Lecture 20 EN.553.744 Luana Ruiz

Today:
- a note on transferability rates
- learning by transference – leveraging transferability in ERM

(THM) Given $(G_n, x_n) \sim (W, x)$ and convolutions

$$H(G_n) = \sum_{k=0}^{\infty} h_k \left(\frac{A_n}{n}\right)^k \quad \text{and} \quad T_H = \sum_{k=0}^{\infty} h_k T_W^{(k)},$$ under

assumptions 3 & 4, we have:

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

$$+ (L_c + 2) \|x - x_n\| + 2l_c$$

↳ Actual transferability rates (in n) depend on $\|T_W - T_{W_n}\|$ and $\|x - x_n\|$

- For $\|T_W - T_{W_n}\|$, we know that the relation btw the operator norm and the cut norm,

combined with the graphon sampling lemma, give us the bound:

$$\|T_W - T_{W_n}\| \leq \sqrt{8\|W - W_n\|_\square} \leq \sqrt{\frac{8.22}{\sqrt{\log n}}}$$

$$= \frac{4\sqrt{11}}{\sqrt[4]{\log n}}$$

- For $\|x - x_n\|_2$, we have $\|x - x_n\|_2 \leq \|x - x_n\|_1$

Further, Levie et al. (2023) defined a cut distance for graphon signals (akin to the graphon cut distance), which is equivalent to the L_1 norm in the following sense:

$$\|x\|_\square \leq \|x\|_1 \leq 2\|x\|_\square$$

Leveraging these together with the graphon signal sampling lemma (Levie et al., 2023), we get:

$$\|x - x_n\|_2 \leq \frac{2.15}{\sqrt{\log n}} = \frac{30}{\sqrt{\log n}}$$

$\Rightarrow$ transferability rate of order $\mathcal{O}\left((\log n)^{-\frac{1}{2}}\right)$

↳ Though applicable to general (w, x), these rates are extremely slow.

↳ We can obtain better rates by making stricter assumptions on the graphon & graphon signal

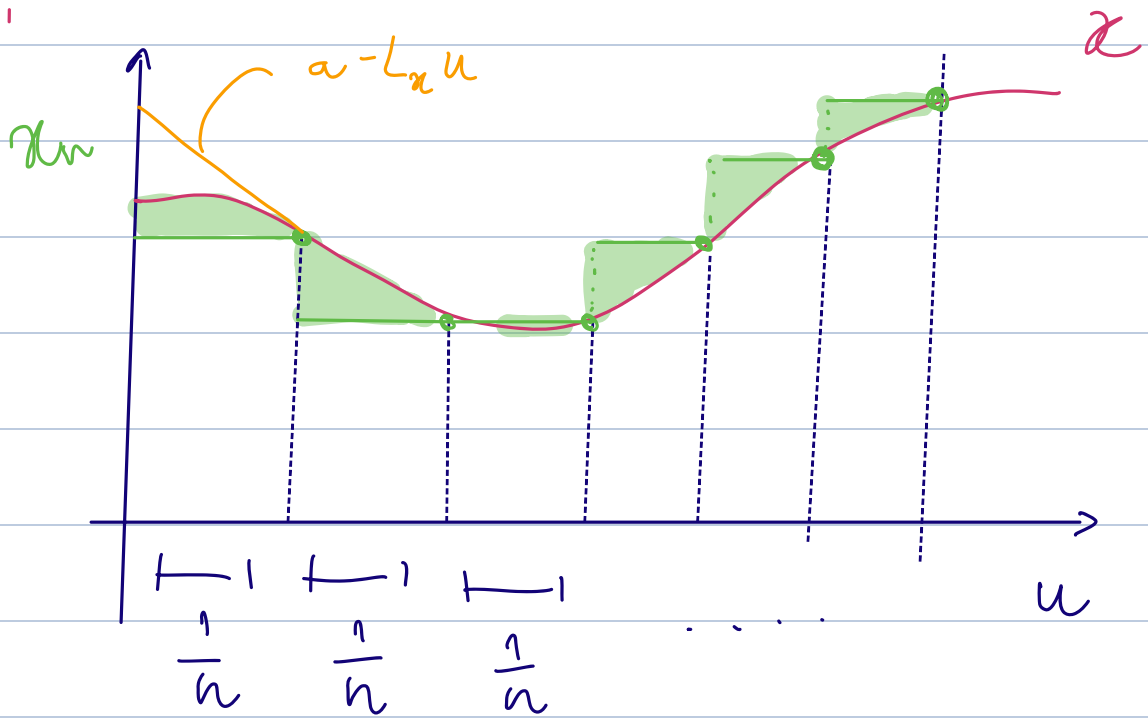
- For L_w-Lipschitz W & L_x-Lipschitz x, we can get

$$\|T_{w_n} - T_w\| \leq \|W - W_n\|_{HS} \leq \frac{L_w}{n}$$

$$\|x - x_n\| \leq \frac{L_x}{n}$$

(For deterministic G_n, x_n)

