A DIFFERENT DESIGN OF GRAPH WAVELETS

There are many off-the-shelf, well-developed graph wavelets we can choose. They mainly focus on extracting features from multiple frequency bands of input signal spectrum. Some of them are shown as follows.

Monic Cubic wavelets. Monic Cubic wavelets (Hammond et al., 2011) define the kernel function $h(\lambda)$ as

$$h(\lambda) = \begin{cases} \lambda & \text{for} \quad \lambda < 1; \\ -5 + 11\lambda - 6\lambda^2 + \lambda^3 & \text{for} \quad 1 \leq \lambda \leq 2; \\ 2/\lambda & \text{for} \quad \lambda > 2. \end{cases}$$

Different scales of filters are implemented by scaling and translation of above kernel function.

Itersine wavelets. Itersine wavelets define the kernel function at scale j as

$$h_j(\lambda) = \sin\left(\frac{\pi}{2}\cos^2(\pi(\lambda - \frac{j-1}{2}))\right) \mathbb{1}\left[\frac{j}{2} - 1 \leq \lambda \leq \frac{j}{2}\right].$$

Itersine wavelets form tight frames.

Geometric scattering wavelets. Geometric scattering wavelet filter bank (Gao et al., 2019) contains a set of filters based on lazy random walk matrix. The filter at scale j is defined as $\mathbf{H}_j(\mathbf{S}) = \mathbf{S}^{2^{j-1}} - \mathbf{S}^{2^j} = \mathbf{S}^{2^{j-1}}(\mathbf{I} - \mathbf{S}^{2^{j-1}})$, where $\mathbf{S} = \frac{1}{2}(\mathbf{I} + \mathbf{A}\mathbf{D}^{-1})$ is the lazy random walk matrix and $\mathbf{D}$ is the degree matrix.

Note that one is also allowed to customize either spatial or temporal graph wavelets, once they conform a frame and satisfy integral Lipschitz constraint shown as follows

$$A^2\|\mathbf{x}\|^2 \leq \sum_{j=1}^{J} \|\mathbf{H}_j\mathbf{x}\|^2 \leq B^2\|\mathbf{x}\|^2, \quad |\lambda h'(\lambda)| \leq \text{const} \ \ \forall\lambda,$$

where A, B are scalar constants and $h'(\cdot)$ is the gradient of the kernel function.

B PROOFS

B.1 PROOF OF LEMMA 1

By reshaping the signal from $\mathbf{Z}$ to $\mathbf{z}$ with $\mathbf{Z}_{s,t} = \mathbf{z}_{(s-1)T+t}$, we can have that

$$\sum_{j_1,j_2=1}^{J_s,J_t} \|(\mathbf{H}_{j_1}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2}(\mathbf{S}_t))\mathbf{z}\|^2 = \sum_{j_1,j_2=1}^{J_s,J_t} \left\|\mathbf{H}_{j_1}(\mathbf{S}_s)\mathbf{Z}\mathbf{G}_{j_2}^{\top}(\mathbf{S}_t)\right\|^2.$$

Since $\mathbf{S}_s$ and $\mathbf{S}_t$ do not change over computation process, we just use $\mathbf{H}_{j_1}$ and $\mathbf{G}_{j_2}$ to represent $\mathbf{H}_{j_1}(\mathbf{S}_s)$ and $\mathbf{G}_{j_2}(\mathbf{S}_t)$, respectively. Suppose $\mathbf{H}_{j_1} = \begin{pmatrix} h_{11} & & h_{1N} \\ & \ddots & \\ h_{N1} & & h_{NN} \end{pmatrix} \in \mathbb{R}^{N\times N}$, then we have the Kronecker product as $\mathbf{H}_{j_1} \otimes \mathbf{G}_{j_2} = \begin{pmatrix} h_{11}\mathbf{G}_{j_2} & & h_{1N}\mathbf{G}_{j_2} \\ & \ddots & \\ h_{N1}\mathbf{G}_{j_2} & & h_{NN}\mathbf{G}_{j_2} \end{pmatrix}$. Apply it to vector $\mathbf{z}$ and we can have a filtered signal $\mathbf{y}_{j_1,j_2} = (\mathbf{H}_{j_1} \otimes \mathbf{G}_{j_2})\mathbf{z} \in \mathbb{R}^{NT}$. The first T elements of $\mathbf{y}$ can also be written as

$$\mathbf{y}_{j_1,j_2}(1:T) = \sum_{i=1}^{N} h_{1i}\mathbf{G}_{j_2} \begin{pmatrix} \mathbf{Z}_{i,1} \\ \mathbf{Z}_{i,2} \\ \vdots \\ \mathbf{Z}_{i,T} \end{pmatrix} = \mathbf{G}_{j_2} \sum_{i=1}^{N} h_{1i} \begin{pmatrix} \mathbf{Z}_{i,1} \\ \mathbf{Z}_{i,2} \\ \vdots \\ \mathbf{Z}_{i,T} \end{pmatrix}.$$

