▷ Lecture 19 EN.553.744 Luana Ruiz

(THM) Non-asymptotic convergence of graph convolutions.

Given $(G_n, x_n) \sim (w, x)$ and convolutions $H(G_n) = \sum_{k=0}^{\infty} h_k \left(\frac{A_n}{n}\right)^k$ and $T_H = \sum_{k=0}^{\infty} h_k T_w^{(k)}$, under assumptions 3 & 4, we have:

$$\|y_n - y\| \leq \left(L + \frac{\pi n_c}{\delta_c}\right) \|T_w - T_{w_n}\| + (L_c + 2)\|x - x_n\| + 2l_c$$

=> Convergence $(G_n, x_n) \to (w, x)$ (w/ appropriate node labeling) means approximation improves w/ n as expected

=> Convergence-discriminability tradeoff is explicit; larger L & smaller c (= more discriminative filters) lead to higher error bound.

=> In the finite-sample regime, unless $\ell = 0$, there is always leftover "nontransferable energy" $\geq \ell_c$ corresponding to spectral components w/ $|\lambda_i| < c$, which do not converge.

Pf. We give a simplified version here, but the ideas are the same.

The proof is constructive; we leverage the fact that any filter can be written as the sum of a low pass and a high pass filters.

Specifically,

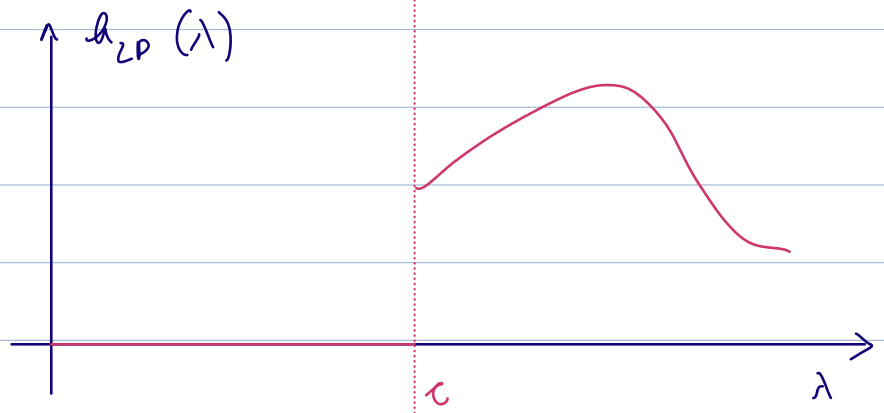
$$h(\lambda) = h_{LP}(\lambda) + h_{HP}(\lambda)$$

where

$$h_{LP}(\lambda) = \begin{cases} h(\lambda), & |\lambda| \geq c \\ 0 & \text{o.w.} \end{cases}$$

$$h_{HP}(\lambda) = \begin{cases} 0, & |\lambda| \geq c \\ h(\lambda) & \text{o.w} \end{cases}$$

E.g.:

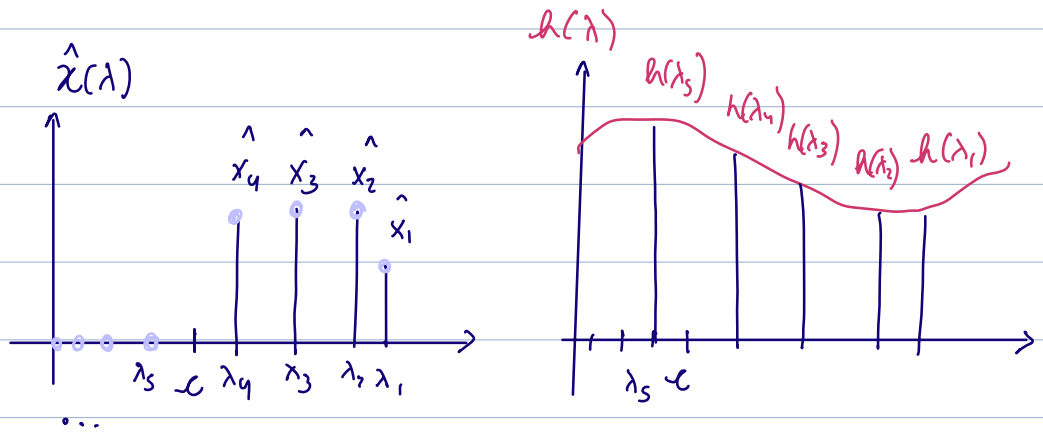


↳ The approximation error of h can then be upper bounded by bounding the approximation errors of h_{LP} & h_{HP}

- We start by upper bounding the approximation error of the low pass component.

