I Mathematical Privilege for GCNN

The GCNN architecture training works in the following sequence **Forward Propogation** → **Loss Calculation** → **Backward Propogation** → **Weight Update**.The analysis is performed for a single GCNN layer but can be very well extended to multiple layers for analysis easily.The architecture considered for this analysis is a single GCNN layer , a MLP(Multi Layer Perceptron) and a regression layer of FCNN(Fully Connected Neural Network).The training process beneath is described via forward and backward propagation and predicting the ipr value:-

Forward Propagation :-
GCN Layer :- $H^{(1,i)} = \sigma(\hat{A}H^{(0,i)}W)$
GCN Final Layer :- $z^{(i)} = \frac{1}{n_i}\Sigma_{j=1}^{n_i} h_j^{(1,i)}$
Linear layer :- $\hat{y}^{(i)} = z^{(i)}W^{(lin)} + b$
Loss Function :- $L = \frac{1}{N}\Sigma_{i=1}^{N}(\hat{y}^{(i)} - y^{(i)})^2$
In the above $\hat{A}^{(i)}$ is the normalized adjacency matrix.$H^{(0,i)}$ is the initial feature matrix.Further $h_j^{(1,i)} = \sigma(\Sigma_{k=1}^{n_i}\hat{A}_{jk}^{(i)}h_k^{(0,i)}W)$ is the feature vector of node j in graph i and the j^{th} row of the updated feature matrix $H^{(1,i)}$.$W^{(lin)}$ and b are the learnable weights of the linear layer.

Backward Propagation :- Inorder to compute the gradients to update weight matrices the chain rule is applied to propagate the error from output layer back through network layers.The gradient of loss with repect to the output of the linear layer is calculated as :-

$$\frac{\partial L}{\partial \hat{y}^{(i)}} = \frac{2}{N}(\hat{y}^{(i)} - y^{(i)}) \tag{1}$$

The gradients of the linear layer are being calculated and the and it is known that each graph i contributes to the overall loss L.We accumulate gradient contributions from each graph when computing gradient of loss with respect to the weight matrix $W^{(lin)}$.To obtain the gradient of the loss with repsect to the weights $W^{(lin)}$, the chain rule is applied :-
Model Architeure Racapitulation
GCN layers:-

$$H^{(l)} = \sigma\Bigg(\hat{A}H^{(l-1)}W^{(l-1)}\Bigg) \quad ,here \quad l = 1,2,3 \tag{2}$$

Average pooling (readout) after $H^{(3)}$:

$$z = \frac{1}{n}\Sigma_{j=1}^{n} h_j^{(3)} \tag{3}$$

First regression layer(linear transform):

$$a = zW_1 + b_1 \tag{4}$$

Second or final regression layer:

$$\hat{y} = aW_2 + b_2 \tag{5}$$

Loss(MSE):

$$L = \frac{1}{N}\Sigma_{i=1}^{N}\Bigg(\hat{y}^{(i)} - y^{(i)}\Bigg)^2 \tag{6}$$

Where $W_1 \in \mathbf{R}^{k_2 \times d_1}, b_1 \in \mathbf{R}^{d_1}$ and $W_2 \in \mathbf{R}^{d_1 \times 1}, b_2 \in \mathbf{R}$ and $a \in \mathbf{R}^{d_1}, \hat{y} \in \mathbf{R}$ **Forward Pass**

$$z \to a = zW_1 + b_1 \to \hat{y} = aW_2 + b_2 \tag{7}$$

Backward Pass : Gradients for the last regression layer
gradient with respect to $\hat{y}$

$$\frac{\partial L}{\partial \hat{y}^{(i)}} = \Sigma_{i=1}^{N} \frac{2}{N}\left(\hat{y}^{(i)} - y^{(i)}\right)$$

Now considering gradient with respect to W_2 we have the following:-

$$\frac{\partial L}{\partial W_2} = \Sigma_{i=1}^{N} \frac{\partial L}{\partial \hat{y}^{(i)}} . \frac{\partial \hat{y}^{(i)}}{\partial W_2}$$