Therefore we have

$$A_2^2 \left\| \sum_{i=1}^{N} h_{1i} \begin{pmatrix} \mathbf{Z}_{i,1} \\ \mathbf{Z}_{i,2} \\ \vdots \\ \mathbf{Z}_{i,T} \end{pmatrix} \right\|^2 \leq \sum_{j_2} \|\mathbf{y}_{j_1,j_2}(1:T)\|^2 \leq B_2^2 \left\| \sum_{i=1}^{N} h_{1i} \begin{pmatrix} \mathbf{Z}_{i,1} \\ \mathbf{Z}_{i,2} \\ \vdots \\ \mathbf{Z}_{i,T} \end{pmatrix} \right\|^2.$$

Thus $\sum_{j_2} \|\mathbf{y}_{j_1,j_2}\|^2$ can be sandwiched as

$$A_2^2 \sum_{k=1}^{N} \left\| \sum_{i=1}^{N} h_{ki} \begin{pmatrix} \mathbf{Z}_{i,1} \\ \mathbf{Z}_{i,2} \\ \vdots \\ \mathbf{Z}_{i,T} \end{pmatrix} \right\|^2 \leq \sum_{j_2} \|\mathbf{y}_{j_1,j_2}\|^2 \leq B_2^2 \sum_{k=1}^{N} \left\| \sum_{i=1}^{N} h_{ki} \begin{pmatrix} \mathbf{Z}_{i,1} \\ \mathbf{Z}_{i,2} \\ \vdots \\ \mathbf{Z}_{i,T} \end{pmatrix} \right\|^2.$$

By definition of vector ℓ_2 norm, we can rewrite the upper and lower bound in Eq. (6) as

$$A_2^2 \sum_{i=1}^{T} \left\| \mathbf{H}_{j_1} \begin{pmatrix} \mathbf{Z}_{1,i} \\ \mathbf{Z}_{2,i} \\ \vdots \\ \mathbf{Z}_{N,i} \end{pmatrix} \right\|^2 \leq \sum_{j_2} \|\mathbf{y}_{j_1,j_2}\|^2 \leq B_2^2 \sum_{i=1}^{T} \left\| \mathbf{H}_{j_1} \begin{pmatrix} \mathbf{Z}_{1,i} \\ \mathbf{Z}_{2,i} \\ \vdots \\ \mathbf{Z}_{N,i} \end{pmatrix} \right\|^2.$$

Summing above quantity over j_1 gives us that

$$A_1^2 A_2^2 \|\mathbf{Z}\|^2 = A_1^2 A_2^2 \sum_{i=1}^{T} \left\| \begin{pmatrix} \mathbf{Z}_{1,i} \\ \mathbf{Z}_{2,i} \\ \vdots \\ \mathbf{Z}_{N,i} \end{pmatrix} \right\|^2 \leq \sum_{j_1,j_2} \|\mathbf{y}_{j_1,j_2}\|^2 \leq B_1^2 B_2^2 \sum_{i=1}^{T} \left\| \begin{pmatrix} \mathbf{Z}_{1,i} \\ \mathbf{Z}_{2,i} \\ \vdots \\ \mathbf{Z}_{N,i} \end{pmatrix} \right\|^2 = B_1^2 B_2^2 \|\mathbf{Z}\|^2,$$

which completes the proof. Lemma 1 is a very handful result. It shows that we can easily construct new spatio-temporal wavelets just by combining spatio and temporal ones. Moreover, the constants for new frame bound can be easily obtained once we know the characteristics of the wavelets in each domain. In particular, it also provides us a convenient way to build tight frames for spatio-temporal data analysis with $A = B$, because we just need to choose tight frames for spatial and temporal domain separately without considering possible correlations.