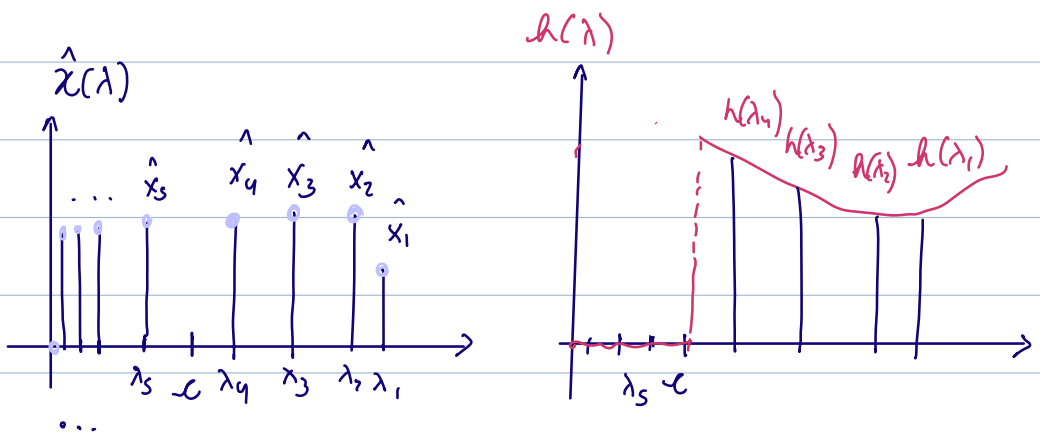
To do that, we note that convergence of convolutions for LP filters is equivalent to convergence of general filters for band limited input signals.

E.g.:



$$\hat{y}_1 = \hat{x}_1 \cdot h(\lambda_1)$$
$$y_2 = \hat{x}_2 \cdot h(\lambda_2)$$
$$\vdots$$
$$\hat{y}_5 = \hat{x}_5 \cdot h(\lambda_5) = 0 \qquad \text{regardless of } h(\lambda)$$
$$y_6 = \hat{x}_6 \cdot h(\lambda_6) = 0$$
$$0$$

This is equivalent to:



Therefore, we can leverage the following step from the proof of asymptotic convergence:

$$\|y - y_n\| \leq \Big\| \sum_{j \in e} g(\lambda_j)\, \hat{x}_j\, \varphi_j - \sum_{j \in e} g(\lambda_j^n)\, [\hat{x}_n]_j\, \varphi_j^n \Big\| \quad (1)$$

Adding and subtracting $\sum_{j \in e} g(\lambda_j^n)\, \hat{x}_j\, \varphi_j$, we get:

$$\|y - y_n\| \leq \Big\| \sum_{j \in e} (g(\lambda_j) - g(\lambda_j^n))\, \hat{x}_j\, \varphi_j \Big\| + \Big\| \sum_{j \in e} g(\lambda_j^n)\, (\hat{x}_j \varphi_j - [\hat{x}_n]_j\, \varphi_j^n) \Big\|$$

$$\leq \underbrace{L}_{\text{Lipschitz}} \cdot \|T_w - T_{w_n}\| \cancel{\|x\|}^{\leq 1} + \sum_{j \in e} \|g(\lambda_j)\| \Big(\|\hat{x}_j\|^{\leq 1} \|\varphi_j - \varphi_j^n\|^{\leq 1} + \Big\| \sum_{j \in e} g(\lambda_j)\, \varphi_j^n\, (\hat{x}_j - [\hat{x}_n]_j) \Big\| \Big)$$

$$+ \underbrace{\big(|\lambda_i - \lambda_i'| \leq \|T - T'\|\big)}_{\text{Lipschitz}}$$