Since $\hat{y} = aW_2 + b_2$ and W_2 is $d_1 \times 1$ so we have on differentiation :-

$$\frac{\partial \hat{y}^{(i)}}{\partial W_2} = a^{(i)T} \tag{8}$$

hence we have :-

$$\frac{\partial L}{\partial W_2} = \Sigma_{i=1}^{N} \frac{2}{N}\left(\hat{y}^{(i)} - y^{(i)}\right) . a^{(i)T}, \quad with \quad shape = d_1 \times 1 \tag{9}$$

for taking gradient with respect to the second adjusting parameter b_2 we have :-

$$\frac{\partial \hat{y}^{(i)}}{\partial b_2} = 1 \quad and \quad \frac{\partial L}{\partial b_2} = \Sigma_{i=1}^{N} \frac{2}{N}\left(\hat{y}^{(i)} - y^{(i)}\right) \tag{10}$$

Next heading on to taking gradient with respect to First Regression Layer Parameters we need $\frac{\partial L}{\partial a_i}$ to proceed backward and we have :-

$$\frac{\partial L}{\partial a^{(i)}} = \frac{\partial L}{\partial \hat{y}^{(i)}} \frac{\partial \hat{y}^{(i)}}{\partial a^{(i)}} \quad and \quad \frac{\partial \hat{y}^{(i)}}{\partial a^{(i)}} = W_2^T \tag{11}$$

Since $\hat{y} = aW_2 + b_2$ so we have :-

$$\frac{\partial L}{\partial a^{(i)}} = \frac{2}{N}(\hat{y}^{(i)} - y^{(i)}) . W_2^T \quad with \quad the \quad shape \quad 1 \times d_1 \tag{12}$$

Gradient with respect to W_1

$$\frac{\partial L}{\partial W_1} = \Sigma_{i=1}^{N} \frac{\partial L}{\partial a^{(i)}} . \frac{\partial a^{(i)}}{\partial W_1} \quad , \frac{\partial a^{(i)}}{\partial W_1} = z^{(i)T} \tag{13}$$

$$\frac{\partial L}{\partial a^{(i)}} = \delta_a^{(i)} \tag{14}$$

as $\delta_a^{(i)}$ is $1 \times d_1$, $z^{(i)}$ is $1 \times k_2$, so $z^{(i)T}$ is $k_2 \times 1$ so, $z^{(i)T} \delta_a^{(i)}$ is $k_2 \times d_1$, which matches W_1 shape, hence the final expression stands as :-

$$\frac{\partial L}{\partial W_1} = \Sigma_{i=1}^{N} x^{(i)T} \left[\frac{2}{N}(\hat{y}^{(i)}) - y^{(i)} W_2^T \right] \tag{15}$$

Now working on the gradient with respect to b_1:-

$$\frac{\partial a^{(i)}}{\partial b_1} = I_{d_1} \quad , \frac{\partial L}{\partial b_1} = \Sigma_{i=1}^{N} \frac{\partial L}{\partial a^{(i)}} = \Sigma_{i=1}^{N} \frac{2}{N}(\hat{y}^{(i)} - y^{(i)}) W_2^T \tag{16}$$

Now finally working out with respect to the final readout layer $z^{(i)}$ we have :-

$$\frac{\partial L}{\partial z^{(i)}} = \frac{\partial L}{\partial a^{(i)}} \frac{\partial a^{(i)}}{\partial z^{(i)}} \quad , \frac{\partial a^{(i)}}{\partial z^{(i)}} = W_1^T \tag{17}$$

$$\frac{\partial L}{\partial z^{(i)}} = \left[\frac{2}{N}(\hat{y}^{(i)} - y^{(i)})W_2^T\right]W_1^T \tag{18}$$

followed by simplification and reordering :-

$$\frac{\partial L}{\partial z^{(i)}} = \left[\frac{2}{N}(\hat{y}^{(i)} - y^{(i)})\right](W_1 W_2)^T \tag{19}$$

