) \cdot A_{G_j} C_{G_j}]_{u,v} \\ &= \tilde{\Sigma}^{(1,0)}(G_i, G_j) = \langle \tilde{h}_u, \tilde{h}_v \rangle, \end{aligned} \quad (12)$$

where the first step follows from Section 3.2, the second step follows from $\tilde{K}^{(0,1)}(G_i, G_j) = \tilde{\Sigma}^{(0,1)}(G_i, G_j)$ (see Eq. (10)).

Compute $l = 1, r = 1$ variables. Next we compute $[\tilde{A}^{(1,1)}(G_i, G_j)]_{u,v} \in \mathbb{R}^{2 \times 2}$ for any $u \in V_i$ and $v \in V_j$, we have

$$[\tilde{A}^{(1,1)}(G_i, G_j)]_{u,v} = \begin{pmatrix} [\tilde{\Sigma}^{(1,0)}(G_i, G_i)]_{u,u} & [\tilde{\Sigma}^{(1,0)}(G_i, G_j)]_{u,v} \\ [\tilde{\Sigma}^{(1,0)}(G_j, G_i)]_{v,u} & [\tilde{\Sigma}^{(1,0)}(G_j, G_j)]_{v,v} \end{pmatrix} = \begin{pmatrix} 1 & \langle \tilde{h}_u, \tilde{h}_v \rangle \\ \langle \tilde{h}_v, \tilde{h}_u \rangle & 1 \end{pmatrix}, \quad (13)$$

where the first step follows from Section 3.2, the second step follows from Eq. (11).

Then we compute $\tilde{\Sigma}^{(1,1)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$. For any $u \in V_i$ and $v \in V_j$,

$$\begin{aligned} [\tilde{\Sigma}^{(1,1)}(G_i, G_j)]_{u,v} &= c_\phi \cdot \mathbb{E}_{(a,b) \sim \mathcal{N}(0, [\tilde{A}^{(1,1)}(G_i, G_j)]_{u,v})} [\dot{\phi}(a) \cdot \dot{\phi}(b)] \\ &= \frac{1}{2\pi} \left(\pi - \arccos(\langle \tilde{h}_u, \tilde{h}_v \rangle) \right), \end{aligned} \quad (14)$$

where the first step follows from Section 3.2, the second step follows from Eq. (13) and $c_\phi = 2$ (Definition C.5).

And similarly we compute $[\tilde{\Sigma}^{(1,1)}(G_i, G_j)]_{u,v} \in \mathbb{R}^{N_i \times N_j}$. For any $u \in V_i$ and $v \in V_j$,

$$\begin{aligned} [\tilde{\Sigma}^{(1,1)}(G_i, G_j)]_{u,v} &= c_\phi \cdot \mathbb{E}_{(a,b) \sim \mathcal{N}(0, [\tilde{A}^{(1,1)}(G_i, G_j)]_{u,v})} [\phi(a) \cdot \phi(b)] \\ &= \frac{1}{2\pi} \left(\pi - \arccos(\langle \tilde{h}_u, \tilde{h}_v \rangle) + \sqrt{1 - \langle \tilde{h}_u, \tilde{h}_v \rangle^2} \right), \end{aligned} \quad (15)$$

where the first step follows from Section 3.2, the second step follows from Eq. (13) and $c_\phi = 2$ (Definition C.5).

Then we compute $\tilde{K}^{(1,1)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$. We decompose $\tilde{K}^{(1,1)}(G_i, G_j)$ as follows:

$$\tilde{K}^{(1,1)}(G_i, G_j) = \tilde{K}_1^{(1,1)}(G_i, G_j) + \tilde{\Sigma}^{(1,1)}(G_i, G_j), \quad (16)$$

where we define

$$[\tilde{K}_1^{(1,1)}(G_i, G_j)]_{u,v} := [\tilde{K}^{(1,0)}(G_i, G_j)]_{u,v} \cdot [\tilde{\Sigma}^{(1,1)}(G_i, G_j)]_{u,v}, \quad \forall u \in V_i, v \in V_j. \quad (17)$$

The above equation follows from the definition of $\tilde{K}^{(1,1)}(G_i, G_j)$ (see Section 3.2).

We then have

$$\begin{aligned} [\tilde{K}_1^{(1,1)}(G_i, G_j)]_{u,v} &= \langle \tilde{h}_u, \tilde{h}_v \rangle \cdot \frac{1}{2\pi} \left(\pi - \arccos(\langle \tilde{h}_u, \tilde{h}_v \rangle) \right) \\ &= \langle \tilde{h}_u, \tilde{h}_v \rangle \cdot \frac{1}{2\pi} \left(\pi - \left(\frac{\pi}{2} - \sum_{l=0}^{\infty} \frac{(2l-1)!!}{(2l)!!} \cdot \frac{\langle \tilde{h}_u, \tilde{h}_v \rangle^{2l+1}}{2l+1} \right) \right) \\ &= \frac{1}{4} \langle \tilde{h}_u, \tilde{h}_v \rangle + \frac{1}{2\pi} \sum_{l=1}^{\infty} \frac{(2l-3)!!}{(2l-2)!!(2l-1)} \cdot \langle \tilde{h}_u, \tilde{h}_v \rangle^{2l} \\ &= \frac{1}{4} \langle \tilde{h}_u, \tilde{h}_v \rangle + \frac{1}{2\pi} \sum_{l=1}^{\infty} \frac{(2l-3)!!}{(2l-2)!!(2l-1)} \cdot \langle \Phi^{(2l)}(\tilde{h}_u), \Phi^{(2l)}(\tilde{h}_v) \rangle, \end{aligned} \quad (18)$$

where the first step follows from plugging Eq. (12) and (14) into Eq. (17), the second step follows from the Taylor expansion that $\arccos(x) = \frac{\pi}{2} - \sum_{l=0}^{\infty} \frac{(2l-1)!!}{(2l)!!} \cdot \frac{x^{2l+1}}{2l+1}$, the third step

follows from merging terms, the fourth step follows from $\langle x, y \rangle^{2l} = \langle \Phi^{(2l)}(x), \Phi^{(2l)}(y) \rangle$ since $\Phi^{(2l)}(\cdot)$ is the feature map of the polynomial kernel of degree $2l$ (Definition C.5).

Compute kernel matrix \tilde{K} . Finally we compute $\tilde{K} \in \mathbb{R}^{n \times n}$. We decompose \tilde{K} as follows:

$$\tilde{K} = \tilde{K}_1 + \tilde{K}_2,$$

where we define $\tilde{K}_1, \tilde{K}_2 \in \mathbb{R}^{n \times n}$ such that for any two graphs $G_i = (V_i, E_i)$ and $G_j = (V_j, E_j)$,

$$\begin{aligned}\tilde{K}_1(G_i, G_j) &= \sum_{u \in V_i, v \in V_j} [\tilde{K}_1^{(1,1)}(G_i, G_j)]_{u,v} \in \mathbb{R}, \\ \tilde{K}_2(G_i, G_j) &= \sum_{u \in V_i, v \in V_j} [\tilde{\Sigma}^{(1,1)}(G_i, G_j)]_{u,v} \in \mathbb{R}.\end{aligned}$$

