Lecture 13 EN.553.744 Prof. Luana Ruiz

Today: finish stability of GNNs

∘) Integral Lipschitz filters: $\exists\, C$ s.t.

$$|h(\lambda) - h(\lambda')| \leq C \frac{|\lambda' - \lambda|}{|\lambda + \lambda'|/2} \quad \forall \lambda, \lambda'$$

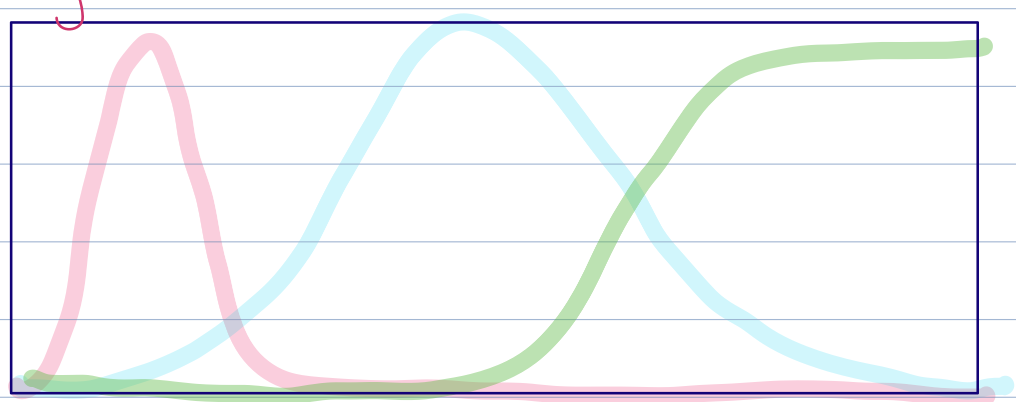
(within an interval)

Lipschitz with a constant inversely proportional to the interval's midpoint
↳ $2C/|\lambda + \lambda'|$

Letting $\lambda' \to \lambda$, we get $\lambda h'(\lambda) \leq C$

the filter can't change for large λ $\Leftarrow$ $h'(\lambda) \leq \frac{C}{\lambda} \to 0$ as $\lambda \to \infty$

E.g.:



At medium frequencies, $\approx$ Lipschitz
At low frequencies, arbitrarily thin
At high frequencies, flat $\Rightarrow$ lose discriminability

C controls discriminability at medium freqs;
but does not affect it close to 0 (arb. thin no matter C)
or for large λ (flat no matter C)

Onto the proof – integral Lipschitz filters are stable to scalings / dilations

Pf.: $$H(s') - H(s) = \sum_{k=0}^{\infty} h_k s'^k - \sum_{k=0}^{\infty} h_k s^k$$

since $s' = (1+\varepsilon)s$,

$$\Rightarrow H(s) - H(s') = \sum_{k=0}^{\infty} h_k \left[(1+\varepsilon)^k s^k - s^k \right]$$

binomial expansion:
$$(1+\varepsilon)^k = \sum_{i=0}^{\infty} \binom{k}{i} \varepsilon^k = 1 + k\varepsilon + O_k(\varepsilon)$$

$$\Rightarrow H(S) - H(S') = \sum_{k=0}^{\infty} h_k \left[(\cancel{1} + k\varepsilon) S^k - \cancel{S^k} \right] + O(\varepsilon)$$

$O(\varepsilon)$ satisfies $0 < \lim_{\varepsilon \to 0} \frac{\| O(\varepsilon) \|}{\varepsilon^2} < \infty$

because filter is analytic $O(\varepsilon)$ of order $\Theta(\varepsilon^2)$ big-oh

$$\Rightarrow H(S) - H(S') = \sum_{k=0}^{\infty} h_k k \varepsilon S^k + O(\varepsilon)$$