B.2 Proof of Theorem 1

We are considering pooling operator $U(\cdot)$ as average in spatial domain in this proof, so $\mathbf{U} = \frac{1}{N}\mathbf{1}_{1\times N}$ and $\boldsymbol{\phi} = \mathbf{U}\mathbf{Z} \in \mathbb{R}^T$. The proof techniques can be easily generalized to any form of $U(\cdot)$. When reshaping $\mathbf{Z} \in \mathbb{R}^{N\times T}$ to $\mathbf{z} \in \mathbb{R}^{NT}$, the new pooling operator can be simply represented as

$$\mathbf{U}' = \frac{1}{N}(\mathbf{I}_T, \mathbf{I}_T, \cdots, \mathbf{I}_T) \in \mathbb{R}^{T\times NT}, \quad \boldsymbol{\phi} = \mathbf{U}'\mathbf{z}.$$

Note that $\|\mathbf{U}'\|_2 = \frac{1}{\sqrt{N}}$. Consider scattering tree nodes at the last layer $L-1$. Suppose they are indexed from 1 to J^{L-1} associated with signal $\mathbf{a}_1, \cdots, \mathbf{a}_{J^{L-1}}$, and their parent nodes are indexed from 1 to J^{L-2} associated with signal $\mathbf{b}_1, \cdots, \mathbf{b}_{J^{L-2}}$. When the input data $\mathbf{X}$ is perturbed, all signals in scattering tree will change correspondingly. Here we simply denote them as $\widetilde{\mathbf{a}}, \widetilde{\mathbf{b}}$. Then for the change of feature vector located at node with $\mathbf{a}_1$, it holds that

$$\|\boldsymbol{\phi}_{\mathbf{a}_1} - \boldsymbol{\phi}_{\widetilde{\mathbf{a}}_1}\|^2 = \|\mathbf{U}'(\mathbf{a}_1 - \widetilde{\mathbf{a}}_1)\|^2 \leq \|\mathbf{U}'\|^2 \|\mathbf{a}_1 - \widetilde{\mathbf{a}}_1\|^2 \leq \frac{1}{N}\|\sigma((\mathbf{H}_{j_1} \otimes \mathbf{G}_{j_2})(\mathbf{b}_1 - \widetilde{\mathbf{b}}_1))\|^2, \quad (6)$$

where $j_1 = j_2 = 1$. The last inequality holds because we are using absolute value function as nonlinear activation, which is non-expansive. Summing above quantity over j_1, j_2 and by the frame bound proved in Lemma 1, we can have that

$$\sum_{i=1}^{J^{L-1}} \|\boldsymbol{\phi}_{\mathbf{a}_i} - \boldsymbol{\phi}_{\widetilde{\mathbf{a}}_i}\|^2 \leq \frac{B^2}{N} \sum_{i=1}^{J^{L-2}} \|\mathbf{b}_i - \widetilde{\mathbf{b}}_i\|^2. \quad (7)$$

Note that for sum of square norm of change at layer $L-2$ it is

$$\sum_{i=1}^{J^{L-2}} \|\boldsymbol{\phi}_{\mathbf{b}_i} - \boldsymbol{\phi}_{\widetilde{\mathbf{b}}_i}\|^2 \leq \frac{1}{N} \sum_{i=1}^{J^{L-2}} \|\mathbf{b}_i - \widetilde{\mathbf{b}}_i\|^2. \tag{8}$$

Compare Eq. (7) and (8). The upper bound only differs with a factor B^2. Then by induction we can have that

$$\|\Phi(\mathbf{S}_s, \mathbf{S}_t, \mathbf{X}) - \Phi(\mathbf{S}_s, \mathbf{S}_t, \widetilde{\mathbf{X}})\|^2 \leq \frac{1}{N} \sum_{\ell=0}^{L-1} B^{2\ell} \|\mathbf{x} - \widetilde{\mathbf{x}}\|^2 = \frac{1}{N} \sum_{\ell=0}^{L-1} B^{2\ell} \|\boldsymbol{\Delta}\|^2.$$

Normalize it with the dimension of final feature map, we have

$$\frac{\|\Phi(\mathbf{S}_s, \mathbf{S}_t, \mathbf{X}) - \Phi(\mathbf{S}_s, \mathbf{S}_t, \widetilde{\mathbf{X}})\|}{\sqrt{T \sum_{\ell=0}^{L-1} J^\ell}} \leq \frac{1}{\sqrt{NT}} \sqrt{\frac{\sum_{\ell=0}^{L-1} B^{2\ell}}{\sum_{\ell=0}^{L-1} J^\ell}} \|\boldsymbol{\Delta}\|. \tag{9}$$

B.3 Proof of Theorem 2

Perturbations on the underlying graph usually happen when the graph is unknown or when the graph changes over time (Segarra et al., 2017). Take skeleton-based action recognition as an example. Some joints may be misrecognized with others due to measurement noise of devices during certain frames, thus the location signals of those joints are interchanged. This leads to different spatial graph structures at those time stamps. Since such kind of perturbations usually happen in spatial domain, here we simply consider the structure perturbations on the spatial graph only. But the results can be extended to more general cases.