Davis-Kahan

$$\|\varphi_j - \varphi_j^n\| \leq \frac{\pi \|T_W - T_{W_n}\|}{2\delta_c} \qquad \forall j \in \mathcal{C}$$

$$\left\| \sum_{j \in \mathcal{C}} a(\lambda_j) \varphi_j^n \left(\hat{x}_j - (\hat{x}_n)_j \right) \right\| = \left\| \sum_{j \in \mathcal{C}} a(\lambda_j) \varphi_j^n \langle \varphi_j - \varphi_j^n, x \rangle \right.$$

$$\left. + \sum_{j \in \mathcal{C}} a(\lambda_j) \varphi_j^n \langle (x - x_n), \varphi_j^n \rangle \right\|$$

$$\leq n_c \|\varphi_j - \varphi_j^n\| + \|X - X_n\|$$

Leading to:

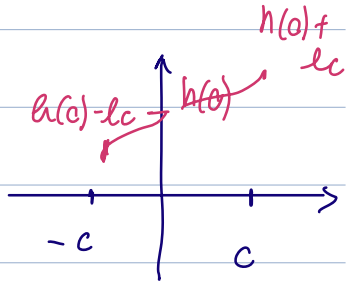
$$\|y - y_n\| \leq L \|T_W - T_{W_n}\| + n_c \left(\frac{\pi \|T_W - T_{W_n}\| \cdot \cancel{2}}{\cancel{2}\delta_c} \right)$$

$$+ \|X - X_n\|$$

(a)

- Next, we consider the high-pass component. We reuse the argument from the proof of asymptotic convergence.

$$\left\| \sum_{j \in \mathbb{Z} \setminus \{0\}} a(\lambda_j)\, \hat{x}_j\, \varphi_j - \sum_{j \in \mathbb{Z} \setminus \{0\}} a(\lambda_j^n)\, (\hat{x}_n)_j\, \varphi_j^n \right\|$$



$$\leq \| (a(0) + l_c)\, x - (a(0) - l_c)\, x_n \|$$

$$\leq |a(0)| \, \| x - x_n \| + l_c \, \| x + x_n \|$$

($|a(0)| \leq 1$)

$$\leq \| x - x_n \| + l_c \, \| 2x + x_n - x \| \qquad (b)$$

$$\leq (1 + l_c) \, \| x - x_n \| + 2 \| x \| \, l_c$$

($\| x \| \leq 1$)

Combining (a) & (b) concludes the proof.

→ From convergence to transferability

In practice, we are not trying to approximate a graphon convolution, but to trans-

fer it; e.g., design on a smaller / offline graph & transfer it to a larger / online (dynamic) graph.

Transferability error bound can be derived from approx. error bound via Δ inequality.

↳ given (G_n, x_n); $(G_m, x_m) \sim (w, x)$ and a graph convolution w/ coeffs. $h_0, h_1, h_2, \ldots$, we have:

$$\| y_n - y_m \| \leq \| y_n - y \| + \| y_m - y \|$$

$\| y_n - y_m \|$: induced graphon signals, in order to compare graph signals of $\neq$ sizes

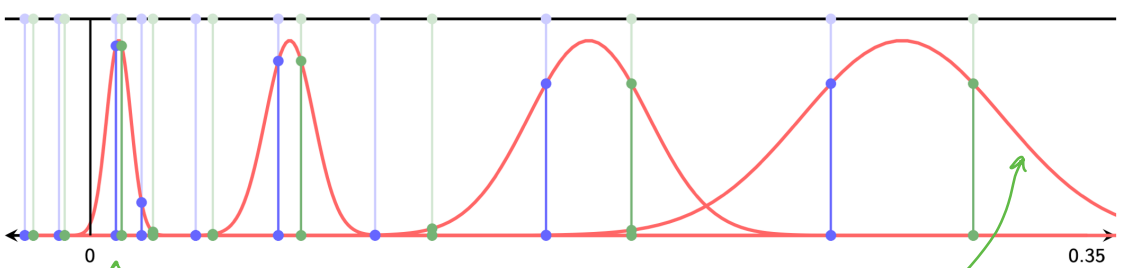
$\| y_n - y \| + \| y_m - y \|$: sum of the approx. error bounds on G_n & G_m

The theorem is essentially the same, since it will often be dominated by the approximation

error of the smallest graph, twice the approximation error for $\min\{n, m\}$ is a safe upper bound for the transferability error.