Since $\| H(S) - H(S') \| = \max_{\|x\|=1} \frac{\| (H(S) - H(S'))x \|}{\|x\|}$,

we'll prove $\| (H(S) - H(S'))x \| \leq C\varepsilon \quad \forall x : \|x\| = 1$

Recall the iGFT of x: $x = \sum_{i=1}^{n} [\hat{x}]_i v_i$

$$\Rightarrow (H(S) - H(S'))x = \sum_{k=0}^{\infty} h_k k \varepsilon S^k \sum_{i=1}^{n} [\hat{x}]_i v_i$$

$$= \varepsilon \sum_{k=0}^{\infty} h_k k \sum_{i=1}^{n} \lambda_i^k v_i [\hat{x}]_i$$

$$= \varepsilon \sum_{i=1}^{n} \sum_{k=0}^{\infty} k h_k \lambda_i^{k} [\hat{x}]_i v_i$$

$$\lambda_i h'(\lambda_i) \rightarrow |\lambda_i h'(\lambda_i)| \leq C$$

$$\Rightarrow \|(H(S) - H(S'))x\|^2 = \varepsilon^2 \sum_{i=1}^{n} \hat{x}_i^2 \left(\lambda_i h'(\lambda_i)\right)^2$$

$$\leq \varepsilon^2 C^2$$

$$\Rightarrow \|(H(S) - H(S'))x\| < C.\varepsilon \quad \blacksquare$$

universal for all graphs of any size; property of the graph convolution independent of the underlying graph

We can control C → design stable filters (low C), or learn them while penalizing for large C

Yet, regarless of C, integral Lipschitz filter becomes non-discriminative for large λ; in convolutions, it's the price we pay for stability to dilations

2) Additive perturbations

$$\tilde{S} = S + E \Rightarrow E = \tilde{S} - S, \quad 0 < \|E\| \leq \mathcal{E}$$

↳ problematic, as $\tilde{S} = P^T S P$ might lead to $\|E\| > 0$

Additive perturbations modulo permutations:

Define $\mathcal{E}(S, \tilde{S}) = \{E : P^T \tilde{S} P = S + E, \ P \in \mathcal{P}\}$

Given $\tilde{S}$, error is E w/ the smallest norm,

$$\tilde{E} = \underset{E \in \mathcal{E}(S,\tilde{S})}{\operatorname{argmin}} \|E\|, \text{ i.e.,}$$

$$\|\tilde{E}\| = d(S, \tilde{S})$$

operator distance modulo permutations

Graph convolutions are stable to additive perturbations - provided they have Lipschitz spectral response

(Thm) Let $P^T \tilde{S} P = S + \tilde{E}$ with $\|\tilde{E}\| = \epsilon$ and assume h is Lipschitz with constant C.

$$\boxed{\text{Then, } \|H(\tilde{S}) - H(S)\|_p \leqslant C(1 + \delta\sqrt{n})\,\epsilon + O(\epsilon^2)}$$

where δ is the "eigenvector misalignment" between S & $\tilde{E}$.

(Def.) Eigenvector misalignment δ :

Let $S = V \Lambda V^T$ and $E = U M U^T$. Then,

$$\boxed{\delta = (\|U - V\| + 1)^2 - 1}$$

Since $\|U\| = \|V\| = 1$, $\delta \leq 8$

Back to the theorem:

$$\boxed{\| H(\tilde{S}) - H(S) \|_p \leq C(1+\delta\sqrt{n})\varepsilon + \mathcal{O}(\varepsilon^2)}$$