Thus from the above calculation it is evident that the with the increase in the number of regression layers and applying pooling function before passing it through both of the read outlayers increase chances of the framework to deal with mixed topologies and also deal with complex ipr patterns more efficiently. Since the final layer of GCN is utilized for implementing the average function as the pooling function thus includin it in our calculations we have :-

$$\frac{\partial L}{\partial h_j^{(1,i)}} = \frac{\partial L}{\partial z^{(i)}}\frac{\partial z^{(i)}}{\partial h_j^{(1,i)}} = \Sigma_{i=1}^{N}\frac{2}{N}(\hat{y}^{(i)} - y^{(i)})(W_1 W_2)^T\frac{1}{n_i} \tag{20}$$

The total gradient with respect to **W** accumulates the contribution from all nodes in all graphs:-

$$\frac{\partial L}{\partial W} = \Sigma_{i=1}^{N}\left(\frac{\partial L}{\partial h_j^{(1,i)}}\frac{\partial h_j^{(1,i)}}{\partial W}\right) \quad proceeding \quad with := \frac{\partial h_j^{(1,i)}}{\partial W} \tag{21}$$

The layer output for the i^{th} graph is represented as $H^{(1,i)} = \sigma(\hat{A}H^{(0,i)}W)$.Hence for a single node j in the graph i , its node representation after GCN layer is:-

$$h_j^{(1,i)} = \sigma\left(\Sigma_{k=1}^{n_i}\hat{A}_{jk}^{(i)}h_k^{(0,i)}W\right) = \sigma(q_j^{(i)}), \quad here \quad q_j^{(i)} = \Sigma_{k=1}^{n_i}\hat{A}_{jk}^{(i)}h_k^{(0,i)}W \tag{22}$$

In order to compute $\frac{\partial h_h^{(1,i)}}{\partial W}$, we apply the chain rule and the result is as follows :-

$$\frac{\partial h_j^{(1,i)}}{\partial W} = \frac{\partial h_j^{(1,i)}}{\partial q_j^{(i)}}\frac{\partial q_j^{(i)}}{\partial W} \tag{23}$$

Partial derivative can be calculated with respect to $q_j^{(i)}$ as :-

$$\frac{\partial h_j^{(1,i)}}{\partial q_j^{(i)}} = \sigma'(q_j^{(i)}) \tag{24}$$

$q_j^{(i)}$ is a linear combination of the rows $H^{(0,i)}$ weighted by $\hat{A}_j^{(i)}$.In the matrix notation we can write as :-

$$\frac{\partial q_j^{(i)}}{\partial W} = \hat{A}_j^{(i)}H^{(0,i)} \tag{25}$$

where $A_j^{(i)}$ is the j^{th} row of $\hat{A}^{(i)}$.Now, we can combine the results of the chain rule and get:-

$$\frac{\partial h_j^{(1,i)}}{\partial W} = \sigma'(q_j^{(i)})\hat{A}_j^{(i)}H^{(0,i)} \tag{26}$$

Now combining equations (19) and (26) into equation (21) we have as follows :-

$$\frac{\partial L}{\partial W} = \left[\frac{2}{N}(\hat{y}^{(i)} - y^{(i)})\right](W_1 W_2)^T\frac{1}{n_i}.\sigma'(q_j^{(i)})\hat{A}_j^{(i)}H^{(0,i)} \tag{27}$$

Finally it is beign observed that the learnable weight matrices are updated using gradient descent as follows :-

$$W \leftarrow W - \eta\frac{\partial L}{\partial W}, \quad W_{lin}^{(1)} \leftarrow W_{lin}^{(1)} - \eta\frac{\partial L}{W_{lin}^{(1)}}, \quad W_{lin}^{(2)} \leftarrow W_{lin}^{(2)} - \eta\frac{\partial L}{W_{lin}^{(2)}}, \tag{28}$$

$$b_1 \leftarrow b_1 - \eta\frac{\partial L}{\partial b_1}, \quad b_2 \leftarrow b_2 - \eta\frac{\partial L}{\partial b_2} \tag{29}$$

The above process includes the parameter η which is the learning parameter.The above process is repeated iteratively as mentioned earlier which includes a forward propagation followd by loss calculation followed by backward propagation and finally a weight update.The above derivation clearly portrays the mathematical upperhand over other GCNN architectures interms of an increased number of learnable weight matrices that helps acclimatize the architecture to complex network topologies and the gradient descent used in updating the parameters suggest a better convergence.