Consider the original spatio-temporal graph as $\mathcal{G}$ with spatial graph shift matrix $\mathbf{S}_s$ and temporal one $\mathbf{S}_t$, and the perturbed graph as $\widehat{\mathcal{G}}$ with $\widehat{\mathbf{S}}_s$ and $\mathbf{S}_t$. We first show that ST-GST is invariant to node permutations in spatial domain, where the set of permutation matrices is defined as $\mathcal{P} = \{\mathbf{P} \in \{0,1\}^{N \times N} : \mathbf{P}1 = 1, \mathbf{P}^\top 1 = 1, \mathbf{P}\mathbf{P}^\top = \mathbf{I}_N\}$. Note that we are considering average in spatial domain for $U(\cdot)$, so $\mathbf{U} = \frac{1}{N}\mathbf{1}_{1\times N}$ and $\boldsymbol{\phi} = \mathbf{U}\mathbf{Z} \in \mathbb{R}^T$, $\widehat{\mathbf{U}} = \mathbf{U}\mathbf{P}$.

Lemma 2. Consider the spatial permutation $\widehat{\mathbf{S}}_s = \mathbf{P}^\top \mathbf{S}_s \mathbf{P}$ and input data $\widehat{\mathbf{X}} = \mathbf{P}^\top \mathbf{X}$ are also permuted in spatial domain correspondingly. Then, it holds that

$$\Phi(\mathbf{S}_s, \mathbf{S}_t, \mathbf{X}) = \Phi(\widehat{\mathbf{S}}_s, \mathbf{S}_t, \widehat{\mathbf{X}}) \tag{10}$$

Proof. Note that the permutation holds for all signals computed in scattering tree; that is to say, $\widehat{\mathbf{Z}}_{(p^{(\ell)})} = \mathbf{P}^\top \mathbf{Z}_{(p^{(\ell)})}$. Suppose for path $p^{(\ell)}$ the last two filter are chosen as $\mathbf{H}(\widehat{\mathbf{S}}_s)$ and $\mathbf{G}(\mathbf{S}_t)$, then the feature vector after pooling with respect to new graph support and data can be written as

$$\begin{aligned}
\boldsymbol{\phi}_{(p^{(\ell)})}(\widehat{\mathbf{S}}_s, \mathbf{S}_t, \widehat{\mathbf{Z}}_{(p^{(\ell)})}) &= \widehat{\mathbf{U}}(\sigma(\mathbf{H}(\widehat{\mathbf{S}}_s)\widehat{\mathbf{Z}}_{(p^{(\ell)})}\mathbf{G}^\top(\mathbf{S}_t))) \\
&= \mathbf{U}\mathbf{P}\sigma(\mathbf{P}^\top \mathbf{H}(\mathbf{S}_s)\mathbf{P}\mathbf{P}^\top \mathbf{Z}_{(p^{(\ell)})}\mathbf{G}^\top(\mathbf{S}_t))
\end{aligned}$$

The last equation holds due to definition of $\mathbf{H}(\mathbf{S})$. Since nonlinear activation is applied element-wise, we can rewrite it as

$$\begin{aligned}
\boldsymbol{\phi}_{(p^{(\ell)})}(\widehat{\mathbf{S}}_s, \mathbf{S}_t, \widehat{\mathbf{Z}}_{(p^{(\ell)})}) &= \mathbf{U}\sigma(\mathbf{P}\mathbf{P}^\top \mathbf{H}(\mathbf{S}_s)\mathbf{P}\mathbf{P}^\top \mathbf{Z}_{(p^{(\ell)})}\mathbf{G}^\top(\mathbf{S}_t)) \\
&= \mathbf{U}\sigma(\mathbf{H}(\mathbf{S}_s)\mathbf{Z}_{(p^{(\ell)})}\mathbf{G}^\top(\mathbf{S}_t)) \\
&= \boldsymbol{\phi}_{(p^{(\ell)})}(\mathbf{S}_s, \mathbf{S}_t, \mathbf{Z}_{(p^{(\ell)})}).
\end{aligned}$$

This conclusion holds independently of specific path $p^{(\ell)}$, so it holds for all feature vector after pooling in scattering tree. Since final feature map is just a concatenation of all feature vectors, the proof is complete. □

Lemma 2 shows that the output of ST-GST is essentially independent of the node ordering in spatial domain, as long as the permutation is consistent across all time stamps. This result is intuitive

because the output of graph convolution should only depend on relative neighborhood structure of each node. Since node reordering will not alter neighborhood topology, the output should remain unchanged.