E.g.:



$$\|x - x_n\|^2 = \int_0^1 (x(u) - x_n(u))^2 \, du$$

$$= \sum_{I_i} \int_{I_i} (x(u) - x_n(u))^2 \, du$$

$$\leq n \cdot \frac{1}{n} \left(\frac{L}{n}\right)^2$$

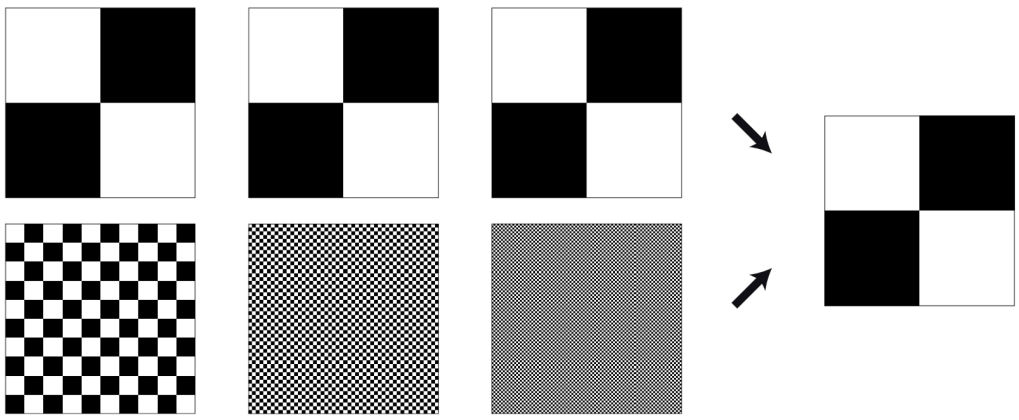
$$\Rightarrow \|x - x_n\| \leq \frac{L_x}{n}$$

(similar example for graphon)

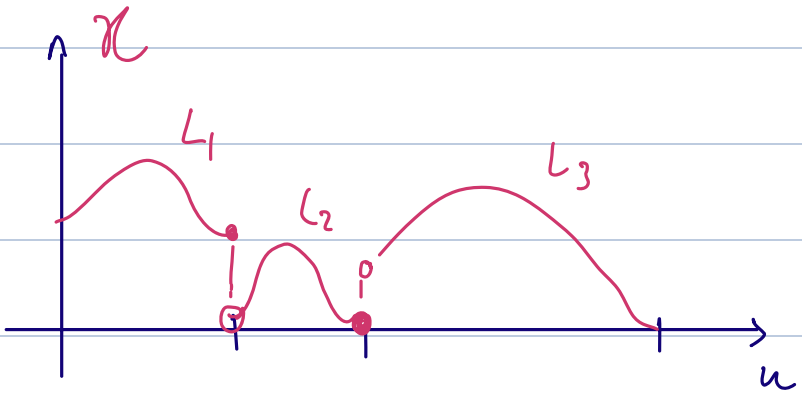
$\Rightarrow$ much better transferability rate, of order $O\left(\frac{1}{n}\right)$

↳ check Ruiz, Chamon, Ribeiro (2023) for stochastic.

However, Lipschitz assumptions on W and $\mathcal{X}$ can be problematic because we lose exchangeability. Conversely, the Lipschitz property is not preserved under measure preserving bijections (not "invariant to permutations")



↳ One compromise is to consider piecewise continuous $W, \mathcal{X}$.



M blocks, $M > 1$

$\Rightarrow$ In this case, we get rates of order $\mathcal{O}\left(\sqrt{\frac{M}{n}}\right)$

(for node-deterministic $G_{m \times n}$)

