Geometric Deep Learning

Richard Oravkin

April 2023

1 Introduction

One reason why I find Geometric Deep Learning so fascinating, is that it connects ideas from various areas of science (especially mathematics) and presents the studied problems in a different light. This often results in a deeper and more intuitive understanding of the problem. Inspired by this, I attempted to do the same. I searched for a paper which intrigued me, carefully studied the findings of the paper and then attempted to create theoretical foundations of the problem using fundamental theorems in mathematics. The work presented in this report is not supposed to "improve" the findings of the paper but merely show them in a new light, one which I pesonally found more intuitive. In fact, this work does not improve the findings of the paper - we will encounter situations where our method is inferior. Overall, what is discussed in this report should be thought of as a first step.

The paper that I decided to analyze is the recent work by members of our department about the zero one laws of graph neural networks. The reason why I chose this is that I like probability theory and I have always been fascinated by Kolmogorov's 0-1 law[1]. I was not aware of any such results in the field of graph neural networks and this paper sounded very original, so it was the obvious choice. Additionally, the paper has important implications about the expressive power of graph neural networks, which is an extremely important problem in geometric deep learning, as was often demonstrated in the course. In this report, I will not discuss these implications, even though I would have liked to, because of the page limit. I leave this task to the reader.

Let me now explicitly define the goals of this report:

1. Understand and explain the findings of the paper both on a technical and an intuitive level.
2. Propose a novel approach for proving the results stated in the paper and analyze the benefits and shortcomings of this approach.
3. Analyze the assumptions of the theorems in the paper and see if they can be modified.
4. Discuss problems the paper does not consider using the new method. If possible, verify them experimentally.

Before we begin, some remarks. I include an appendix with some mathematical results which will be used repeatedly. Theorems which are not in the appendix, are my own, meaning that I proposed and proved the theorems and did not copy the proofs from any textbooks. This is why I include the proofs in the main part of the document and not the appendix. I spent a significant amount of time coming up with the proofs and I don't include the many failed attempts, so please consider this when marking. Also, I exceed the page limit quite significantly, but this is mainly because I restate theorems and definitions from the paper (for the reader's convenience) and my report contains a lot of maths which significantly inflates the page limit, because of the formatting. So in fact, the amount of relevant information in this report is around 6 pages, but I wanted to make it easy to read in case the person marking this is not familiar with the discussed paper.

[1] https://en.wikipedia.org/wiki/Kolmogorov%27s_zero–one_law

2 Zero One Laws of Graph Neural Networks

The general overview of the problem considered in the paper is the following: You assign labels $l \in \{0, 1\}$ to graphs using a GNN model (e.g a GCN) , and you are interested in the asymptotic behaviour, as the number of nodes in the graph increases. More precisely, you fix the parameters of your model (so the model is NOT learning) and you sample graphs G_n, of increasing size n, from the Erdos-Renyi model $\mathbb{G}(n, r)$. For each node in the graph, you sample the node features from a distribution D with mean μ. The features are i.i.d and the distribution D does not change throughout the problem. You feed these node features into your model to obtain a graph label $l \in \{0, 1\}$. Now you ask the question: What is the probability that G_n is classified as 0 (or 1), as n tends to infinity? What the authors show is that under certain conditions, this probability tends to zero or one. In my opinion, this sounds like magic. Let's demystify the result.

First, let us state one of the theorems from the paper and give an outline of the proof:

Theorem 1. *Let M be a MEANGNN used for binary graph classification and take $r \in [0, 1]$. Then, M satisfies a zero-one law with respect to $G(n, r)$ and D with mean μ assuming the following conditions hold:*

1. *D is sub-Gaussian.*
2. *σ is Lipschitz continuous.*
3. *The graph-level representation uses average pooling*
4. *The classifier is non-splitting.*

On a first glance, this looks daunting. To understand what is going on, we must look at the proof.