This equality follows from $\tilde{K}(G_i, G_j) = \sum_{u \in V_i, v \in V_j} [\tilde{K}^{(1,1)}(G_i, G_j)]_{u,v}$ (see Section 3.2) and Eq. (16). For \tilde{K}_1 , we further have

$$\begin{aligned}\tilde{K}_1(G_i, G_j) &= \frac{1}{4} \langle \sum_{u \in V_i} \tilde{h}_u, \sum_{v \in V_j} \tilde{h}_v \rangle + \frac{1}{2\pi} \sum_{l=1}^{\infty} \frac{(2l-3)!!}{(2l-2)!!(2l-1)} \cdot \langle \sum_{u \in V_i} \Phi^{(2l)}(\tilde{h}_u), \sum_{v \in V_j} \Phi^{(2l)}(\tilde{h}_v) \rangle \\ &= \frac{1}{4} (\tilde{H}^{(1)})^\top \cdot \tilde{H}^{(1)} + \frac{1}{2\pi} \sum_{l=1}^{\infty} \frac{(2l-3)!!}{(2l-2)!!(2l-1)} \cdot (\tilde{H}^{(2l)})^\top \cdot \tilde{H}^{(2l)},\end{aligned}$$

where the first step follows from Eq. (18), the second step follows from the definition of $\tilde{H}^{(t)} \in \mathbb{R}^{d \times n}$ (Eq. (9) in Definition C.5). For \tilde{K}_2 , we have that \tilde{K}_2 is a kernel matrix, so it is positive semi-definite. Let $A \succeq B$ denotes $x^\top A x \leq x^\top B x$ for all x . Thus we have $\tilde{K} \succeq \tilde{K}_1$. \square

C.3 BOUND ON $y^\top \tilde{K}^{-1} y$

Lemma C.8 (Bound on $y^\top \tilde{K}^{-1} y$, generalization of Theorem 4.2 of Du et al. (2019a)). *Following the notations of Definition C.5 and under the assumptions of Assumption C.6, we have*

$$\|y\|_{\tilde{K}^{-1}} \leq 4 \cdot |\alpha_1| \|\beta_1\|_2 + \sum_{l=1}^T 4\sqrt{\pi}(2l-1) \cdot |\alpha_{2l}| \|\beta_{2l}\|_2.$$

Proof. Decompose $\|y\|_{\tilde{K}^{-1}}$. From Part 1 of Assumption C.6, we have $\forall i \in [n]$

$$\begin{aligned}y_i &= \alpha_1 \sum_{u \in V_i} \langle \bar{h}_u, \beta_1 \rangle + \sum_{l=1}^T \alpha_{2l} \sum_{u \in V_i} \langle \bar{h}_u, \beta_{2l} \rangle^{2l} \\ &= \alpha_1 \sum_{u \in V_i} \langle \bar{h}_u, \beta_1 \rangle + \sum_{l=1}^T \alpha_{2l} \langle \sum_{u \in V_i} \Phi^{(2l)}(\bar{h}_u), \Phi^{(2l)}(\beta_{2l}) \rangle \\ &= y_i^{(1)} + \sum_{l=1}^T y_i^{(2l)},\end{aligned}\tag{19}$$

where the second step follows from $\Phi^{(2l)}$ is the feature map of the polynomial kernel of degree $2l$, i.e., $\langle x, y \rangle^{2l} = \langle \Phi^{(2l)}(x), \Phi^{(2l)}(y) \rangle$ for any $x, y \in \mathbb{R}^d$, and the third step follows from defining vectors $y^{(1)}, y^{(2l)} \in \mathbb{R}^n$ for $l \in [T]$ such that $\forall i \in [n]$,

$$y_i^{(1)} := \alpha_1 \langle \sum_{u \in V_i} \bar{h}_u, \beta_1 \rangle \in \mathbb{R},\tag{20}$$

$$y_i^{(2l)} := \alpha_{2l} \langle \sum_{u \in V_i} \Phi^{(2l)}(\bar{h}_u), \Phi^{(2l)}(\beta_{2l}) \rangle \in \mathbb{R}, \quad \forall l \in [T],\tag{21}$$

And we have

$$\|y\|_{\tilde{K}^{-1}} \leq \|y\|_{\tilde{K}_1^{-1}} \leq \|y^{(1)}\|_{\tilde{K}_1^{-1}} + \sum_{l=1}^T \|y^{(2l)}\|_{\tilde{K}_1^{-1}}, \quad (22)$$

which follows from Lemma C.7 that $\tilde{K} \succeq \tilde{K}_1$ and thus $\tilde{K}^{-1} \preceq \tilde{K}_1^{-1}$, the second step follows from Eq. (19) and triangle inequality.

Upper bound $\|y^{(1)}\|_{\tilde{K}_1^{-1}}$. Recall the definitions of the two matrices $\bar{H}^{(1)}, \tilde{H}^{(1)} \in \mathbb{R}^{d \times n}$ in Eq. (8) and (9) of Definition C.5:

$$\bar{H}^{(1)} := \left[\sum_{u \in V_1} \bar{h}_u, \sum_{u \in V_2} \bar{h}_u, \dots, \sum_{u \in V_n} \bar{h}_u \right], \quad \tilde{H}^{(1)} := \left[\sum_{u \in V_1} \tilde{h}_u, \sum_{u \in V_2} \tilde{h}_u, \dots, \sum_{u \in V_n} \tilde{h}_u \right].$$

Note that from Eq. (20) we have $y^{(1)} = \alpha_1 \cdot (\bar{H}^{(1)})^\top \beta_1 \in \mathbb{R}^n$. Also, from Lemma C.7 we have $\tilde{K}_1 \succeq \frac{1}{4}(\tilde{H}^{(1)})^\top \tilde{H}^{(1)} \in \mathbb{R}^{n \times n}$. Using these two equations, we have

$$\begin{aligned} \|y^{(1)}\|_{\tilde{K}_1^{-1}}^2 &= (y^{(1)})^\top \tilde{K}_1^{-1} y^{(1)} \\ &\leq 4\alpha_1^2 \cdot \beta_1^\top \bar{H}^{(1)} \cdot ((\tilde{H}^{(1)})^\top \tilde{H}^{(1)})^{-1} \cdot (\bar{H}^{(1)})^\top \beta_1. \end{aligned} \quad (23)$$

Next we want to prove that $(1 - \frac{1}{2})(\bar{H}^{(1)})^\top \bar{H}^{(1)} \preceq (\tilde{H}^{(1)})^\top \tilde{H}^{(1)} \preceq (1 + \frac{1}{2})(\bar{H}^{(1)})^\top \bar{H}^{(1)}$. For any $i, j \in [n]$, we have

$$\begin{aligned} [(\bar{H}^{(1)})^\top \bar{H}^{(1)}]_{i,j} &= \left(\sum_{u \in V_i} \bar{h}_u^\top \right) \cdot \left(\sum_{v \in V_j} \bar{h}_v \right) = \mathbf{1}_{[N_i]}^\top C_{G_i}^\top A_{G_i}^\top H_{G_i}^\top \cdot H_{G_j} A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}, \\ [(\tilde{H}^{(1)})^\top \tilde{H}^{(1)}]_{i,j} &= \left(\sum_{u \in V_i} \tilde{h}_u^\top \right) \cdot \left(\sum_{v \in V_j} \tilde{h}_v \right) = \mathbf{1}_{[N_i]}^\top C_{G_i}^\top A_{G_i}^\top (S_{G_i}^\top S_{G_i}) H_{G_i}^\top \cdot H_{G_j} (S_{G_j}^\top S_{G_j}) A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}, \end{aligned}$$

where $\mathbf{1}_{[N_i]} \in \mathbb{R}^{N_i}$ is the all one vector, and the second steps of the two equations follow from $\bar{h}_u = [H_{G_i} A_{G_i} C_{G_i}]_{*,u}$ and $\tilde{h}_u = [H_{G_i} S_{G_i}^\top S_{G_i} A_{G_i} C_{G_i}]_{*,u}$ (see Definition C.5).