Based on Lemma 2, we use a relative perturbation model for structure modifications (Gama et al., 2019b), which focuses more on the change of neighborhood topology compared to absolute perturbations adopted in Levie et al. (2019). Define the set of permutations that make $\mathbf{S}_s$ and $\widehat{\mathbf{S}}$ the closet as $\mathcal{P}_s := \arg\min_{\mathbf{P}\in\mathcal{P}} \|\mathbf{P}^\top\widehat{\mathbf{S}}_s\mathbf{P} - \mathbf{S}_s\|_2$. Consider the set of perturbation matrices $\mathcal{E}(\mathbf{S},\widehat{\mathbf{S}}) = \{\mathbf{E}|\mathbf{P}^\top\widehat{\mathbf{S}}_s\mathbf{P} = \mathbf{S}_s + \mathbf{E}^\top\mathbf{S}_s + \mathbf{S}_s\mathbf{E}, \mathbf{P}\in\mathcal{P}_s, \mathbf{E}\in\mathbb{R}^{N\times N}\}$. Then the relative distance to measure structure perturbations can be defined as

$$d(\mathbf{S}_s,\widehat{\mathbf{S}}_s) = \min_{\mathbf{E}\in\mathcal{E}(\mathbf{S}_s,\widehat{\mathbf{S}}_s)} \|\mathbf{E}\|_2$$

Note that if $\widehat{\mathbf{S}}_s = \mathbf{P}^\top\mathbf{S}_s\mathbf{P}$, meaning that the structure perturbation is purely permutation, then the relative distance $d(\mathbf{S}_s,\widehat{\mathbf{S}}_s) = 0$, which is consistent with result shown in Lemma 2. Therefore, without loss of generality, we can assume that $\mathbf{P} = \mathbf{I}_N$ and $\widehat{\mathbf{S}}_s = \mathbf{S}_s + \mathbf{E}^\top\mathbf{S}_s + \mathbf{S}_s\mathbf{E}$ in later context. With this formulation, we are ready to prove Lemma 3.

Lemma 3. Suppose eigenvalues $\{m_i\}_{i=1}^N$ of $\mathbf{E}$ are organized in order such that $|m_1| \leq |m_2| \leq \cdots \leq |m_N|$, satisfying $|m_N| \leq \epsilon/2$ and $|m_i/m_N - 1| \leq \epsilon$ for $\epsilon > 0$. For spatial graph filter $\mathbf{H}(\mathbf{S}_s)$ and temporal graph filter $\mathbf{G}(\mathbf{S}_t)$, denote their kernel functions as $h(\lambda)$ and $g(\lambda)$, respectively. If for all λ, $h(\lambda)$ is chosen to satisfy integral Lipschitz constraint $|\lambda h'(\lambda)| \leq C$ and $g(\lambda)$ has bounded spectral response $|g(\lambda)| \leq D$. Then it holds that

$$\|\mathbf{H}(\mathbf{S}_s)\otimes\mathbf{G}(\mathbf{S}_t) - \mathbf{H}(\widehat{\mathbf{S}}_s)\otimes\mathbf{G}(\mathbf{S}_t)\|_2 \leq \epsilon CD + O(\epsilon^2). \tag{11}$$

Proof. From Proposition 2 in Gama et al. (2019b) we can have that when $\mathbf{E}$ satisfies above conditions, $\|\mathbf{H}(\mathbf{S}_s) - \mathbf{H}(\widehat{\mathbf{S}}_s)\|_2 \leq \epsilon C + O(\epsilon^2)$. So

$$\begin{aligned}\|\mathbf{H}(\mathbf{S}_s)\otimes\mathbf{G}(\mathbf{S}_t) - \mathbf{H}(\widehat{\mathbf{S}}_s)\otimes\mathbf{G}(\mathbf{S}_t)\|_2 &= \|(\mathbf{H}(\mathbf{S}_s) - \mathbf{H}(\widehat{\mathbf{S}}_s))\otimes\mathbf{G}(\mathbf{S}_t)\|_2\\ &\leq \|\mathbf{H}(\mathbf{S}_s) - \mathbf{H}(\widehat{\mathbf{S}}_s)\|_2\|\mathbf{G}(\mathbf{S}_t)\|_2\\ &\leq \epsilon CD + O(\epsilon^2),\end{aligned}$$