II Mathematical Superiority of ARMA GNN Vs GCNN

To begin with proving the superiority of ARMA over GCNN revolves around the fact that rational functions have better expressive capacity and shows rapid convergence compared to polynomials.To prove the aforementioned we will be heavily relying on weierstrass approximation theorem to show that continuous real valued functions on a compact interval can be unformly approximated by polynomials.In short words polynomials are uniformly dense in $C([a,b]\mathbf{R})$ with respect to sup-norm(the absolute value of the function in the provided domain). The proof of Stone - Weierstrass theorem requires few preliminaries to be established which go on as follows:-

Definition(1) (Unital Sub-Algebra , Separating Points):- Let K be a compact metric space .Considering the banach algebra beneath:-

$$C(K,\mathbf{R}) := [f : K \to \mathbf{R}| \quad f \quad is \quad continuous] \tag{30}$$

equpped with the sup norm ,

$$||f||_\infty := sup_{t\in K}|f(t)| \tag{31}$$

Then the following two cases arise ,

- $A \subset C(K,\mathbf{R})$ is a unital sub-algebra if $1 \in A$ and if $f, g \in A, \alpha, \beta \in \mathbf{R}$ implies that $\alpha f + \beta g \in A$ and $fg \in A$.
- $A \subset C(K,\mathbf{R})$ separates points of K if for all $s, t \in K$ with $s \neq t$, there exists $f \in A$ such that $f(s) \neq f(t)$.

Generalization can follow for the K which can be redefined from compact metric space to being a compact topological space.
Before moving on to the stone-weistrass theorem a proof has to devised for lattice with significance of it being used when defining classes of function in compact metric space.

Definition(2) (Lattice):-A subset $S \in C(K,\mathbf{R})$ is a lattice if, for all $f, g \in S$, , $f\mathrm{V}g \in S$ and $f \wedge g \in S$, where $(f\mathrm{V}g)$(x) $:= max[f(x), g(x)]$ and $(f \wedge g)(x) := min(f(x), g(x))$, to work on the above lemma the taylor series of the function $\sqrt{(1 - g(t))}$, where $0 \leq g(t) \leq 1$, is investigated.It is useful to first study the taylor series of $\sqrt{1-t}$. Formally,

$$\sqrt{1-t} = 1 - \Sigma_{n=1}^{\infty} a_n t^n \tag{32}$$

where, for $n \in N$,

$$a_n = (-1)^{n-1}\left({}^{\frac{1}{2}}C_n\right) = \frac{(-1)^{n-1}}{n!}\Pi_{k=0}^{n-1}\left(\frac{1}{2} - k\right) = 2^{1-2n}\frac{(2n-2)!}{n!(n-1)!}, \quad n \in N, a_n \geq 0 \tag{33}$$

We observe for the ratio test for convergence that :-

$$lim_{n\to\infty}\left|\frac{a_{n+1}}{a_n}\right| = lim_{n\to\infty}\frac{2n-1}{2(n+1)} = 1 \tag{34}$$

Thus the convergence test proves that it happens in pointwise fashion within the interval $t \in (-1,1)$.Now defining $\psi(t) = 1 - \Sigma_{n=1}^{\infty} a_n t^n$.so $\psi(t)$ converges for $t \in (-1,1)$,Evaluating we obtain :-

$$\psi(t) = -2(1-t)\frac{d\psi}{dt} \implies -\frac{1}{2}\int \frac{dt}{1-t} = \int \frac{d\psi}{\psi} \implies \psi(t) = c\sqrt{1-t}, \quad for \quad some \quad c \in \mathbf{R} \tag{35}$$

Hence evaluating both sides at $t = 0$ gives $c = 1$.Thus, we have $\psi(t) = \sqrt{1-t}$ pointwise for $t \in [0,1)$.