For any $i, j \in [n]$, using Lemma 5.4 we have that with probability $1 - 1/N^4$,

$$\begin{aligned} &|[(\tilde{H}^{(1)})^\top \tilde{H}^{(1)}]_{i,j} - [(\bar{H}^{(1)})^\top \bar{H}^{(1)}]_{i,j}| \\ &\leq \frac{O(\log^{1.5} N)}{\sqrt{b_i}} \cdot \|A_{G_i} C_{G_i} \mathbf{1}_{[N_i]}\|_2 \cdot \|H_{G_i}^\top H_{G_j} A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}\|_2 \\ &\quad + \frac{O(\log^{1.5} N)}{\sqrt{b_j}} \cdot \|A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}\|_2 \cdot \|H_{G_j}^\top H_{G_i} A_{G_i} C_{G_i} \mathbf{1}_{[N_i]}\|_2 \\ &\quad + \frac{O(\log^3 N)}{\sqrt{b_i b_j}} \cdot \|A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}\|_2 \|A_{G_i} C_{G_i} \mathbf{1}_{[N_i]}\|_2 \cdot \|H_{G_j}^\top H_{G_i}\|_F \\ &\leq \frac{1}{10\gamma T} \cdot [(\bar{H}^{(1)})^\top \bar{H}^{(1)}]_{i,j}, \end{aligned} \quad (24)$$

where the last step follows from Part 3 of Assumption C.6 that

$$\begin{aligned} &\|A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}\|_2 \|A_{G_i} C_{G_i} \mathbf{1}_{[N_i]}\|_2 \cdot \|H_{G_j}^\top H_{G_i}\|_F \\ &\leq O(1) \cdot \frac{\min\{\sqrt{b_i}, \sqrt{b_j}\}}{\gamma T \log^3 N} \cdot \mathbf{1}_{[N_i]}^\top C_{G_i}^\top A_{G_i}^\top H_{G_i}^\top H_{G_j} A_{G_j} C_{G_j} \mathbf{1}_{[N_j]}. \end{aligned}$$

Then we have that

$$\begin{aligned} \|(\tilde{H}^{(1)})^\top \tilde{H}^{(1)} - (\bar{H}^{(1)})^\top \bar{H}^{(1)}\|_2^2 &\leq \|(\tilde{H}^{(1)})^\top \tilde{H}^{(1)} - (\bar{H}^{(1)})^\top \bar{H}^{(1)}\|_F^2 \\ &= \sum_{i=1}^n \sum_{j=1}^n |[(\tilde{H}^{(1)})^\top \tilde{H}^{(1)}]_{i,j} - [(\bar{H}^{(1)})^\top \bar{H}^{(1)}]_{i,j}|^2 \end{aligned}$$

$$\begin{aligned}
&\leq \frac{1}{100\gamma^2 T^2} \sum_{i=1}^n \sum_{j=1}^n [(\bar{H}^{(1)})^\top \bar{H}^{(1)}]_{i,j}^2 \\
&= \frac{1}{100\gamma^2 T^2} \|(\bar{H}^{(1)})^\top \bar{H}^{(1)}\|_F^2 \\
&\leq \frac{1}{100T^2} \|(\bar{H}^{(1)})^\top \bar{H}^{(1)}\|_2^2,
\end{aligned}$$

where the third step follows from Eq. (24), the fifth step follows from Part 2 of Assumption C.6 that $\|(\bar{H}^{(1)})^\top \bar{H}^{(1)}\|_F \leq \gamma \cdot \|(\bar{H}^{(1)})^\top \bar{H}^{(1)}\|_2$.

Thus we have proven that

$$(1 - \frac{1}{10})(\bar{H}^{(1)})^\top \bar{H}^{(1)} \preceq (\tilde{H}^{(1)})^\top \tilde{H}^{(1)} \preceq (1 + \frac{1}{10})(\bar{H}^{(1)})^\top \bar{H}^{(1)}.$$

Using this fact, we can bound $\|y^{(1)}\|_{\tilde{K}_1^{-1}}$ as follows:

$$\begin{aligned}
\|y^{(1)}\|_{\tilde{K}_1^{-1}} &\leq (4\alpha_1^2 \cdot \beta_1^\top \bar{H}^{(1)} \cdot ((\tilde{H}^{(1)})^\top \tilde{H}^{(1)})^{-1} \cdot (\bar{H}^{(1)})^\top \beta_1)^{1/2} \\
&\leq (8\alpha_1^2 \cdot \beta_1^\top \bar{H}^{(1)} \cdot ((\bar{H}^{(1)})^\top \bar{H}^{(1)})^{-1} \cdot (\bar{H}^{(1)})^\top \beta_1)^{1/2} \\
&\leq 4 \cdot \alpha_1 \|\beta_1\|_2.
\end{aligned} \tag{25}$$

where the first step follows from Eq. (23).

Upper bound $\|y^{(2l)}\|_{\tilde{K}_1^{-1}}$. Consider some $l \in [T]$. Recall the definitions of the two matrices $\bar{H}^{(2l)}, \tilde{H}^{(2l)} \in \mathbb{R}^{d \times n}$ in Eq. (8) and (9) of Definition C.5:

$$\begin{aligned}
\bar{H}^{(2l)} &:= \left[\sum_{u \in V_1} \Phi^{(2l)}(\bar{h}_u), \sum_{u \in V_2} \Phi^{(2l)}(\bar{h}_u), \dots, \sum_{u \in V_n} \Phi^{(2l)}(\bar{h}_u) \right], \\
\tilde{H}^{(2l)} &:= \left[\sum_{u \in V_1} \Phi^{(2l)}(\tilde{h}_u), \sum_{u \in V_2} \Phi^{(2l)}(\tilde{h}_u), \dots, \sum_{u \in V_n} \Phi^{(2l)}(\tilde{h}_u) \right].
\end{aligned}$$

Note that from Eq. (21) we have $y^{(2l)} = \alpha_{2l} \cdot (\bar{H}^{(2l)})^\top \Phi^{(2l)}(\beta_{2l}) \in \mathbb{R}^n$. Also, from Lemma C.7 we have $\tilde{K}_1 \succeq \frac{(2l-3)!!}{2\pi(2l-2)!!(2l-1)} (\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)} \in \mathbb{R}^{n \times n}$. Using these two equations, we have

$$\begin{aligned}
\|y^{(2l)}\|_{\tilde{K}_1^{-1}}^2 &= (y^{(2l)})^\top \tilde{K}_1^{-1} y^{(2l)} \\
&\leq \frac{2\pi(2l-2)!!(2l-1)}{(2l-3)!!} \alpha_{2l}^2 \cdot \beta_{2l}^\top \bar{H}^{(2l)} \cdot ((\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)})^{-1} \cdot (\bar{H}^{(2l)})^\top \beta_{2l}.
\end{aligned} \tag{26}$$