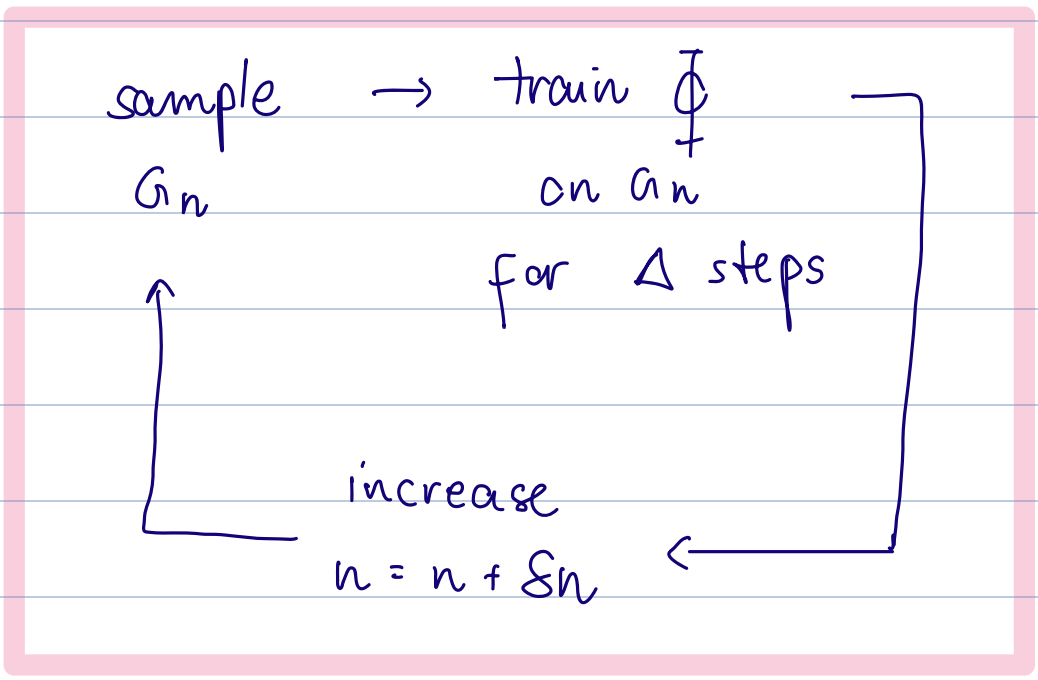
Check Avella-Medina et al. (2018) for stochastic.

→ learning by transference

Instead of:

sampling G_n → training Φ on G_n → running Φ on G_m, $m > n$

We do:



→ run Φ on G_m $m > \max(n)$

- We do this to leverage graph convergence in the training process of the GNN.

- Helpful because, ideally, we would like to solve the ~~ERM~~ problem on the graphon:

$$\min_{\mathcal{H}} \sum_{i=1}^{|\mathcal{D}|} \ell\left(y^{(i)}, \Phi_{\mathcal{H}}(x^{(i)}, W)\right)$$

with gradient updates:

$$\mathcal{H}_{k+1} = \mathcal{H}_k - \eta_k \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \nabla_{\mathcal{H}} \ell\left(y^{(i)}, \Phi_{\mathcal{H}}(x^{(i)}, W)\right)$$

η_k: step size

- But in practice can only compute gradients on graphs:

$$\mathcal{H}_{k+1} = \mathcal{H}_k - \eta_k \frac{1}{|\mathcal{D}|} \sum_{i=1}^{|\mathcal{D}|} \nabla_{\mathcal{H}} \ell\left(y_n^{(i)}, \Phi_{\mathcal{H}}(x_n^{(i)}, G_n)\right)$$

Still, if the graph gradients

$$\nabla_{\mathcal{H}} \ell\left(y_n^{(i)}, \Phi_{\mathrm{GE}}(x_n^{(i)}, G_n)\right)$$

converge to the graphon gradients

$$\nabla_{\mathcal{H}} \ell\left(y^{(i)}, \Phi_{\mathrm{GE}}(x^{(i)}, W)\right)$$

we can hope to learn $\mathcal{H}$'s that are stationary points of the graphon ERM.

(THM) Let $(G_n, x_n) \to (W, x)$ in the usual sense. Then, under the same assumptions needed for GNN transferability, and assuming the loss ℓ and its gradients are Lipschitz, and $\ell \geq 0$, we have:

$$\mathbb{E}\Big(\| \nabla_{\mathcal{H}} \ell\left(y_n^{(i)}, \Phi_{\mathrm{GE}}(x_n^{(i)}, G_n)\right) - \nabla_{\mathcal{H}} \ell\left(y^{(i)}, \Phi_{\mathrm{GE}}(x^{(i)}, W)\right) \| \Big)$$

$$= \leq C_1 \cdot \epsilon + C_2 \cdot \| W - W_n \|$$

where C_1 & C_2 are constants depending on D, F, the Lipschitz constants, and the graphon spectrum.

I.e., the graph gradients converge to the graphon gradient.