Now we must show that $\psi(1) = 0$.Hence Stirling's inequality states that :-

$$e^{\frac{7}{8}-n} n^{\frac{1}{2}} < n! < e^{1-n} n^{n+\frac{1}{2}}.$$

Hence, for $n \geq 2$,

$$a_n < 2^{1-2n} \frac{1}{e^{\frac{7}{8}} n^{\frac{1}{2}}} \frac{(2n-2)^{2n-\frac{3}{2}}}{(n-1)^{n-\frac{1}{2}}} < \frac{1}{\sqrt{2}e^{\frac{7}{8}}} \frac{(n-1)^{2n-\frac{3}{2}}}{(n-1)^{2n}} = \frac{1}{\sqrt{2}e^{\frac{7}{8}}} \frac{1}{(n-1)^{\frac{3}{2}}}.$$

So,

$$\sum_{n=1}^{\infty} a_n < \frac{1}{2} + \frac{1}{\sqrt{2}e^{\frac{7}{8}}} \sum_{n=2}^{\infty} \frac{1}{(n-1)^{\frac{3}{2}}} < \infty$$

where the comparison test was used in the first inequality and p-series test in the second. Since $\left(\sum_{n=1}^{k} a_n\right)_{k \geq 1}$ is monotonically increasing and bounded above, $\psi(1)$ exists. Then, by **Abel's Theorem**[1],

$$\psi(1) = 1 - \sum_{n=1}^{\infty} a_n = 1 - \lim_{t \to 1^-} \sum_{n=1}^{\infty} a_n t^n = \lim_{t \to 1^-} \sqrt{1-t} = 0.$$

Hence, again using Abel's Theorem, $\psi(t) = \sqrt{1-t}$ uniformly for $t \in [0,1]$.

Lemma Let $A \subset C(\mathbf{K}, \mathbf{R})$ be a closed unital sub-algebra. Then

i) if $f \in A$ and $f \geq 0$, then $\sqrt{f} \in A$;

ii) if $f \in A$, then $|f| \in A$;

iii) A is a lattice.

Proof. To see i), consider without restriction

$$0 \leq f \leq 1.$$

Then, we can write

$$f = 1 - g$$

with

$$0 \leq g \leq 1.$$

Using a Taylor series expansion, we can write, formally,

$$\sqrt{f(t)} = \sqrt{1-g(t)} = 1 - \sum_{n=1}^{\infty} a_n g^n(t)$$

[1] Theorem (Abel). *Let $f(x) := \sum_{n=0}^{\infty} c_n (x-x_0)^n$, and assume that the series converges at $x = x_0 + R$, for some $R \in (0,\infty)$. Then the series is uniformly convergent on $[x_0, x_0+R]$ and*

$$\lim_{x \to (x_0+R)^-} f(x) = f(x_0+R) = \sum_{n=0}^{\infty} c_n R^n.$$

where the coefficients are as in the above remark. The Taylor series approximates $\sqrt{f}$ uniformly in $\| \cdot \|_\infty$. Indeed,

$$\left\| \sqrt{f} - \left(1 - \sum_{n=1}^{N} a_n g^n \right) \right\|_\infty \leq \sup_{t \in K} \left| \sqrt{f(t)} - \left(1 - \sum_{n=1}^{N} a_n g^n(t) \right) \right|$$
$$= \sup_{z \in [0,1]} \left| \sqrt{z} - \left(1 - \sum_{n=1}^{N} a_n z^n \right) \right|.$$

So, given $\epsilon > 0$, by the uniform convergence of the Taylor series of $\sqrt{1-t}$ on $[0, 1]$, there exists $N \in \mathbf{N}$ such that

$$\left\| \sqrt{f} - \left(1 - \sum_{n=1}^{N} a_n g^n \right) \right\|_\infty < \epsilon \quad \forall N \geq \bar{N}.$$

That is,

$$\lim_{N \to \infty} \left\| \sqrt{f} - \left(1 - \sum_{n=1}^{N} a_n g^n \right) \right\|_\infty = 0.$$