Next we want to prove that $(1 - \frac{1}{2})(\bar{H}^{(2l)})^\top \bar{H}^{(2l)} \preceq (\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)} \preceq (1 + \frac{1}{2})(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}$. For any $i, j \in [n]$, we have

$$[(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}]_{i,j} = \left(\sum_{u \in V_i} \Phi^{(2l)}(\bar{h}_u)^\top \right) \cdot \left(\sum_{v \in V_j} \Phi^{(2l)}(\bar{h}_v) \right) = \sum_{u \in V_i} \sum_{v \in V_j} (\bar{h}_u^\top \bar{h}_v)^{2l} \tag{27}$$

$$[(\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)}]_{i,j} = \left(\sum_{u \in V_i} \Phi^{(2l)}(\tilde{h}_u)^\top \right) \cdot \left(\sum_{v \in V_j} \Phi^{(2l)}(\tilde{h}_v) \right) = \sum_{u \in V_i} \sum_{v \in V_j} (\tilde{h}_u^\top \tilde{h}_v)^{2l} \tag{28}$$

where the second steps of both equations follow from $\Phi^{(2l)}(x)^\top \Phi^{(2l)}(y) = (x^\top y)^{2l}$.

Note that in Eq. (24) we have proven that

$$\left| \sum_{u \in V_i} \sum_{v \in V_j} (\bar{h}_u^\top \bar{h}_v - \tilde{h}_u^\top \tilde{h}_v) \right| \leq \frac{1}{10\gamma T} \sum_{u \in V_i} \sum_{v \in V_j} \bar{h}_u^\top \bar{h}_v.$$

Thus we have

$$\left| \sum_{u \in V_i} \sum_{v \in V_j} ((\bar{h}_u^\top \bar{h}_v)^{2l} - (\tilde{h}_u^\top \tilde{h}_v)^{2l}) \right| = \left| \sum_{u \in V_i} \sum_{v \in V_j} \sum_{i=0}^{2l-1} (\bar{h}_u^\top \bar{h}_v - \tilde{h}_u^\top \tilde{h}_v) (\bar{h}_u^\top \bar{h}_v)^i (\tilde{h}_u^\top \tilde{h}_v)^{2l-1-i} \right|$$

$$\begin{aligned}
&\leq 2l \cdot \left| \sum_{u \in V_i} \sum_{v \in V_j} (\bar{h}_u^\top \bar{h}_v - \tilde{h}_u^\top \tilde{h}_v) \right| \cdot (\max_{u,v} \{|\bar{h}_u^\top \bar{h}_v|, |\tilde{h}_u^\top \tilde{h}_v|\})^{2l-1} \\
&\leq 2l \cdot \frac{1}{10\gamma T} \cdot (1 + \frac{1}{10\gamma T})^{2l} \cdot (\sum_{u \in V_i} \sum_{v \in V_j} \bar{h}_u^\top \bar{h}_v)^{2l} \\
&\leq \frac{1}{2\gamma} (\sum_{u \in V_i} \sum_{v \in V_j} \bar{h}_u^\top \bar{h}_v)^{2l}, \tag{29}
\end{aligned}$$

where the first step follows from $x^{2l} - y^{2l} = \sum_{i=0}^{2l-1} (x-y)x^i y^{2l-1-i}$, the third step follows from $\max_{u,v} \{|\bar{h}_u^\top \bar{h}_v|\} \leq \sum_{u \in V_i} \sum_{v \in V_j} \bar{h}_u^\top \bar{h}_v$ since h_u and h_v are non-negative, and the previous inequality from Eq. (24), the fourth step follows from $l \leq T$ and $(1 + \frac{1}{10\gamma T})^{2l} \leq (1 + \frac{1}{10l})^{2l} \leq \sqrt{e}$.

Thus we have proven that for any $i, j \in [n]$, we have

$$\begin{aligned}
|[(\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)}]_{i,j} - [(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}]_{i,j}| &= \left| \sum_{u \in V_i} \sum_{v \in V_j} (\bar{h}_u^\top \bar{h}_v)^{2l} - \sum_{u \in V_i} \sum_{v \in V_j} (\tilde{h}_u^\top \tilde{h}_v)^{2l} \right| \\
&\leq \frac{1}{2\gamma} [(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}]_{i,j}. \tag{30}
\end{aligned}$$

where the first step follows from Eq. (27) and (28), the second step follows from Eq. (29).

Now similar to how we bound $\|y^{(1)}\|_{\tilde{K}^{-1}}$, we have that

$$\begin{aligned}
\|(\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)} - (\bar{H}^{(2l)})^\top \bar{H}^{(2l)}\|_2^2 &\leq \|(\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)} - (\bar{H}^{(2l)})^\top \bar{H}^{(2l)}\|_F^2 \\
&= \sum_{i=1}^n \sum_{j=1}^n |[(\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)}]_{i,j} - [(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}]_{i,j}|^2 \\
&\leq \frac{1}{4\gamma^2} \sum_{i=1}^n \sum_{j=1}^n [(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}]_{i,j}^2 \\
&= \frac{1}{4\gamma^2} \|(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}\|_F^2 \\
&\leq \frac{1}{4} \|(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}\|_2^2,
\end{aligned}$$

where the third step follows from Eq. (30), the fifth step follows from Part 2 of Assumption C.6 that $\|(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}\|_F \leq \gamma \cdot \|(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}\|_2$.

Thus we have proven that

$$(1 - \frac{1}{2})(\bar{H}^{(2l)})^\top \bar{H}^{(2l)} \preceq (\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)} \preceq (1 + \frac{1}{2})(\bar{H}^{(2l)})^\top \bar{H}^{(2l)}.$$

Using this fact, we can bound $\|y^{(2l)}\|_{\tilde{K}^{-1}}$ as follows:

$$\begin{aligned}
\|y^{(2l)}\|_{\tilde{K}^{-1}} &\leq \left(\frac{2\pi(2l-2)!!(2l-1)}{(2l-3)!!} \alpha_{2l}^2 \cdot \beta_{2l}^\top \bar{H}^{(2l)} \cdot ((\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)})^{-1} \cdot (\bar{H}^{(2l)})^\top \beta_{2l} \right)^{1/2} \\
&\leq \left(\frac{8\pi(2l-2)!!(2l-1)}{(2l-3)!!} \alpha_{2l}^2 \cdot \beta_{2l}^\top \bar{H}^{(2l)} \cdot ((\tilde{H}^{(2l)})^\top \tilde{H}^{(2l)})^{-1} \cdot (\bar{H}^{(2l)})^\top \beta_{2l} \right)^{1/2} \\
&\leq 4\sqrt{\pi}(2l-1) \cdot \alpha_{2l} \|\beta_{2l}\|_2, \tag{31}
\end{aligned}$$

where the first step follows from Eq. (26).