↳ Takeaways are the same

The most important: tradeoff between transf. & discriminability



($\uparrow L$)

sharper filters are more discriminative but less transferable

($\downarrow L$)

smoother filters are more transferable but less discriminative

To improve the tradeoff, we need GNNs.

→ Transferability of GNNs

GNNs inherit the transferability property from graph filters.

(AS 5) The nonlinearities σ are normalized Lipschitz ($L=1$). → satisfied by ReLU, sigmoid, tanh ...

(THM) GNN transferability

Given $(G_n, x_n) \sim (W, x)$ and $(G_m, x_m) \sim (W, x)$ + GNNs $\Phi_{\mathcal{H}}(G_n, x_n)$ and $\Phi_{\mathcal{H}}(G_m, x_m)$ with the same set of weights $\mathcal{H}$; under assumptions AS3, AS4, AS5:

$$\|y_n - y_m\| \leq 2DF^{D-1}\left(L + \frac{\pi n_c}{\min_{n,m}(\delta_c)}\right) \max_{j=n,m} \|W - W_j\|$$

$$+ 2(L_c + 2) \max_{j=n,m} \|x - x_j\| + 4DF^{D-1} l_c$$

where D is depth and F the width of the GNN.

↳ The proof follows easily from the proof of transferability of graph convolutions.

- Note that if we have:

$$y = \sigma(T_H x) \quad \& \quad y_n = \sigma(T_H^n x_n),$$

$$\|y - y_n\| \leq L_\sigma \cdot \|T_H x - T_H^n x_n\|$$

L_σ: 1 by Ass 5

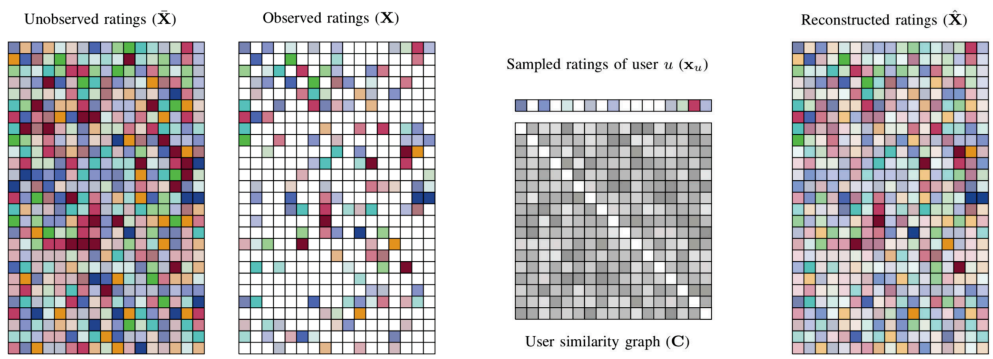
$\|T_H x - T_H^n x_n\|$: ↳ transf. of convolution

- D and F show up by recursively bounding $\|x_l - x_{ln}\|$ in terms of $\|x_{l-1} - x_{l-1,n}\|$ for $1 \leq l \leq D$.

E.g.: Transferability in recommender systems

M movies

U users



→ Based on the observed rating matrix X, we can build either a user or movie similarity graph from the users'/movies' correlations

→ In the case where the graph connects similar movies (i.e., with similar ratings):

$$\sigma_{m_1 m_2} = \frac{1}{|\mathcal{U}_{m_1 m_2}|} \sum_{u \in \mathcal{U}_{m_1 m_2}} (x_{u m_1} - \mu_{m_1 m_2})(x_{u m_2} - \mu_{m_2 m_1})$$

where m_1, m_2 are movies, $\mathcal{U}_{m_1 m_2}$ is the set users having rated m_1 & m_2, and μ_{ij} is the rating rating of movie i by $u \in \mathcal{U}_{ij}$

- The graph is defined as:

$$[A]_{m_1,m_2} = \frac{\sigma_{m_1,m_2}}{\sqrt{\sigma_{m_1,m_1}\,\sigma_{m_2,m_2}}}$$