The proof goes as follows: Let M be a MEANGNN with the first layer preactivations[2] given by:

$$y_v^1 = \frac{1}{N(v)} W_1 \sum_{u \in N(v)} x_u + \frac{1}{n} W_2 \sum_{u \in V} x_u + b$$

By using a concentration bound on the x_u, which are sampled from a sub-gaussian distribution, and a concentration bound on the size of the neighborhood, they obtain a bound on the probability that y_v^1 is further than ϵ from its mean $(W_1 + W_2)\mu$. Using this bound, they show this probability tends to 0 as n increases and this is done for every $\epsilon > 0$. This proves that the preactivations y_v^1 converge to the value $(W_1 + W_2)\mu$, **in probability**. Since the activation function σ is assumed to be Lipschitz continuous the output of the first layer converges to $\sigma((W_1 + W_2)\mu)$ in probability. A similar technique is used for the second layer, and the result is extended by induction. Assumption 4 is just a small technicality - you simply don't want the value $\sigma((W_1 + W_2)\mu)$ (and similarly for the higher layers) to land exactly on the classification boundary, because then the convergence to this value would not imply that the graph receives the same label. I omitted some details, but this is the general overview.

3 A small idea

I personally don't find concentration bounds very intuitive. Perhaps there is a more fundamental reason why this convergence occurs. Consider the preactivations once again:

$$y_v^1 = \frac{1}{N(v)} W_1 \sum_{u \in N(v)} x_u + \frac{1}{n} W_2 \sum_{u \in V} x_u + b$$

In particular, focus on the second term. We are summing n independent and identically distributed random variables and dividing by n. Note that the random variables are $W_2 x_u$, and since $f(x) = W_2 x$ is a linear transformation, these are integrable and have mean $W_2\mu$. This means we can directly apply the

[2]Shortly before submitting, I noticed that I forgot about the bias term. This means that all of the arguments in this report are missing a $+b$ (but the logic is exactly the same).

strong law of large numbers, one of the most fundamental results in probability theory, to get convergence to the mean. Moreover, we get a stroner type of convergence, namely **almost sure convergence**, which implies convergence in probability.

What about the first term? Here it is exactly the same idea but now we have a problem - the sum we are taking is in fact random, because the neighborhood of the vertex v follows a $Bin(n,r)$[3] distribution because the graphs are sampled from the Erdos-Renyi model. Can we still use the strong law of large numbers? Intuitively, the answer is a resounding yes, because for large n, $N(v)$ will be large, so there will be enough terms in the sum. Let's prove it.

Theorem 2. *Let (X_k) be a sequence of integrable i.i.d random variables and let (B_n) be a independent sequence with $B_n \sim Bin(n,r)$ for $r > 0$. Then*

$$\frac{X_1 + X_2 + \cdots + X_{B_n}}{B_n} \to \mathbb{E}[X] \text{ a.s}$$

Proof. Suppose $\epsilon > 0$ is given. If we show that:

$$\sum_{n=1}^{\infty} \mathbb{P}\left(\left|\frac{X_1 + \cdots + X_{B_n}}{B_n} - \mathbb{E}[X]\right| > \epsilon\right) < \infty$$

then by the first Borel Cantelli lemma, we would be done.
Using the law of total probability and conditioning, we have:

$$\mathbb{P}\left(\left|\frac{X_1 + \cdots + X_{B_n}}{B_n} - \mathbb{E}[X]\right| > \epsilon\right) = \sum_{i=0}^{n} \mathbb{P}\left(\left|\frac{X_1 + \cdots + X_i}{i} - \mathbb{E}[X]\right| > \epsilon | B_n = i\right) \mathbb{P}(B_n = i)$$

Now we will do a small trick - we will turn the sum into an infinite one but introduce indicators:

$$\sum_{i=0}^{\infty} \mathbb{P}\left(\left|\frac{X_1 + \ldots X_i}{i} - \mathbb{E}[X]\right| > \epsilon\right) \mathbb{P}(B_n = i) \mathbb{1}(i \leq n)$$

Now what we need to show is:

$$\sum_{n=1}^{\infty}\sum_{i=0}^{\infty} \mathbb{P}\left(\left|\frac{X_1 + \cdots + X_i}{i} - \mathbb{E}[X]\right| > \epsilon\right) \mathbb{P}(B_n = i) \mathbb{1}(i \leq n) < \infty$$