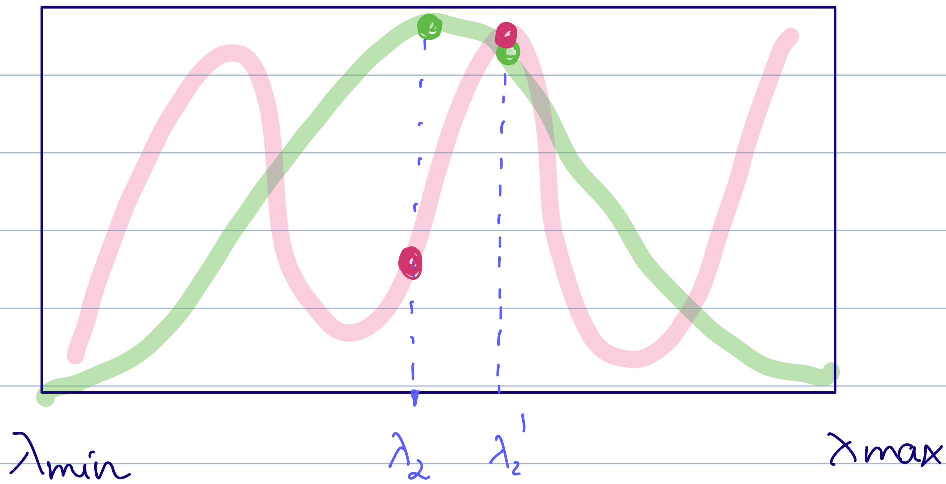
$\hookrightarrow$ Lipschitz stability to absolute perturbations with constant $C(1+\delta\sqrt{n}) \leq C(1+\delta\sqrt{n})$

a) not bad for small n, but terrible for large graphs (unless $\delta \downarrow$ as $n \uparrow$)

b) universal for all graphs of size n; property of the graph convolution independent of the underlying graph

c) We can control C $\rightarrow$ design stable filters (low C), or learn them while penalizing for large C

d) Stability-discriminability tradeoff:



$\uparrow C$, higher discriminability, lower stability

$\downarrow C$, vice-versa

Pf. Exercise. Check Gama et al., 2019

3) Relative perturbations

Unlike dilations, absolute perturbations do not take the graph's edge weights into account
$\hookrightarrow$ not meaningful

Meaningful perturbations are a combination of the previous two:

Relative perturbations modulo permutations:

symmetric

$$\mathcal{E}(S,\tilde{S}) = \{E : P^T\tilde{S}P = S + ES + SE, \; P \in \mathcal{P}\}$$

Given $\tilde{S}$, error is E w/ the smallest norm,

$$\tilde{E} = \underset{E \in \mathcal{E}(S,\tilde{S})}{\operatorname{argmin}} \|E\|, \text{ i.e.,}$$

$$\|\tilde{E}\| = d(S,\tilde{S})$$

operator distance modulo permutations

↳ relative measure of how far $\tilde{S}$ is from being a permutation of S

Locally, we have:

$$(P^T\tilde{S}P)_{ij} = S_{ij} + (\tilde{E}S)_{ij} + (S\tilde{E})_{ij}$$

$$= S_{ij} + \sum_{k \in N(j)} \tilde{E}_{ik} S_{kj} + \sum_{k \in N(i)} S_{ik} \tilde{E}_{kj}$$

$\Rightarrow$ edge changes are proportional to the local structure of the graph (degrees)

(Thm) Let $P^T \tilde{S} P = S + \tilde{E} S + S \tilde{E}$ with $\|\tilde{E}\| = \varepsilon$ and assume h is integral Lipschitz w/ constant C.

$$\boxed{\text{Then, } \|H(\tilde{S}) - H(S)\|_{\mathcal{P}} \leq 2C\,(1 + \delta\sqrt{n})\,\varepsilon + \mathcal{O}(\varepsilon^2)}$$