Since for all $n \in \mathbf{N} \cup \{0\}$ $g^n \in A$, and A is a sub-algebra, $1 - \sum_{n=1}^{N} a_n g^n \in A$. And, since A is closed by hypothesis, $\sqrt{f} \in A$. This completes the proof of i). To prove ii), note that

$$|f| = \sqrt{f^2}.$$

So for $f \in A$, $f.f = f^2 \in A$, since A is an algebra.Applying i) to f^2 implies $|f| \in A$.To prove iii) , it is to be noted that,

$$f \wedge g = \frac{1}{2}(f + g - |f - g|) \quad and \quad f \vee g = \frac{1}{2}(f + g + |f - g|) \tag{36}$$

Applying ii) to the above identities gives the desired result.
Its high time now I must disclose the proof of the Stone weistrass theorem which justifies the effeciency of the polynomial nature of GCNN and followed by the superiority of ARMA as a rational function to function better than the GCNN

Stone-Weierstrass Theorem (1937).:-
Let K be a compact metric space and $A \subset C(K, \mathbf{R})$ a unital sub-algebra which separates points of K .Then A is dense in $C(K, \mathbf{R})$.
An equivalent statement is that if A is a closed unital subalgebra which separates points of a compact set K, and $A \subset C(K, \mathbf{R})$,then $A = C(K, \mathbf{R})$.We will proceed using this formulation.
Proof. Let $A \subset C(K, \mathbf{R})$ be a closed unital sub-algebra that separates points of K. Let $\epsilon > 0$ be given. For any $f \in C(K, \mathbf{R})$ we will show that there exists $g \in A$ such that

$$\|f - g\|_\infty < \epsilon.$$

Consider points $s, t \in K$. Since A separates points, there exists $h \in A$ such that $h(s) \neq h(t)$. For some $\lambda, \mu \in \mathbf{R}$, define $\tilde{h} : K \to \mathbf{R}$ by

$$\tilde{h}(v) := \mu + (\lambda - \mu) \frac{h(v) - h(t)}{h(s) - h(t)} \quad \forall v \in K.$$

Note that $\tilde{h} \in A$ and $\tilde{h}(s) = \lambda$, $\tilde{h}(t) = \mu$. Thus, for $s \neq t$, there exists $f_{s,t} \in A$ such that

$$f_{s,t}(s) = f(s)$$

and $f_{s,t}(t) = f(t)$.

So, $f_{s,t}$ approximates f in neighbourhoods around s and t. Now, fix s, and let t vary. Put

$$U_t := \{v \in K | f_{s,t}(v) < f(v) + \epsilon\}.$$

U_t is open because it is the pre-image of an open set. Also, $t \in U_t$. So, $\bigcup_{t\in K} U_t$ is clearly an open cover of K. By the compactness of K, there exists finitely many $t_1, \ldots, t_n \in K$ such that

$$K \subset \bigcup_{i=1}^{n} U_{t_i}.$$

Put

$$h_s := \min_{1\leq i\leq n} f_{s,t_i}.$$

Then,

$$\begin{array}{rcl} h_s & \in & A \\ h_s(s) & = & f(s) \\ h_s & < & f + \epsilon. \end{array}$$

Now, define

$$V_s := \{v \in K | h_s(v) > f(v) - \epsilon\}.$$

Note that V_s is open and $K \subset \bigcup_{s\in K} V_s$. By compactness, there exists finitely many $s_1, \ldots, s_m \in K$ such that

$$K \subset \bigcup_{j=1}^{m} V_{s_j}.$$

Put $g = \max_{1\leq j\leq m} h_{s_j}$. Then, $g \in A$ and

$$f - \epsilon < g < f + \epsilon.$$

That is,

$$\|f - g\|_\infty < \epsilon.$$

So, A is dense in $C(K, \mathbf{R})$. And, since A is closed, $A = C(K, \mathbf{R})$. **Corollary:-** Let X be a compact subset of $\mathbf{R}^n$ for some $n \in \mathbf{N}$. Then the algebra of all polynomials $P(X, \mathbf{R})$ in the coordinates $x_1, \ldots, x_n$ is dense in $C(K, \mathbf{R})$.**Remarks** The case in which $n = 1$ in the above corollary is the Weierstrass Approximation Theorem.