Exchanging the order of the sums, we get:

$$\sum_{i=0}^{\infty}\sum_{n=i}^{\infty} \mathbb{P}\left(\left|\frac{X_1 + \cdots + X_i}{i} - \mathbb{E}[X]\right| > \epsilon\right) \mathbb{P}(B_n = i)$$
$$\sum_{i=0}^{\infty} \mathbb{P}\left(\left|\frac{X_1 + \cdots + X_i}{i} - \mathbb{E}[X]\right| > \epsilon\right) \sum_{n=i}^{\infty} \mathbb{P}(B_n = i)$$

Observe that the second sum is just:

[3]Actually it is a $1 + Bin(n-1,r)$distribution because the self loop is always present, but it makes no difference to the analysis

$$\sum_{n=i}^{\infty} \mathbb{P}(B_n = i) = \sum_{n=i}^{\infty} \binom{n}{i} r^i (1-r)^{n-i}$$
$$= r^i \sum_{n=i}^{\infty} \binom{n}{i} (1-r)^{n-i}$$
$$= r^i \frac{1}{r^{i+1}} = 1/r$$

where we used the Taylor series for $1/(1-x)^\alpha$. This means we can rewrite our sum as

$$1/r \sum_{i=0}^{\infty} \mathbb{P}\left(\left| \frac{X_1 + \cdots + X_i}{i} - \mathbb{E}[X] \right| > \epsilon \right)$$

and we need to show this converges.

But this is easy! We know this converges from the **ordinary** strong law of large numbers. □

Remark 1. *One should always be careful when exchanging the order of limits/sums/integrals. However, in the above we use the following fact: Let a_{mn} be a double sequence in the extended real numbers. Then $\sup_{mn} a_{mn} = \sup_n \sup_m a_{mn} = \sup_m \sup_n a_{mn}$*

Remark 2. *To be able to deduce that the last sum converges, you must actually look at the proof of the strong law of large numbers. We cannot deduce that the above sum converges just from the statement of the strong law, because the implication in Borel Cantelli lemmas is the other way (unless you have independence between the events which are being summed, which we don't (I don't mean independence between the variables)).*

Remark 3. *We could consider other random graph models (i.e $B_n \not\sim Bin(n,r)$). All one has to do is show:*

$$\sum_{i=0}^{\infty} \sum_{n=i}^{\infty} \mathbb{P}(B_n = i) < \infty$$

which is easy to do (if not analytically you can check this numerically).

Remark 4. *Notice that we did not impose any restrictions on the distribution D, only that it must have a mean. In the paper, the authors require the distribution to be sub-Gaussian, but we see this is not necessary.*

So what is the point? The main point of this section (and this entire work) is to give an intuitive understanding of the zero one laws and understand, from a theoretical point of view, the driving mechanism behind the phenomenon. In this section, we discovered that a possible explanation of the zero one laws is the strong law of large numbers.

This is further supported by the fact that we do not need assumptions on the distribution D.

Let's summarize this section:

1. We presented a novel approach to the proof of the zero one law.
2. We obtained a stronger type of convergence.
3. We discovered a sufficient condition for other random graph models.
4. We discovered that assumption 1 is not needed - the distribution need not be subgaussian.
5. We achieved an intuitive understanding of the mechanism.

Note that we are still only considering one layer. We will comment on the general case later. Let us now move on to the next model, which is where we will see the real benefits of our approach.

4 SUMGNN

Equipped with the theoretical tools from the previous section, let us analyze SUMGNN. The theorem presented in the paper states that a SUMGNN model also satisfies a zero one law but now the assumptions are slightly different. Firstly, we don't require the nonlinearity to be Lipschitz, but rather that it is eventually constant in both directions (with "eventual values" denoted σ_∞ and $\sigma_{-\infty}$). More importantly, we get a new condition, which states the model must be "synchronously saturating". This is defined as follows:

Definition 3. *Let M be a SUMGNN model using the aggregation:*

$$z_v^{t+1} = \sigma\left(W_1^t z_v^t + W_2^t \sum_{u \in N(v)} z_u^t + W_3^t \sum_{u \in V} z_u^t + b\right)$$

where $z_u^0 = x_u$. We say M is synchronously saturating if:

1. $(rW_2^t + W_1^t)\mu \neq 0$
2. $(rW_2^t + W_1^t)z_t \neq 0$

We would like to understand why this condition is necessary and whether it can be removed. This is for two reasons: Firstly, this condition imposes restrictions on the weights. Perhaps this is not too problematic, because if the weight matrices have a high rank, the kernel is low dimensional, and will have 0 measure. In my opinion, the bigger problem is the first condition, because it additionally imposes a condition on the mean μ of the feature distribution D - it cannot be zero. However, zero mean is very common (in many ML applications, it is even desired) so I find condition 1 a little problematic. The next couple of pages will focus on whether conditions 1 and 2 can be eliminated. Here we will see that our approach, which we illustrated on the case of MeanGNN, really pays off. However, we must do a bit more work.

4.1 Random walks

It is clear that for the laws of large numbers to work, you must normalize the sum of the variables (compute the average), as was done in the case of MEANGNN. So what happens if you remove the normalization, as in SUMGNN? What can you say about the following sum:

$$S_n = X_1 + X_2 + \cdots + X_n$$

where (X_i) are i.i.d, integrable random variables. Well, this is just a random walk, and in some cases, understanding the behaviour is very easy:

Rewriting we have:

$$S_n = \frac{S_n}{n} n = \frac{X_1 + X_2 + \cdots + X_n}{n} n$$

Recall that with almost sure convergence, we have the algebra of limits. Thus taking limits:

$$\begin{aligned} \lim_{n\to\infty} S_n &= \lim_{n\to\infty} \frac{X_1 + X_2 + \cdots + X_n}{n} \lim_{n\to\infty} n \quad \text{a. s} \\ &= \mathbb{E}[X] \lim_{n\to\infty} n \quad \text{a. s} \end{aligned}$$

where we used the strong law of large numbers. Now the behaviour is trivial: If $\mathbb{E}[X] < 0$, $S_n \to -\infty$. If $\mathbb{E}[X] > 0$, $S_n \to \infty$. But if $\mathbb{E}[X] = 0$, we have an expression of type $0 \times \infty$, which doesn't tell us much.

Let's try to look at the last case more closely. There must be something we can say. From an intuitive point of view, one would expect the walk to oscillate. But it is not clear what this means - does it stay close to 0? How often does it go far away? What's going on? One way to answer these questions is to look at extremal values of the walk. This inspires the following theorem.

Theorem 4. *Let (X_n) be a sequence of* ***1 dimensional*** *i.i.d square integrable random variables with mean 0 and variance σ^2. Let $S_k = \sum_{n=1}^{k} X_n$. Then $\mathbb{P}(\limsup S_k = \infty) = 1$ and $\mathbb{P}(\liminf S_k = -\infty) = 1$.*

Remark 5. *Notice that we require square integrability. This is because we will use the central limit theorem in our proof, which requires square integrability. This is a weaker assumption than sub-gaussian.*

Proof. Let us only prove the lim sup case (the other case is similar).

First, observe that $\{\limsup S_n = \infty\}$ is a tail event (meaning it belongs to the tail sigma algebra of the independent random variables (X_i)). By Kolmogorov's 0-1 law, it either has probability 0 or 1 (but we don't know which). The proof will rely on this fact.

By CLT, we know

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n}{\sqrt{n}} > 1\right) = 1 - \Phi(1) > 0$$

where Φ is the cdf of $N(0, \sigma^2)$. Now a trivial observation: for any $m \in \mathbb{N}$ we have:

$$\mathbb{P}\left(\frac{S_m}{\sqrt{m}} > 1\right) \leq \mathbb{P}\left(\bigcup_{k=m}^{\infty} \frac{S_k}{\sqrt{k}} > 1\right)$$

since the event on the right is larger. Taking limits we have:

$$\lim_{m\to\infty} \mathbb{P}\left(\frac{S_m}{\sqrt{m}} > 1\right) \leq \lim_{m\to\infty} \mathbb{P}\left(\bigcup_{k=m}^{\infty} \frac{S_k}{\sqrt{k}} > 1\right)$$

but by the previous fact, we know the left hand side is strictly greater than zero. Thus we can conclude that:

$$\lim_{m\to\infty} \mathbb{P}\left(\bigcup_{k=m}^{\infty} \frac{S_k}{\sqrt{k}} > 1\right) > 0$$