Upper bound $\|y\|_{\tilde{K}^{-1}}$. Plugging Eq. (25) and (31) into Eq. (22), we have

$$\|y\|_{\tilde{K}^{-1}} \leq 4 \cdot \alpha_1 \|\beta_1\|_2 + \sum_{l=1}^T 4\sqrt{\pi}(2l-1) \cdot \alpha_{2l} \|\beta_{2l}\|_2.$$

□

C.4 BOUND ON TRACE OF \tilde{K}

Lemma C.9 (Bound on trace of \tilde{K} , generalization of Theorem 4.3 of [Du et al. \(2019a\)](#)). *Following the notations of Definition C.5 and under the assumptions of Assumption C.6, we have*

$$\text{tr}[\tilde{K}] \leq 2nN^2.$$

Proof. From Lemma C.7 we can decompose $\tilde{K} \in \mathbb{R}^{n \times n}$ into

$$\tilde{K} = \tilde{K}_1 + \tilde{K}_2 \succeq \tilde{K}_1.$$

And for each $i \in [n]$, Lemma C.7 gives the following bound on the diagonal entries of \tilde{K}_1 :

$$\begin{aligned} \tilde{K}_1(G_i, G_i) &= \sum_{u \in V_i} \sum_{v \in V_i} \langle \tilde{h}_u, \tilde{h}_v \rangle \cdot \frac{1}{2\pi} (\pi - \arccos(\langle \tilde{h}_u, \tilde{h}_v \rangle)) \\ &\leq \sum_{u \in V_i} \sum_{v \in V_i} \frac{1}{2} = \frac{N_i^2}{2}, \end{aligned}$$

where the second step follows from \tilde{h}_u and \tilde{h}_v are unit vectors and that $\arccos(\cdot) \in [0, \pi]$.

Lemma C.7 also gives the following bound on the diagonal entries of \tilde{K}_2 :

$$\begin{aligned} \tilde{K}_2(G_i, G_i) &= \sum_{u \in V_i} \sum_{v \in V_i} \langle \tilde{h}_u, \tilde{h}_v \rangle \cdot \frac{1}{2\pi} \left(\pi - \arccos(\langle \tilde{h}_u, \tilde{h}_v \rangle) + \sqrt{1 - \langle \tilde{h}_u, \tilde{h}_v \rangle^2} \right) \\ &\leq \sum_{u \in V_i} \sum_{v \in V_i} \frac{1}{2\pi} (\pi + 1) \leq N_i^2. \end{aligned}$$

where the second step follows from \tilde{h}_u and \tilde{h}_v are unit vectors and that $\arccos(\cdot) \in [0, \pi]$.

Thus we have

$$\text{tr}[\tilde{K}] = \text{tr}[\tilde{K}_1] + \text{tr}[\tilde{K}_2] \leq 2 \sum_{i=1}^n N_i^2 \leq 2nN^2.$$

□

D RUNNING TIME

Theorem D.1 (Main running time theorem). *Consider a GNN with L levels of BLOCK operations, and R hidden layers in each level. We compute the kernel matrix using n graphs $G_1 = (V_1, E_1), \dots, G_n = (V_n, E_n)$ with $|V_i| = N_i$. Let $b_i \leq N_i$ be the sketch size of G_i . Let $d \in \mathbb{N}_+$ be the dimension of the feature vectors.*

The total running time to compute the approximate GNTK is

$$\sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, d, N_j) + O(L) \cdot \sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, N_j, b_i) + O(LR) \cdot \left(\sum_{i=1}^n N_i \right)^2.$$

When assuming $N_i \leq N$ and $b_i \leq b$ for all $i \in [n]$, the total running time is

$$O(n^2) \cdot (\mathcal{T}_{\text{mat}}(N, N, d) + L \cdot \mathcal{T}_{\text{mat}}(N, N, b) + LR \cdot N^2).$$

Proof. Preprocessing time. When preprocessing, we compute $A_{G_i} S_{G_i}^\top$ for each $i \in [n]$ in $\mathcal{T}_{\text{mat}}(N_i, N_i, b_i)$ time. So in total we need $\sum_{i=1}^n \mathcal{T}_{\text{mat}}(N_i, N_i, b_i)$ time.

We also compute the initial matrices $\tilde{\Sigma}^{(0,R)}(G_i, G_j), \tilde{K}^{(0,R)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$ for each $i, j \in [n]$ with $[\tilde{\Sigma}^{(0,R)}(G_i, G_j)]_{u,v} = [K^{(0,R)}(G_i, G_j)]_{u,v} = \langle h_u, h_v \rangle, \forall u \in V_i, v \in V_j$. Computing

each $\tilde{\Sigma}^{(0,R)}(G_i, G_j)$ corresponds to multiplying the concatenation of the feature vectors of G_i with that of G_j , which is multiplying a $N_i \times d$ matrix with a $d \times N_j$ matrix, and this takes $\mathcal{T}_{\text{mat}}(N_i, d, N_j)$ time. So computing all initial matrices takes $\sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, d, N_j)$ time.

Thus the total preprocessing time is

$$\sum_{i=1}^n \mathcal{T}_{\text{mat}}(N_i, N_i, b_i) + \sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, d, N_j).$$

BLOCK operation: aggregation time. In the l -th level of BLOCK operation, $\forall i, j \in [n]$ we compute $\tilde{\Sigma}^{(l,0)}(G_i, G_j), \tilde{K}^{(l,0)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$ by computing

$$\begin{aligned} \tilde{\Sigma}^{(\ell,0)}(G_i, G_j) &:= C_{G_i}(A_{G_i} S_{G_i}^\top) \cdot S_{G_i} \cdot \tilde{\Sigma}^{(\ell-1,R)}(G_i, G_j) \cdot S_{G_j}^\top \cdot (S_{G_j} A_{G_j}) C_{G_j}, \\ \tilde{K}^{(\ell,0)}(G, H) &:= C_{G_i}(A_{G_i} S_{G_i}^\top) \cdot S_{G_i} \cdot \tilde{K}^{(\ell-1,R)}(G_i, G_j) \cdot S_{G_j}^\top \cdot (S_{G_j} A_{G_j}) C_{G_j}. \end{aligned}$$

This takes $O(\mathcal{T}_{\text{mat}}(N_i, N_j, b_i) + \mathcal{T}_{\text{mat}}(N_i, N_j, b_j))$ time.

Thus the total time of all aggregation operations of L levels is

$$O(L) \cdot \sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, N_j, b_i).$$

BLOCK operation: hidden layer time. In the l -th level of BLOCK operation, in the r -th hidden layer, for each $i, j \in [n]$ we compute $\tilde{\Sigma}^{(l,r)}(G_i, G_j), \tilde{K}^{(l,r)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$ by computing each entry $[\tilde{\Sigma}^{(l,r)}(G_i, G_j)]_{u,v}, [\tilde{K}^{(l,r)}(G_i, G_j)]_{u,v} \in \mathbb{R}$ for $u \in V_i, v \in V_j$. Computing each entry takes $O(1)$ time, which follows trivially from their definitions (see Section 3.1 and 3.2). Thus the total time of all R hidden layers operations of L levels is

$$O(LR) \cdot \left(\sum_{i=1}^n N_i \right)^2.$$

READOUT operation time. Finally we compute kernel matrix $K \in \mathbb{R}^{n \times n}$ such that for $i, j \in [n]$, $K(G_i, G_j) \in \mathbb{R}$ is computed as

$$K(G_i, G_j) = \sum_{u \in V_i, v \in V_j} [K^{(L,R)}(G_i, G_j)]_{u,v}.$$

Thus the total time of READOUT operation is

$$\left(\sum_{i=1}^n N_i \right)^2.$$

Total time. Thus the total running time to compute the approximate GNTK is

$$\sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, d, N_j) + O(L) \cdot \sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, N_j, b_i) + O(LR) \cdot \left(\sum_{i=1}^n N_i \right)^2.$$

When assuming $N_i \leq N$ and $b_i \leq b$ for all $i \in [n]$, the total running time is

$$O(n^2) \cdot (\mathcal{T}_{\text{mat}}(N, N, d) + L \cdot \mathcal{T}_{\text{mat}}(N, N, b) + LR \cdot N^2).$$

□

For comparison we state the running time of computing GNTK of [Du et al. \(2019a\)](#).