The proof follows easily (modulo some algebra from Lipschitz continuity of ℓ and $\nabla\ell$ and transferability of Φ.

- Gradient convergence implies that algorithm (☆) converges to the solution of the graphon ERM:

(THM) Under the previous assumptions, and further assuming that the GNN Φ and $\nabla_{\mathcal{H}}\Phi$ are stable to perturbations of $\mathcal{H}$, we have that, if at each step k Alg. (☆) satisfies:

$$\mathbb{E}\left(\left\| \nabla_{\mathcal{H}_k} \hat{\Phi}(\mathbf{W}_n) - \nabla_{\mathcal{H}_k} \Phi(\mathbf{W}) \right\|\right) \leq \left\| \nabla_{\mathcal{H}_k} \Phi(\mathbf{W}) \right\| + \varepsilon,$$

then under the appropriate step sizes

η_k, Alg. (★) converges to an ϵ-neighborhood of the solution of the graphon ERM in $k^* = \mathcal{O}\left(\frac{1}{\epsilon^2}\right)$ iterations, at most.

Pf.: The proof relies on the following lemma proved in (Cervino, 2023):

(Lemma) If at step k n is such that:

$$\mathbb{E}\left(\left\| \nabla_{g_k} \ell(y, \phi_{g_k}(w)) - \nabla_{g_k} \ell(y_n, \phi_{g_k}(w_n)) \right\|\right) - \left\| \nabla_{g_k} \ell(y, \phi_{g_k}(w)) \right\| \leq \epsilon$$

then under the appropriate step size η_k, the next iterate g_{k+1} satisfies:

$$\mathbb{E}\left(\ell(y, \phi_{g_{k+1}}(w))\right] < \ell(y, \phi_{g_k}(w)) - \frac{\eta_k}{2}\,(\delta\nabla\ell) \qquad (\leq 0)$$

where $\delta\nabla\ell = \left\| \nabla_{g_k} \ell(y, \phi_{g_k}(w)) \right\|^2$

$$- \mathbb{E}\left(\left\| \nabla_{g_k} \ell(y, \phi_{g_k}(w)) - \nabla_{g_k} \ell(y_n, \phi_{g_k}(w_n)) \right\|^2\right) \geq \epsilon^2$$

Proof of theorem:

For every $\epsilon > 0$, we define the stopping time:

$$k^* = \min_{k \geq 0} \left\{ \mathbb{E}\left(\| \nabla_{\mathcal{H}_k} \ell(Y, \phi_{\mathcal{H}_k}(w)) \| \right) \leq C_1 \epsilon + \epsilon \right\}$$

We have:

$$\mathbb{E}[\ell(Y, \Phi_{\mathcal{H}_0}(w) - \ell(Y, \Phi_{\mathcal{H}^*}(w)) \mid k^*]$$

$$\mathbb{E}\left[\sum_{k=1}^{k^*} \ell(Y, \Phi_{\mathcal{H}_{k-1}}(w)) - \ell(Y, \Phi_{\mathcal{H}_k}(w)) \mid k^* \right]$$

$$\Rightarrow \mathbb{E}[\ell(Y, \Phi_{\mathcal{H}_0}(w) - \ell(Y, \Phi_{\mathcal{H}^*}(w))]$$

$$= \sum_{t=0}^{\infty} \mathbb{E}\left[\sum_{k=1}^{t} \ell(Y, \Phi_{\mathcal{H}_{k-1}}(w)) - \ell(Y, \Phi_{\mathcal{H}_k}(w)) \right] P(k^* = t)$$

From Lemma 1, for any $k \leq k^*$, we have:

$$\mathbb{E}\left[\sum_{k=1}^{t} \ell(Y, \Phi_{\mathcal{H}_{k-1}}(w)) - \ell(Y, \Phi_{\mathcal{H}_k}(w)) \right] \geq \frac{\eta}{2} \epsilon^2$$

And so:

$$\mathbb{E}[\ell(Y, \Phi_{k_0}(W)) - \ell(Y, \Phi_{k^*}(W))] \geq \frac{\eta}{2}\varepsilon^2 \sum_{t=0}^{\infty} t\, P(k^* = t)$$

$$= \frac{\eta}{2}\varepsilon^2 \mathbb{E}(k^*)$$

Since the loss function is non-negative,

$$\frac{\mathbb{E}[\ell(Y, \Phi_{k_0}(W))]}{\frac{\eta}{2}\varepsilon^2} \geq \mathbb{E}(k^*)$$

$$\Rightarrow \quad k^* = \Theta\left(\frac{1}{\varepsilon^2}\right) \quad \blacksquare$$