A careful reader may notice that it is not yet clear whether the expression is actually well defined (the limit may not exist). But observe that the events

$$A_m := \bigcup_{k=m}^{\infty} \frac{S_k}{\sqrt{k}} > 1$$

are decreasing, meaning that $A_1 \supset A_2 \supset A_3 \supset \dots$ (this is because as we increase m, the union gets smaller). Thus we can use continuity of the probability measure (it's finite) to conclude that

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} A_m\right) = \lim_{m\to\infty} \mathbb{P}(A_m)$$

meaning that the limit exists and we know what it is. By what we showed previously, it is strictly greater than zero. Plugging back the definition of A_m into the LHS we have:

$$\mathbb{P}\left(\bigcap_{m=1}^{\infty} \bigcup_{k=m}^{\infty} \frac{S_k}{\sqrt{k}} > 1\right) > 0$$

but this implies:

$$\mathbb{P}\left(\limsup \frac{S_k}{\sqrt{k}} \geq 1\right) > 0$$

by definition of lim sup.

Now we are almost done. Clearly, if $\limsup \frac{S_k}{\sqrt{k}} \geq 1$, then $\limsup S_k = \infty$. This means the latter is a larger event. Thus we have:

$$0 < \mathbb{P}\left(\limsup \frac{S_k}{\sqrt{k}} \geq 1\right) \leq \mathbb{P}\left(\limsup S_k = \infty\right)$$

But we know that the RHS is either zero or one, because it is a tail event (this was said at the beginning). Since we showed it is greater than zero, it therefore must be one. Thus

$$\mathbb{P}\left(\limsup S_k = \infty\right) = 1$$

as required. The $\liminf$ case is done similarly. □

Remark 6. *I found this proof incredibly satisfying. We are using 0-1 laws of probability theory to talk about 0-1 laws of graph neural networks. Very cool.*

To summarize, we have exactly three possibilities:

1. If $\mathbb{E}[X] > 0$, $S_k \to \infty$ almost surely
2. If $\mathbb{E}[X] < 0$, $S_k \to -\infty$ almost surely
3. If $\mathbb{E}[X] = 0$, then $\limsup S_k = \infty$ **AND** $\liminf S_k = -\infty$ almost surely. This means the walk oscillates - it has diverging maximal and minimal values with probability 1.

Let's now relate it to the theorem in the paper.

4.2 The zero one law

Recall that the preactivation of the first layer of a SumGNN is given as:

$$y_v^1 = W_1 x_v + W_2 \sum_{u \in N(v)} x_u + W_3 \sum_{u \in V} x_u + b$$

Let us fix an arbitrary component i. Observe that the third term is exactly a 1 dimensional random walk with random variables $(W_3 x_u)_i$. The second term is a random walk with random variables $(W_2 x_u)_i$ and length following a $Bin(n, r)$ distribution. We can use $Ber(r)$ indicator variables to turn it into a length n random walk, and now we can sum the two random walks. This results in a random walk with mean $((rW_2 + W_1)\mu)_i$ (the r comes from the expectation of the indicator variable). Now, if we are in the case that this is > 0, then by our theorem, the random walk diverges to ∞, and thus the i-th component of the preactivation y_v^1 will also diverge to ∞. This means that if we have a nonlinearity σ which is eventually constant with value σ_∞, the i-th component of the output of the first layer is σ_∞. Similarly for the < 0 case. Since the component was arbitrary, the result of the first layer is a vector $z \in \{\sigma_\infty, \sigma_{-\infty}\}^d$ which means the output of the first layer converges. As we showed above, this happens with probability 1. The proof in the paper follows the same logic, but uses concentration bounds and does not mention the connection to random walks.