The second line holds because $\mathbf{H}(\mathbf{S}_s) - \mathbf{H}(\widehat{\mathbf{S}}_s)$ is a symmetric matrix, which can be written as eigen-decomposition as $\mathbf{F}\mathbf{\Omega}\mathbf{F}^\top$. And $(\mathbf{F}\mathbf{\Omega}\mathbf{F}^\top)\otimes(\mathbf{V}\mathbf{\Lambda}\mathbf{V}^T) = (\mathbf{F}\otimes\mathbf{V})(\mathbf{\Omega}\otimes\mathbf{\Lambda})(\mathbf{F}\otimes\mathbf{V})^\top$ holds, which finishes the proof. As for general structural perturbations, where we want to find $\|\mathbf{H}(\mathbf{S}_s)\otimes\mathbf{G}(\mathbf{S}_t) - \mathbf{H}(\widehat{\mathbf{S}}_s)\otimes\mathbf{G}(\widehat{\mathbf{S}}_t)\|_2$, we can add and subtract term $\mathbf{H}(\widehat{\mathbf{S}}_s)\otimes\mathbf{G}(\widehat{\mathbf{S}}_t)$, use triangle inequality and further bound those two terms with more assumptions on $h(\lambda)$ and $g(\lambda)$. □

The bound shown in Lemma 3 indicates that the difference of output caused by changing spatial graph support from $\mathbf{S}_s$ to $\widehat{\mathbf{S}}_s$ is proportional to ϵ, which is a scalar characterizing the level of the perturbation. Constraints on eigenvalues of $\mathbf{E}$ limits the change of graph structure. A more detailed description explaining the necessity of such constraints can be found in Gama et al. (2019b). With Lemma 3 in hand, we are ready to show the change of feature vector after pooling at each node in scattering tree when such structure perturbations happen.

Lemma 4. Consider a ST-GST with L layers and $J = J_s \times J_t$ scales at each layer. Suppose that the graph filter bank forms a frame with upper bound $B = B_1 \times B_2$, where B_1, B_2 are frame bounds for spatial and temporal domain, respectively. Suppose for all λ, spatial wavelet filter bank $\{\mathbf{H}_{j_1}\}_{j_1=1}^{J_s}$ satisfies $\max_i |\lambda h_i'(\lambda)| \leq C$ and temporal wavelet filter bank $\{\mathbf{G}_{j_2}\}_{j_2=1}^{J_t}$ satisfies $\max_i |g_i(\lambda)| \leq D$, and other conditions the same as Lemma 3. Then for the change of feature vector $\boldsymbol{\phi}_{p^{(\ell)}}$ associated with path $p^{(\ell)}$ it holds that

$$\|\boldsymbol{\phi}_{p^{(\ell)}}(\mathbf{S}_s,\mathbf{S}_t,\mathbf{X}) - \boldsymbol{\phi}_{p^{(\ell)}}(\widehat{\mathbf{S}}_s,\mathbf{S}_t,\mathbf{X})\| \leq \frac{1}{\sqrt{N}}\epsilon\ell CDB^{\ell-1}\|\mathbf{X}\|. \tag{12}$$

Proof. Expand $\|\boldsymbol{\phi}_{p^{(\ell)}}(\mathbf{S}_s, \mathbf{S}_t, \mathbf{X}) - \boldsymbol{\phi}_{p^{(\ell)}}(\widehat{\mathbf{S}}_s, \mathbf{S}_t, \mathbf{X})\|$ as

$$\begin{aligned}
&\|\mathbf{U}'(\sigma(\mathbf{H}_{j_1^{(\ell)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}\mathbf{x} - \mathbf{U}'(\sigma(\mathbf{H}_{j_1^{(\ell)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}\mathbf{x}\| \\
\leq &\frac{1}{\sqrt{N}}\|(\sigma(\mathbf{H}_{j_1^{(\ell)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}\mathbf{x} - (\sigma(\mathbf{H}_{j_1^{(\ell)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}\mathbf{x}\|,
\end{aligned}$$

where $\|\mathbf{U}'\|_2 = 1/\sqrt{N}$ and $(\sigma(\mathbf{H}_{j_1^{(\ell)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}$ is a shorthand for applying spatio-temporal filters and nonlinear activation in order to input data ℓ times according to the path $p^{(\ell)}$. Add and subtract term $\sigma(\mathbf{H}_{j_1^{(\ell)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t))\sigma(\mathbf{H}_{j_1^{(\ell-1)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(\ell-1)}}(\mathbf{S}_t)) \cdots \sigma(\mathbf{H}_{j_1^{(1)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(1)}}(\mathbf{S}_t))\mathbf{x}$ and apply triangle inequality, we can have that

$$\begin{aligned}
&\|(\sigma(\mathbf{H}_{j_1^{(\ell)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}\mathbf{x} - (\sigma(\mathbf{H}_{j_1^{(\ell)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}\mathbf{x}\| \\
\leq &\|\sigma(\mathbf{H}_{j_1^{(\ell)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)) \Big((\sigma(\mathbf{H}_{j_1^{(\ell-1)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell-1)}}(\mathbf{S}_t)))_{p^{(\ell-1)}} - \\
&(\sigma(\mathbf{H}_{j_1^{(\ell-1)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(\ell-1)}}(\mathbf{S}_t)))_{p^{(\ell-1)}}\Big)\mathbf{x}\| + \\
&\|\Big(\sigma(\mathbf{H}_{j_1^{(\ell)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)) - \sigma(\mathbf{H}_{j_1^{(\ell)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t))\Big) \cdot \\
&(\sigma(\mathbf{H}_{j_1^{(\ell-1)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(\ell-1)}}(\mathbf{S}_t)))_{p^{(\ell-1)}}\mathbf{x}\|.
\end{aligned}$$