The signals are defined as:

$$x_u \in \mathbb{R}^M \qquad 1 \le u \le U$$

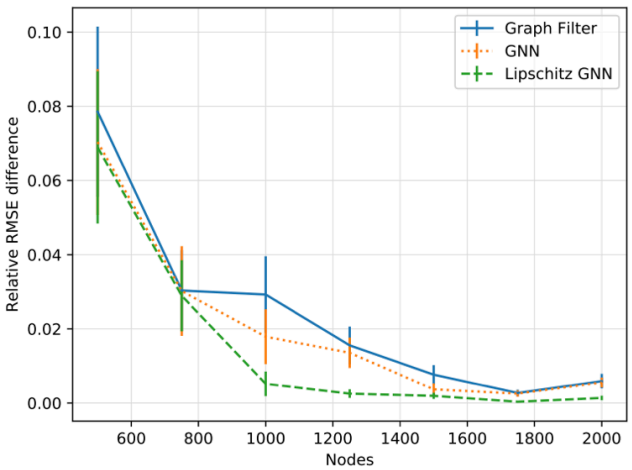
$$[x_u]_m = x_{um} \quad \text{if available} \quad (0 \text{ o.w.})$$

- We hold out some of the existing ratings as labels; i.e., for $u \in \mathcal{U}'$ & $m \in \mathcal{M}'$, we set $(x_u)_m = 0$ and $(y_u)_m = x_{um}$. We then compute the loss over $\mathcal{U}'$ & $\mathcal{M}'$

- Then, we run a transferability experiment (like in HW3): train on G_n, transfer to G_m, $m > n$, while varying n and computing:

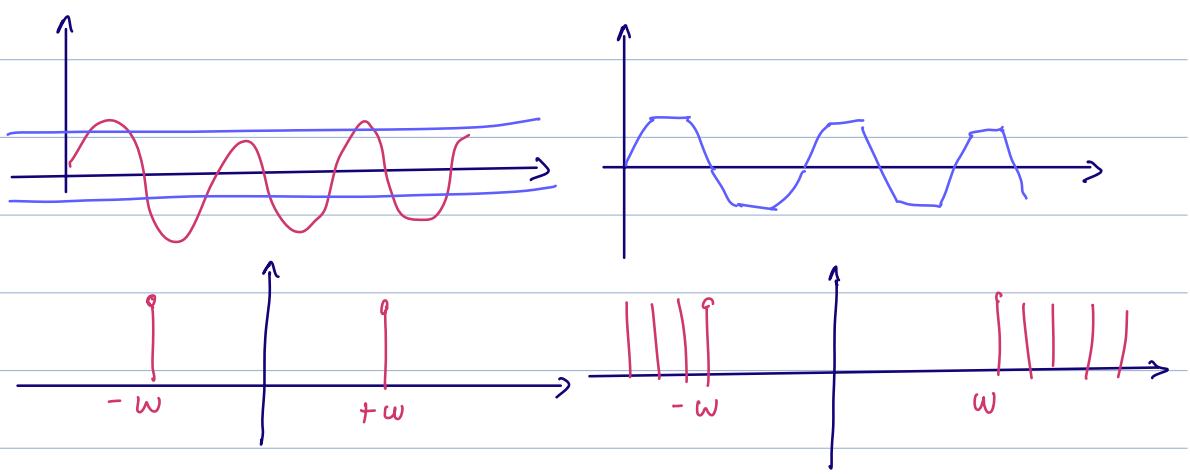
$$\underline{\mathcal{E}} = \frac{\text{error}_{G_n} - \text{error}_{G_m}}{\text{error}_{G_n}}$$

e



If the transferability error of GNNs is essentially the same as that of graph convolutions (modulo scaling by D, F), why are they more transferable?

↳ For the same reason they are more stable, nonlinearities scatter energy $\sum_{i \notin \mathcal{E}} \| x_{\ell-1,i} \|^2$ to $\sum_{i \in \mathcal{E}} \| x_{\ell,i} \|^2$, where components can be discriminated



↳ for the same level of transferability, better discriminability, and vice-versa