Theorem D.2 (Running time of Du et al. (2019a)). *Consider a GNN with L levels of BLOCK operations, and R hidden layers in each level. We compute the kernel matrix using n graphs $G_1 = (V_1, E_1), \dots, G_n = (V_n, E_n)$ with $|V_i| = N_i$. Let $d \in \mathbb{N}_+$ be the dimension of the feature vectors.*

The total running time of Du et al. (2019a) to compute the GNTK is

$$\sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, d, N_j) + O(L) \cdot \left(\sum_{i=1}^n N_i^2 \right)^2 + O(LR) \cdot \left(\sum_{i=1}^n N_i \right)^2.$$

When assuming $N_i \leq N$ and $b_i \leq b$ for all $i \in [n]$, the total running time is

$$O(n^2) \cdot (\mathcal{T}_{\text{mat}}(N, N, d) + L \cdot N^4 + LR \cdot N^2).$$

We include a proof here for completeness.

Proof. Comparing with Theorem D.1, the only different part of the running time is the aggregation time of BLOCK operation. For the other three parts, see the proof of Theorem D.1.

BLOCK operation: aggregation time. In the l -th level of BLOCK operation, $\forall i, j \in [n]$ we compute $\Sigma^{(l,0)}(G_i, G_j), K^{(l,0)}(G_i, G_j) \in \mathbb{R}^{N_i \times N_j}$ by computing

$$\begin{aligned} \text{vec}(\Sigma^{(\ell,0)}(G_i, G_j)) &:= ((C_{G_i} A_{G_i}) \otimes (C_{G_j} A_{G_j})) \cdot \text{vec}(\Sigma^{(\ell-1,R)}(G_i, G_j)) \in \mathbb{R}^{N_i N_j}, \\ \text{vec}(K^{(\ell,0)}(G_i, G_j)) &:= ((C_{G_i} A_{G_i}) \otimes (C_{G_j} A_{G_j})) \cdot \text{vec}(K^{(\ell-1,R)}(G_i, G_j)) \in \mathbb{R}^{N_i N_j}. \end{aligned}$$

Note that the sizes are $((C_{G_i} A_{G_i}) \otimes (C_{G_j} A_{G_j})) \in \mathbb{R}^{N_i N_j \times N_i N_j}$, $\text{vec}(\Sigma^{(\ell-1,R)}(G_i, G_j)) \in \mathbb{R}^{N_i N_j}$. So this takes $O(N_i^2 N_j^2)$ time, even to simply compute $((C_{G_i} A_{G_i}) \otimes (C_{G_j} A_{G_j}))$.

Thus the total time of all aggregation operations of L levels is

$$O(L) \cdot \left(\sum_{i=1}^n N_i^2 \right)^2.$$

Total time. Thus the total running time in Du et al. (2019a) to compute the exact GNTK is

$$\sum_{i=1}^n \sum_{j=1}^n \mathcal{T}_{\text{mat}}(N_i, d, N_j) + O(L) \cdot \left(\sum_{i=1}^n N_i^2 \right)^2 + O(LR) \cdot \left(\sum_{i=1}^n N_i \right)^2.$$

When assuming $N_i \leq N$ and $b_i \leq b$ for all $i \in [n]$, the total running time is

$$O(n^2) \cdot (\mathcal{T}_{\text{mat}}(N, N, d) + L \cdot N^4 + LR \cdot N^2).$$

□

E MISSING PROOFS FOR KRONECKER PRODUCT AND SKETCHING

E.1 PROOFS OF KRONECKER PRODUCT EQUIVALENCE

Fact E.1 (Equivalence between two matrix products and Kronecker product then matrix-vector multiplication). *Given matrices $A \in \mathbb{R}^{n_1 \times d_1}$, $B \in \mathbb{R}^{n_2 \times d_2}$, and $H \in \mathbb{R}^{d_1 \times d_2}$, we have $\text{vec}(AHB^\top) = (A \otimes B) \cdot \text{vec}(H)$.*

Proof. First note that $AHB^\top \in \mathbb{R}^{n_1 \times n_2}$, $A \otimes B \in \mathbb{R}^{n_1 n_2 \times d_1 d_2}$, and $(A \otimes B) \cdot \text{vec}(H) \in \mathbb{R}^{n_1 n_2}$.

For any $i_1 \in [n_1]$, $i_2 \in [n_2]$, define $i := i_1 + (i_2 - 1) \cdot n_1$, we have

$$\text{vec}(AHB^\top)_i = (AHB^\top)_{i_1, i_2}$$

$$= \sum_{j_1 \in [d_1]} \sum_{j_2 \in [d_2]} A_{i_1, j_1} \cdot H_{j_1, j_2} \cdot B_{i_2, j_2},$$

and we also have,

$$\begin{aligned} ((A \otimes B) \cdot \text{vec}(H))_i &= \sum_{\substack{j := j_1 + (j_2 - 1) \cdot d_1, \\ j_1 \in [d_1], j_2 \in [d_2]}} (A \otimes B)_{i, j} \cdot \text{vec}(H)_j \\ &= \sum_{j_1 \in [d_1], j_2 \in [d_2]} A_{i_1, j_1} B_{i_2, j_2} \cdot H_{j_1, j_2}. \end{aligned}$$

Thus we have $\text{vec}(AHB^\top) = (A \otimes B) \cdot \text{vec}(H)$. \square

E.2 PROOF OF SKETCHING BOUND

We will use the following inequality.

Fact E.2 (Khinchine's inequality). *Let $\sigma_1, \sigma_2, \dots, \sigma_n$ be i.i.d. sign random variables, and let $z_1, z_2, \dots, z_n \in \mathbb{R}$. Then there exist constants $C, C' > 0$ such that $\forall t \in \mathbb{R}_+$,*

$$\Pr \left[\left| \sum_{i=1}^n \sigma_i z_i \right| \geq Ct \|z\|_2 \right] \leq e^{-C't^2}.$$

Lemma E.3 (Restatement of Lemma 5.4). *Let $A \in \mathbb{R}^{n \times n}$ be a matrix. Let $R \in \mathbb{R}^{b_1 \times n}$ and $S \in \mathbb{R}^{b_2 \times n}$ be two independent AMS matrices. Let $g, h \in \mathbb{R}^n$ be two vectors. Then with probability at least $1 - \text{poly}(1/n)$, we have*

$$\begin{aligned} &g^\top (R^\top R) A (S^\top S) h - g^\top A h \\ &\leq O\left(\frac{\log^{1.5} n}{\sqrt{b_1}}\right) \|g\|_2 \|Ah\|_2 + O\left(\frac{\log^{1.5} n}{\sqrt{b_2}}\right) \|g^\top A\|_2 \|h\|_2 + O\left(\frac{\log^3 n}{\sqrt{b_1 b_2}}\right) \cdot \|g\|_2 \|h\|_2 \|A\|_F. \end{aligned}$$

Proof. For $i \in [n]$, we use $R_i \in \mathbb{R}^{b_1}$ and $S_i \in \mathbb{R}^{b_2}$ to denote the i -th column of R and S .