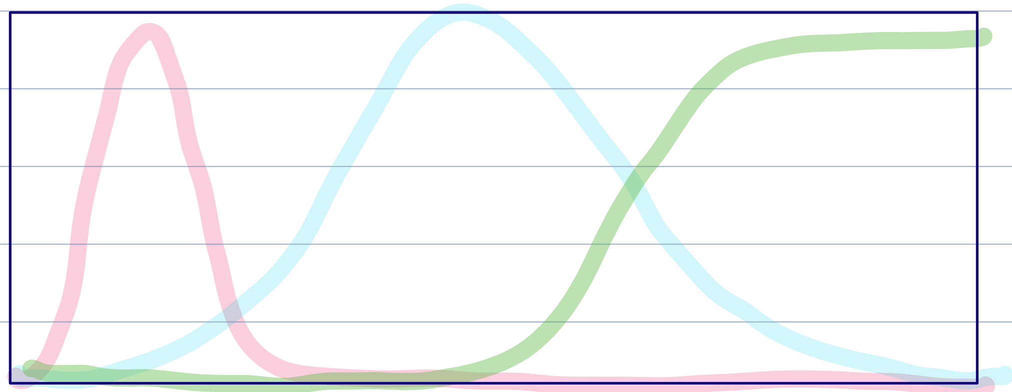
where δ is the "eigenvector misalignment" between S & $\tilde{E}$.

$\hookrightarrow$ Lipschitz stability to relative perturbations w/ constant $2C(1 + \delta\sqrt{n})$

save for a factor of 2, bound is the same as for additive perturbations

$\hookrightarrow$ same comments as before apply ((a) − (c)), except for (d)

For high frequencies λ, there is no stability-discriminability tradeoff → filter is always flat & not discriminative, regardless of C



Pf.: Check Gama et al., 2019

▶ What about GNNs?

GNNs inherit the stability properties of their convolutions.

(Thm). Let $\phi(S, h)$ be an $\overset{L\text{-layer}}{\text{GNN}}$ and $\tilde{S}$ a graph perturbation (modulo permutations)

1) If $\tilde{S} = S + \varepsilon S$ and all h integral Lipschitz,

$$\|\phi(S, h) - \phi(\tilde{S}, h)\|_p \leq L C \varepsilon + O(\varepsilon^2)$$

2) If $P^T \tilde{S} P = S + \tilde{E}$ and all h Lipschitz,

$$\|\phi(S, h) - \phi(\tilde{S}, h)\|_p \leq L\, C\, (1 + 8\sqrt{n})\, \varepsilon + O(\varepsilon^2)$$

3) If $P^T \tilde{S} P = S + \tilde{E}S + S\tilde{E}$ and all h integral Lipschitz,

$$\|\phi(S, h) - \phi(\tilde{S}, h)\|_p \leq 2L\, C\, (1 + 8\sqrt{n})\, \varepsilon + O(\varepsilon^2)$$

↳ only difference is number of layers L

Pf. Non-restrictive assumptions:

1) $\|x_\ell\| \leq 1$ $\forall \ell$: normalized input at all layers (easy to achieve w/ non-amplifying h, $\|H\| = 1$)

2) σ normalized Lipschitz (Lipschitz constant 1, satisfied for most NL – ReLU, sigmoid, etc.)

Let $\|\tilde{E}\| = \varepsilon$ (regardless of perturbation type), and let filters h be stable to perturbation $\tilde{E}$ with

$$\|H(\tilde{S}) - H(S)\|_p \leq C_h \cdot \varepsilon$$

Layer ℓ is a perceptron with filter H_ℓ:

$$\|\tilde{x}_\ell - x_\ell\| \overset{(*)}{=} \|\sigma(H_\ell(\tilde{S})\tilde{x}_{\ell-1}) - \sigma(H_\ell(S)x_{\ell-1})\|$$

σ normalized Lipschitz

$$\leq \|H(\tilde{S})\tilde{x}_{\ell-1} - H_\ell(S)x_{\ell-1}\|$$

↳ add and subtract $H_\ell(\tilde{S})x_{\ell-1}$:

$$\|H_\ell(\tilde{S})\tilde{x}_{\ell-1} - H_\ell(\tilde{S})x_{\ell-1} + H_\ell(\tilde{S})x_{\ell-1} - H_\ell(S)x_{\ell-1}\|$$

$$(**) \quad \leq \cancelto{1}{\|H_\ell(\tilde{S})\|}\underbrace{\|\tilde{x}_{\ell-1} - x_{\ell-1}\|}_{\text{same as } (*)} + \cancelto{\leq 1}{\|x_{\ell-1}\|}\underbrace{\|H_\ell(\tilde{S}) - H_\ell(S)\|}_{\leq C_n\varepsilon}$$