The Stone weierstrass theorem states that for rational function they can approximate continous functions more effeciently than polynomials, especially functions with → Sharp Transitions, Poles or Singularities and Rapid decay.A formulation for the same can be assumed to be such as :-
For the same number of parameters (P+Q = K), rational functions can achieve lower approximation error:-

$$inf_{a\in P_K}\Bigg|\Bigg|f - a\Bigg|\Bigg|_\infty \geq inf_{r\in R_{P,Q}}\Bigg|\Bigg|f - r\Bigg|\Bigg|_\infty \quad where \quad P + Q = K \tag{37}$$

Here P_K is the set of polynomials of degree K, and $R_{P,Q}$ is the set of rational functions of order (P,Q).The inequality needs to be proved about a fixed degree K.The need of the hour is to establish a faster convergence rate as $K \to \infty$.

II.1 Asymptotic Convergence Rate

Hence what needs to be checked now is the convergene rates for either of the **Polynomials** comapred to the **Rational Functions**
The best uniform approximation erros are defined as :-

- **Polynomial Error:**$E_K(f) = inf_{a\in P_K}\Bigg|\Bigg|f - a\Bigg|\Bigg|_\infty$
- **Rational Function Error:**$R_N(f) = inf_{r\in R_{N,N}}\Bigg|\Bigg|f - r\Bigg|\Bigg|_\infty$(Here we set P=Q=N so the total degree sums up as K=2N)

The comaprison is for functions f that converge on the interval of approximation that is obey the above theorems.
The proof following is primarily in the field of Approximation Theory and associated with the work of **J.L.Walsh,S.N.Mergelyan**, and later continued by **A.A.Gonchar,E.B.Saff** and **V.K.Totik**Working initially on the polynomial approximation rate we have:-

II.1.1 The Polynomial Approximation Rate(Geometric Convergence)

For a function f that is analytic in a closed interval [a,b] the convergence rate of the best polynomial approximation error $E_K(f)$ is geometric.Let D_ρ be the largest ellipse in the complex plane with foci a and b inside which f is analytic , and let ρ be the sum of its semi-axes(the conformal radius of ellipse).The **Geometric Rate for Polynomials** goes as follows for error $E_K(f)$:-

$$E_K(f) \sim \frac{C_f}{\rho^{K+1}} \quad or \quad more \quad formally \quad \to lim_{K\to\infty} sup[E_K(f)]^{1/K} = \frac{1}{\rho} \tag{38}$$

Where C_f is a constant related to f.This is a geometric convergence rate, determined solely by the distance ρ to the nearest singularity of f in the complex plane. The further the nearest singularity is from the interval, the faster the error decreases.

II.1.2 The Rational Function Approximation(Super Geometric Convergence)

For the same analytic function f, the best rational function approximation error $R_N(f)$ with $P = Q = N$, so the total degree $K = 2N$ can converge much faster.In many cases the best rational approximation error exhibits a double geometric rate of convergence.

$$R_N(f) \sim C'_f.\frac{1}{\rho^{2N+1}} \to lim_{N\to\infty} sup[R_N(f)]^{1/N} = \frac{1}{\rho^2} \tag{39}$$

ρ being the same parameter used for the polynomial converegence.The most advantage arises when the function f has a pole or a strong singularity near the interval.

- **Polynomials** P_K**:**To approximate a function with a singularity z_0, polynomials must use all their coefficients to fight the singularity's influence on the real axis, leading to the geometric rate limited to $\frac{1}{\rho}$.
- **Rational Functions** $R_{N,N}$**:**The denominator polynomial $q(x)$ can be specifically designed to cancel or absorb the effect of the pole at z_0.The optimal rational function $r(x) = \frac{p(x)}{q(x)}$ will have a pole of its own very close to z_0.This "pole cancelling " mechanism frees up to $2N$ parameters to approximate the remaining smooth part of the funciton allowing the overall error to converge much faster.