Recursive quantities can be observed above and the bound can be solved explicitly (Gama et al., 2019b). By induction and conclusion from Lemma 3, we can get that

$$\|(\sigma(\mathbf{H}_{j_1^{(\ell)}}(\mathbf{S}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}\mathbf{x} - (\sigma(\mathbf{H}_{j_1^{(\ell)}}(\widehat{\mathbf{S}}_s) \otimes \mathbf{G}_{j_2^{(\ell)}}(\mathbf{S}_t)))_{p^{(\ell)}}\mathbf{x}\| \leq \ell\epsilon CDB^{\ell-1}\|\mathbf{x}\|.$$

Multiplying the coefficient $1/\sqrt{N}$ caused by pooling gets us the final result. □

Note that the upper bound in Lemma 4 holds for all path of length ℓ. Thus the square norm of change in final feature map can be summarized by the sum of square norm of change at each layer, which finishes the proof of Theorem 2.

C Additional Experiments

C.1 Dataset

MSR Action3D dataset (Li et al., 2010) is a small dataset capturing indoor human actions. It covers 20 action types and 10 subjects, with each subject repeating each action 2 or 3 times. The dataset contains 567 action clips with maximum number of frames 76; however, 10 of them are discarded because the skeleton information are either missing or too noisy (Wang et al., 2012). For each clip, locations of 20 joints are recorded, and only one subject is present. Training and testing set is decided by cross-subject split for this dataset, with 288 samples for training and 269 for testing.

NTU-RGB+D (Liu et al., 2019) is currently the largest dataset with 3D joints annotations for human action recognition task. It covers 60 action types and 40 subjects. The dataset contains 56,880 action clips with maximum number of frames 300, and there are 25 joints for each subject in one clip. Each clip is guaranteed to have at most 2 subjects. The cross-subject benchmark of NTU-RGB+D includes 40,320 clips for training and 16,560 for testing.

Full table of performance on MSR Action3D dataset. The table contains performance comparison for different algorithms with different set of parameters on MSR Action3D dataset. Note that the triple shown after ST-GST represents the value for (J_s, J_t, L). Methods labeled "fixed topology" are modified so as not to use adaptive training of the adjacency matrix in order for the comparison with ST-GST to be fair. Methods labeled "learnable topology" means that we use adaptive training for adjacency matrix to further validate our claim. Other configurations of compared methods are then set by default. From the table we can see that ST-GST outperforms all other methods even when the graph topology can be learned by neural networks. The intuition behind this is that deep learning methods need large amount of training data due to the complex structures, and it can easily

	Method	Accuracy (%)
	GFT+TPM	74.0
	HDM	81.8
GNNs	ST-GCN (fixed topology)	52.0
GNNs	ST-GCN (learnable topology)	56.0
GNNs	Temporal Conv. (resnet)	67.3
GNNs	Temporal Conv. (resnet-v3-gap)	69.9
GNNs	Temporal Conv. (resnet-v4-gap)	72.1
GNNs	MS-G3D (GCN scales=10, G3D scales=6)	80.3
GNNs	MS-G3D (GCN scales=5, G3D scales=5)	81.4
GNNs	MS-G3D (GCN scales=8, G3D scales=5)	82.2
Scattering	Separable ST-GST (5, 5, 3)	73.6
Scattering	Separable ST-GST (5, 5, 4)	72.9
Scattering	Separable ST-GST (5, 10, 3)	81.4
Scattering	Separable ST-GST (5, 15, 3)	85.9
Scattering	Separable ST-GST (5, 20, 3)	**87.0**
Scattering	Joint Kronecker ST-GST (15, 3)	61.0
Scattering	Joint Cartesian ST-GST (15, 3)	59.1
Scattering	Joint Strong ST-GST (15, 3)	61.7

Table 3: Full comparison of classification accuracy (MSR Action3D with 288 training and 269 testing samples).