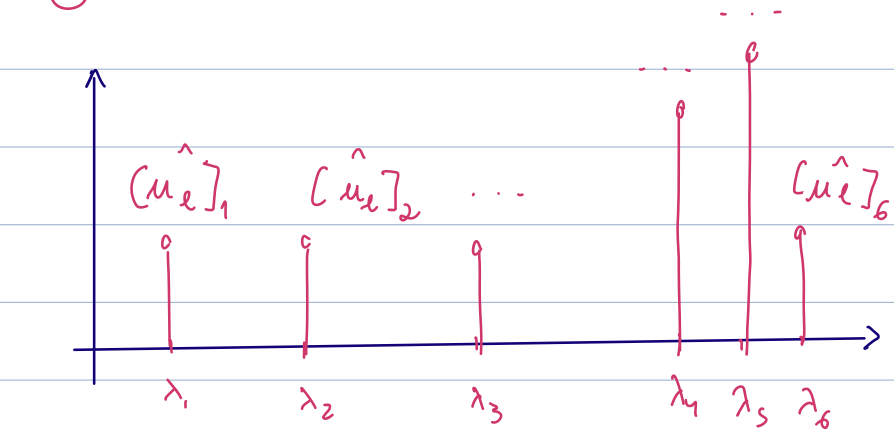
Apply $(**)$ recursively to $\|\tilde{x}_{\ell-1} - x_{\ell-1}\|$, $\|\tilde{x}_{\ell-2} - x_{\ell-2}\|$, ... $\Rightarrow$ L factor appears

↳ same comments as for the respective filter types apply.

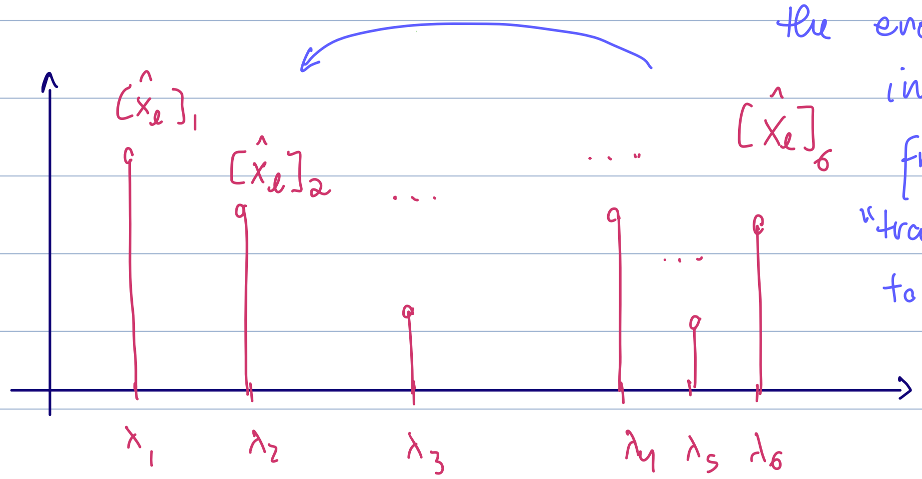
However: while in the node domain the nonlinearity has little effect (normalized Lipschitz), in the spectral domain it is key:

Nonlinearities have the effect of scattering the signal energy across the spectrum

E.g.:



$\downarrow \sigma$



part of the energy in high freqs. "travels" to low freqs

both Lips. & integral Lips. can discriminate

⇒ GNNs are more stable (for same level of discriminability) than convs. (linear GNNs)

⇒ GNNs are more discriminative (for same level of stability) than convs.