To demonstrate the much faster convergence , we compare the error for the sam etotal number of parameters , $K = 2N$ and for $\rho > 1$(necessary for convergence), the rational error $R_N(f)$ is determined by the term $(\rho^2)^{-N}$, while the polynomial error $E_K(f)$ is determined by the slower term ρ^{-2N}.

$$\frac{E_{2N}(f)}{R_N(f)} \approx \frac{\frac{C_f}{\rho^{2N}}}{\frac{C_{f'}}{\rho^{2N}}}.\frac{1}{(other \quad term)} \tag{40}$$

The core result, proven by Donald Newman for the function $f(x) = |x|$ and generalized by others, is that for many functions, the rational error converges at the rate of the square of the polynomial rate:

$$lim_{N\to\infty}\frac{E_{2N}(f)}{R_N(f)} = \infty \tag{41}$$

This demonstrates that the rational function error decays at a super-geometric or double-geometric rate compared to the polynomial error, confirming the phrase "converges much faster".

III Mathematical Privilege of GCAT Vs GCNN

From the equation (63) we have the score α_{ij} to be used directly as an entry for the matrix ϕ but to encourage attention sparsity the α_{ij} is passed through **local soft maximum operator** :-

$$\phi_{ij} = \frac{e^{\alpha_{ij}}}{\left(\Sigma_{j' \in N_i \cup i} e^{\alpha_{ij'}}\right)} \tag{42}$$

Since the soft maximum operator assigns a rational function form for the matrix ϕ which is a layer specific matrix and also convolutional in nature for the GCAT and since being different from the shift operator S the operation takes place as follows:-

$$X_l = \sigma\left(\Sigma_{k=0}^{K} \phi^k X_{l-1} A_k\right), here \quad \phi = \phi_l \quad and A_k = A_{lk} \tag{43}$$

Now coming back to the Newman's proof for best uniform polynomial and rational approximation errors for $f(x) = |x|$ on the interval $[-1, 1]$ is
Polynomial Approximation $= E_N(|x|) \sim \frac{C_1}{n}$, where C_1 is a constant.
And the **Rational Approximation** rate being $= R_{n,n}(|x|) \sim C_2 e^{-C_3\sqrt{n}}$, where C_2, C_3 being constants.

A. The Non-Analytic Case ($f(x) = |x|$)

For $|x|$:

- Polynomial Rate: $O(1/n)$ (algebraic)
- Rational Rate: $O(e^{-\pi\sqrt{n}})$ (root-exponential)

The ratio of the errors shows the enormous difference:

$$\lim_{n\to\infty} \frac{E_n(|x|)}{R_{n,n}(|x|)} = \lim_{n\to\infty} \frac{C_1/n}{C_2 e^{-\pi\sqrt{n}}} = \infty$$

The rational error vanishes incredibly fast, while the polynomial error only decays as $1/n$. The polynomial's failure is due to the **singularity at the origin** ($|x|$ is not differentiable at $x = 0$). A rational function uses the poles of its denominator, clustered exponentially near the singularity, to absorb the non-smoothness.

B. The Analytic Case (The $\sim 1/\rho^{2N}$ Analogy)

The "rate-squared" analogy is **literally true for analytic functions**, which is the more generalized result:

- Polynomial Rate: $E_K(f) \sim \frac{C_E}{\rho^K}$
- Rational Rate: $R_{N,N}(f) \sim \frac{C_R}{\rho^{2N}}$

If we compare them using the same total parameter count, $K \approx 2N$:

$$R_{N,N}(f) \approx \frac{C_R}{\rho^{2N}} = \frac{C_R}{(\rho^N)^2}$$

If the polynomial error for degree N is $E_N(f) \sim 1/\rho^N$, then the rational error $R_{N,N}(f)$ is proportional to the **square of the polynomial error of half its degree**. This is where the analogy originates and why rational approximation is considered "double-geometric."Hence the entries for the matrix ϕ being in rational format depict a more rapid convergence compared to the GCNN hence is superior to it.