Each column R_i of the AMS matrix R has the same distribution as $\sigma_i R_i$, where σ_i is a random sign. The AMS matrix R has the following properties:

$$1. \langle R_i, R_i \rangle = 1, \forall i \in [n]. \quad (32)$$

$$2. \Pr \left[\langle R_i, R_j \rangle \leq \frac{\sqrt{\log(n/\delta)}}{\sqrt{b_1}}, \forall i \neq j \in [n] \right] \geq 1 - \delta. \quad (33)$$

Similarly each column S_i of AMS matrix S has the same distribution as $\sigma'_i S_i$, where σ'_i is a random sign. For more details see Alon et al. (1999).

We have

$$g^\top (R^\top R) A (S^\top S) h = \sum_{i, j, i', j'} g_i h_{j'} \sigma_i \sigma_j \sigma'_{i'} \sigma'_{j'} \langle R_i, R_j \rangle A_{j, i'} \langle S_{i'}, S_{j'} \rangle. \quad (34)$$

Thus we can split the summation of Eq. (34) into three parts: 1. Two pairs of indexes are the same: $i = j$ and $i' = j'$; 2. One pair of indexes are the same: $i = j$ and $i' \neq j'$, or symmetrically $i \neq j$ and $i' = j'$; 3. No pair of indexes are the same: $i \neq j$ and $i' \neq j'$.

Part 1. Two pairs of indexes are the same. We consider the case where $i = j$ and $i' = j'$. We have

$$\sum_{i=j, i'=j'} g_i h_{j'} \sigma_i \sigma_j \sigma'_{i'} \sigma'_{j'} \langle R_i, R_j \rangle A_{j, i'} \langle S_{i'}, S_{j'} \rangle = \sum_{i, i'} g_i h_{i'} A_{i, i'} = g^\top A h, \quad (35)$$

where the first step follows from $\langle R_i, R_i \rangle = \langle S_{i'}, S_{i'} \rangle = 1, \forall i, i' \in [n]$, see Eq. (32).

Part 2. One pair of indexes are the same. We consider the case where $i = j$ and $i' \neq j'$, or the symmetric case where $i \neq j$ and $i' = j'$.

W.l.o.g. we consider the case that $i = j$ and $i' \neq j'$. We have

$$\begin{aligned} \sum_{i=j, i' \neq j'} g_i h_{j'} \sigma_i \sigma_j \sigma'_{i'} \sigma'_{j'} \langle R_i, R_j \rangle A_{j, i'} \langle S_{i'}, S_{j'} \rangle &= \sum_{i, i' \neq j'} g_i h_{j'} \sigma'_{i'} \sigma'_{j'} A_{i, i'} \langle S_{i'}, S_{j'} \rangle \\ &= \sum_{j'} \sigma'_{j'} h_{j'} \sum_{i' \neq j'} \sigma'_{i'} (A^\top g)_{i'} \langle S_{i'}, S_{j'} \rangle, \end{aligned}$$

where the first step follows from $\langle R_i, R_i \rangle = 1, \forall i \in [n]$ (Eq. (32)), the second step follows from $\sum_i g_i A_{i, i'} = (A^\top g)_{i'}$.

Using Khintchine's inequality (Fact E.2) and Union bound, we have that with probability at least $1 - \text{poly}(1/n)$,

$$\begin{aligned} \left(\sum_{j'} \sigma'_{j'} h_{j'} \sum_{i' \neq j'} \sigma'_{i'} (A^\top g)_{i'} \langle S_{i'}, S_{j'} \rangle \right)^2 &\leq O(\log n) \sum_{j'} h_{j'}^2 \left(\sum_{i' \neq j'} \sigma'_{i'} (A^\top g)_{i'} \langle S_{i'}, S_{j'} \rangle \right)^2 \\ &\leq O(\log^2 n) \sum_{j'} h_{j'}^2 \sum_{i' \neq j'} (A^\top g)_{i'}^2 \langle S_{i'}, S_{j'} \rangle^2 \\ &\leq O((\log^3 n)/b_2) \sum_{j'} h_{j'}^2 \sum_{i' \neq j'} (A^\top g)_{i'}^2 \\ &\leq O((\log^3 n)/b_2) \|h\|_2^2 \|A^\top g\|_2^2, \end{aligned}$$

where the first step follows from applying Khintchine's inequality with $t = O(\sqrt{\log n})$, the second step again follows from applying Khintchine's inequality with $t = O(\sqrt{\log n})$, the third step follows from that with probability at least $1 - \text{poly}(1/n)$, $\langle S_i, S_j \rangle \leq O(\sqrt{(\log n)/b_2})$ for all $i \neq j \in [n]$, see Eq. (33).

Plugging this equation into the previous equation, and note that the case that $i' = j', i \neq j$ is symmetric, we have that with probability at least $1 - \text{poly}(1/n)$,

$$\sum_{\substack{i=j, i' \neq j' \\ \text{or } i'=j', i \neq j}} g_i h_{j'} \sigma_i \sigma_j \sigma'_{i'} \sigma'_{j'} \langle R_i, R_j \rangle A_{j, i'} \langle S_{i'}, S_{j'} \rangle \quad (36)$$

$$\leq O(\log^{1.5} n / \sqrt{b_1}) \|g\|_2 \|Ah\|_2 + O(\log^{1.5} n / \sqrt{b_2}) \|g^\top A\|_2 \|h\|_2. \quad (37)$$

Part 3. No pair of indexes are the same. We consider the case where $i \neq j$ and $i' \neq j'$. We prove it by using Khintchine's inequality (Fact E.2) four times. We have that with probability $1 - \text{poly}(1/n)$,

$$\begin{aligned} &\left(\sum_{i \neq j, i' \neq j'} g_i h_{j'} \sigma_i \sigma_j \sigma'_{i'} \sigma'_{j'} \langle R_i, R_j \rangle A_{j, i'} \langle S_{i'}, S_{j'} \rangle \right)^2 \\ &= \left(\sum_i \sigma_i g_i \sum_{j'} \sigma'_{j'} h_{j'} \sum_{i' \neq j'} \sigma'_{i'} \langle S_{i'}, S_{j'} \rangle \sum_{j \neq i} \sigma_i \langle R_i, R_j \rangle A_{j, i'} \right)^2 \\ &\leq O(\log n) \sum_i g_i^2 \left(\sum_{j'} \sigma'_{j'} h_{j'} \sum_{i' \neq j'} \sigma'_{i'} \langle S_{i'}, S_{j'} \rangle \sum_{j \neq i} \sigma_i \langle R_i, R_j \rangle A_{j, i'} \right)^2 \\ &\leq O(\log^2 n) \sum_i g_i^2 \sum_{j'} h_{j'}^2 \left(\sum_{i' \neq j'} \sigma'_{i'} \langle S_{i'}, S_{j'} \rangle \sum_{j \neq i} \sigma_i \langle R_i, R_j \rangle A_{j, i'} \right)^2 \\ &\leq O(\log^3 n) \sum_i g_i^2 \sum_{j'} h_{j'}^2 \sum_{i' \neq j'} \langle S_{i'}, S_{j'} \rangle^2 \left(\sum_{j \neq i} \sigma_i \langle R_i, R_j \rangle A_{j, i'} \right)^2 \\ &\leq O(\log^4 n) \sum_i g_i^2 \sum_{j'} h_{j'}^2 \sum_{i' \neq j'} \langle S_{i'}, S_{j'} \rangle^2 \sum_{j \neq i} \langle R_i, R_j \rangle^2 A_{j, i'}^2 \\ &\leq O((\log^6 n)/(b_1 b_2)) \|g\|_2^2 \|h\|_2^2 \|A\|_F^2, \end{aligned}$$