Method	Accuracy (%)
Separable ST-GST (5, 5, 3)	73.4 ± 0.8
Separable ST-GST (5, 20, 3)	86.7 ± 0.4
Joint Kronecker ST-GST (5, 3)	46.3 ± 1.2
Joint Cartesian ST-GST (5, 3)	42.2 ± 1.1
Joint Strong ST-GST (5, 3)	45.0 ± 1.2
Joint Kronecker ST-GST (15, 3)	59.6 ± 0.5
Joint Cartesian ST-GST (15, 3)	58.6 ± 1.0
Joint Strong ST-GST (15, 3)	60.0 ± 1.0

Table 4: Performance for different methods on MSR Action3D with standard deviations.

be trapped into bad local optima due to overfitting when the size of training set is limited, which is common in practice. Also the good performance of ST-GST in sparse label regime could potentially inspire active learning for processing spatio-temporal data (Bilgic et al., 2010).

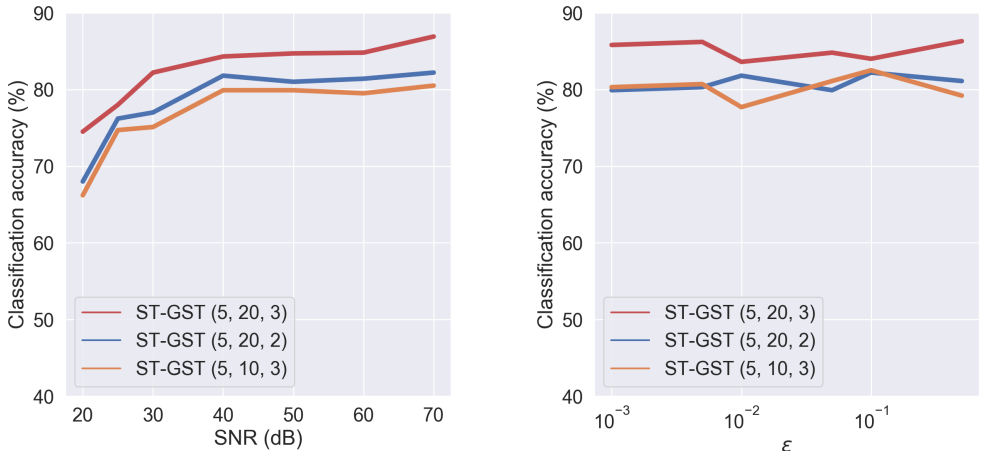
Performance on MSR Action3D dataset with standard deviations. We repeat part of our experiments 20 times on MSR Action3D dataset, especially for joint approaches, to obtain the standard deviations of classification accuracy. The results are shown in Table 4. Note that since ST-GST is a mathematically designed transform, the output features should be the same for different trails, and the randomness comes from classifiers used later (random forest in this case). It can be seen that the standard deviations are comparable in all these methods, and therefore the conclusion that separable ST-GST consistently outperforms joint ST-GST still holds.

Comparison between different choices of wavelets. In practice we find that using graph geometric scattering wavelets (Gao et al., 2019) for both spatial and temporal domain can achieve the best performance, which is reported in main text. Classification accuracy using other type of wavelets is shown here. All experiments performed here are separable ST-GST with $J_s = 5, J_t = 15, L = 3$ on MSR Action3D dataset. An interesting observation is that there is a significant reduction in accuracy when we change temporal wavelet from diffusion based one (Geometric) to spectrum based one (MonicCubic or Itersine). This may caused by the design of different wavelets.

Stability of ST-GST. We also show the classification accuracy under different level of perturbations on spatio-temporal signals and spatial graph structures in Fig. 4. The experiments are con-

Spatial wavelet	Temporal wavelet	Accuracy (%)
Geometric	Geometric	85.9
Geometric	MonicCubic	76.6
Geometric	Itersine	73.6
MonicCubic	Geometric	82.9
Itersine	Geometric	82.5
MonicCubic	MonicCubic	80.7
MonicCubic	Itersine	78.4
Itersine	MonicCubic	76.2
Itersine	Itersine	80.7

Table 5: Performance for different choices of spatial and temporal wavelets (MSR Action3D) with setting (5, 15, 3).



(a) Signal perturbations.

(b) Structure perturbations.

Figure 4: Comparisons on performance under different level of perturbations.

ducted on MSR Action3D dataset. For signal perturbation, signal-to-noise ratio (SNR) is defined as $10 \log \frac{\|\mathbf{X}\|^2}{\|\mathbf{\Delta}\|^2}$. For structure perturbation, $\mathbf{E}$ is set to be a diagonal matrix, whose diagonal elements satisfy corresponding constraints on ϵ. From both Fig. 4(a) and (b) we can see that ST-GST is stable and will not deviate much from original output when the perturbations are small.