where the second step follows from Khintchine's inequality with $t = O(\sqrt{n})$, the third step follows from Khintchine's inequality with $t = O(\sqrt{n})$ for each $i \in [n]$, and combining the n

inequalities using Union bound, the fourth step and the fifth step follows from same reason as the third step, the sixth step follows from that with probability at least $1 - \text{poly}(1/n)$, $\langle S_{i'}, S_{j'} \rangle \leq O(\sqrt{(\log n)/b_2})$ for all $i' \neq j' \in [n]$, and similarly with probability at least $1 - \text{poly}(1/n)$, $\langle R_i, R_j \rangle \leq O(\sqrt{(\log n)/b_1})$ for all $i \neq j \in [n]$, we combine the $2n^2$ such bounds all $i, j, i', j' \in [n]$ using Union bound.

Thus we have that with probability at least $1 - \text{poly}(1/n)$,

$$\sum_{i \neq j, i' \neq j'} g_i h_{j'} \sigma_i \sigma_j \sigma_{i'} \sigma_{j'} \langle R_i, R_j \rangle A_{j, i'} \langle S_{i'}, S_{j'} \rangle \leq O((\log^3 n) / \sqrt{b_1 b_2}) \cdot \|g\|_2 \|h\|_2 \|A\|_F. \quad (38)$$

Combining all parts together. Adding Eq. (35), (36), (38) together and plugging into Eq. (34), using Union bound, we have that with probability at least $1 - \text{poly}(1/n)$,

$$\begin{aligned} & g^\top (R^\top R) A (S^\top S) h - g^\top A h \\ & \leq O\left(\frac{\log^{1.5} n}{\sqrt{b_1}}\right) \|g\|_2 \|A h\|_2 + O\left(\frac{\log^{1.5} n}{\sqrt{b_2}}\right) \|g^\top A\|_2 \|h\|_2 + O\left(\frac{\log^3 n}{\sqrt{b_1 b_2}}\right) \cdot \|g\|_2 \|h\|_2 \|A\|_F. \end{aligned}$$

□

F EXPERIMENT DETAILS

All our experiments are run on an AMD Ryzen 3960X CPU with 128 Gigabytes RAM. We also disable the parallel computing among pairs of graphs for fair running time comparison. In calculating the kernel, we follow the formula described in Section 3.1 and 3.2, using the technique introduced in Section 4.2 and 4.3. Follow Du et al. (2019a), during GNTK learning, we tune the number of AGGREGATE operations, the number of fully connected layers in each COMBINE operation, and the normalization parameter c_u . We also use the C -SVM as the final classifier, and use grid search from 120 values evenly chosen from $[10^{-2}, 10^4]$ to find the best C value.

Note that different choice of hyper-parameters will result in different learning time. Thus, we use the same optimum parameters reported in Du et al. (2019a) and compare the performance in Section 6. Specifically, for social networking datasets COLLAB, IMDBBINARY and IMDBMULTI, we set the number of AGGREGATE operations to be 2, the number of fully connected layers in each COMBINE operation to be 2, and c_u to be 1. And for bioinformatics datasets PTC, NCI1, MUTAG and PROTEINS, we set the number of AGGREGATE operations to be 10, the number of fully connected layers in each COMBINE operation to be 1, and c_u to be $1/|\mathcal{N}(u)|$, where $\mathcal{N}(u)$ is the neighborhood of node u .

For our sketching method, we find that current benchmark datasets for graph classification task are generally small, and matrix decoupling method has already results in a descent kernel learning time. As shown in Table 2, the average number of nodes in the graphs are less than $1k$. On small graphs, after matrix reordering and decoupling, the matrix multiplication time won't dominate the overall calculation time. And for very small graphs, the overhead memory access time introduced by sketching method is even larger than the reduced matrix multiplication time. We observe that when the average number of nodes in the graph reaches $10k$ or more, the matrix multiplication time will dominate the running time of whole algorithm. Thus, for future large scale graph classification tasks, according to Section 4.3 and 5, our sketching method will significantly reduce the running time with strictly bounded generalization error.

We conduct experiments to validate that the error introduced by matrix sketching is strictly bounded. Following Lemma 5.4, we validate the error difference between matrix multiplication with and without the sketching method. Specifically, we randomly generate $[n, n]$ matrix A , G and H . And matrix multiplication without sketching is calculated by $M = G^T A H$. For the sketching method, we randomly generate two AMS matrices R and S with size $[\gamma n, n]$ where γ is the sketching ratio. And matrix multiplication with sketching is calculated by $M_{sk} = G^T R^T R A S^T S H$. The experimental error matrix is calculated by $|M - M_{sk}|$, and the theoretical error matrix is calculated by the RHS of Lemma 5.4. We divide both errors

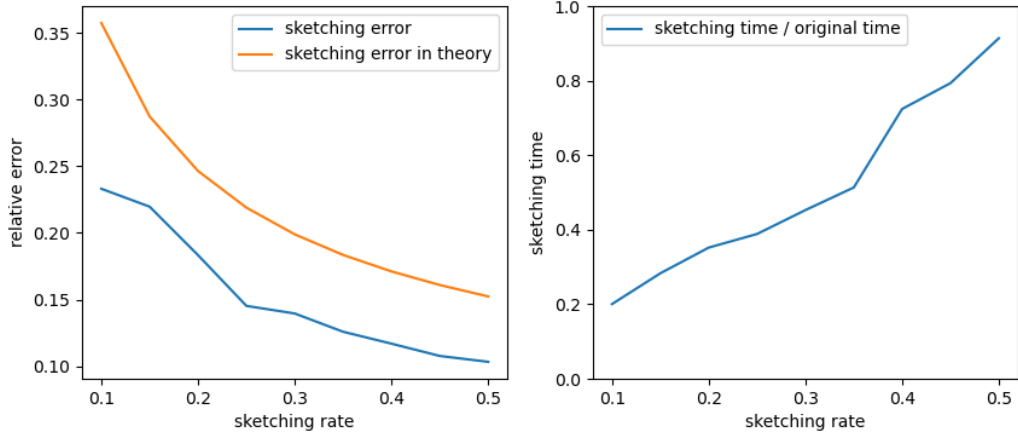


Figure 1: Comparison between theoretical and experimental sketching errors (left) and sketching time (right) under different sketching rates.

by the original matrix M to show the relative error. And we show the final mean error by taking the average over all entries of the error matrices.

The results are shown in Fig. 1. We take $n = 500$, and run experiments under different sketching rates from 0.1 to 0.9. We run each sketching rate for 100 times and calculate the mean error. We also show the comparison between time of matrix multiplication with and without the sketching. Experiments show that our sketching error is always lower than the theoretical bound. When sketching rate gets higher, we lose less information so the error decreases, and in the meantime running time increases because the dimension of the matrix is larger. This experiment validates our Lemma 5.4, showing that our matrix sketching method has a strictly bounded error.