So what can we say about the "non-synchronously saturating" case? Notice that this is exactly case 3 in the previous theorem, meaning that we have a random walk with zero expectation. Now we can utilize the theorem we proved - we know that $\limsup S_k = \infty$ and $\liminf S_k = -\infty$, in such a random walk. This means that each component of the preactivation of the first layer oscillates as n increases - it sometimes drifts to large positive values, then it comes back and drifts to large negative values and so on, and this process is repeated forever, with probability 1. If this is the case, not only it is clear that we cannot conclude a zero one law is satisfied as before, we can in fact conclude that it is not satisfied, because the preactivations oscillate all over the place. This was not proved in the paper and it is where this report extends the result.

Let me mention one other avenue which I found fascinating. One reason why we see this behaviour is because we are applying the non-linearity component-wise. Things would get much more interesting if this was not the case. The reason for that is that once you look at random walks in higher dimensions, the behaviour is much more delicate. In particular, the Chung-Fuchs theorem[4] states that a random walk with increments whose expectation is 0 (the non-synchronously saturating case) always diverges, provided it is in at least 3 dimensions. This means that if we had a nonlinearity, which was eventually constant in the sense that

$$\sigma(x) = C \text{ if } ||x|| \geq R$$

for some radius R, one would observe a very interesting behaviour - the zero one law would be satisfied if the number of dimensions is ≥ 3. I am not sure if such nonlinearities are useful in machine learning applications, but I found this very interesting. Note that this does not contradict the findings in the previous paragraph.

Remark 7. *It is difficult to verify these claims experimentally. Absence of evidence is not evidence of absence, so experimentally proving that something does not converge is difficult. In the experiments shown in the paper, the convergence of SUMGNN was already quite slow.*

Let's summarize:

1. We discussed the necessity of the "synchronously saturating" assumption.
2. We discovered a connection between SumGNN and random walks.
3. We explored the theory of random walks and proved that any square integrable random walk belongs to one of three cases.
4. Using our theorem, we discovered that the "synchronously saturating" assumption must not be removed. Additionally, we formulated the following **conjecture**: A non synchronously saturating SumGNN satisfies a zero one law if and only if the node features have dimension at least 3 and the nonlinearity is eventually constant in the multivariate sense.

To conclude, I believe that this connection to random walks deserves to be studied more thoroughly.

5 GCN and GAT

The last model which is discussed in the paper is the GCN. Additionally, the authors conjecture that GATs also follow a zero one law, but do not explore this. In this section, I want to briefly look into these.

Looking at the GCN proof in the paper, one might wonder how the proof would work for GATs. In the proof, a concentration bound is used to estimate the normalization coefficients of the GCN, whose distribution is known. But how would this proof work for cases where the coefficients are more general, such as in GATs? This motivated the following question which I would like to study:

Question 1. *Suppose we have a GNN with the following aggregation scheme:*

$$y_v^1 = \sum_{u \in N(v)} \alpha_{vu} W_1 x_u$$

Under what conditions on the coefficients α_{vu}, and the random graph distribution, does the GNN satisfy a zero one law?

[4] https://en.wikipedia.org/wiki/Chung–Fuchs_theorem

This is where we can utilize the intuition we developed so far. Firstly, let us denote $X_i := W_1 x_i$ for node $i \in N(v)$ (these are i.i.d integrable), and denote the coefficient by α_i. Now, rewriting the sum, using the same trick as before, we have:

$$y_v^1 = \frac{\alpha_1 X_1 + \alpha_2 X_2 + \cdots + \alpha_{N(v)} X_{|N(v)|}}{|N(v)|} |N(v)| \tag{1}$$

where $|N(v)|$ follows some distribution based on the random graph model.[5] We are interested in the limit as n goes to infinity. Now, the problem is more complicated because the coefficients α_i depend on n. To complicate things further, in models like GAT, the coefficients are not independent. Let's put this aside for now and assume they are pairwise independent. This assumption is unrealistic, but we should start with the easiest case. So the problem we are considering is this:

1. X_i are i.i.d from a fixed distribution D with mean μ
2. For every n, the coefficients α_i are sampled from a distribution D_n and are pairwise independent.
3. The neighborhood $N(v)$ is sampled from a distribution according to the random graph model we choose (e.g $Bin(n, r)$ for Erdos-Renyi)

Let us further rewrite (1). By defining $\beta_i := \alpha_i N(v)$ and bringing the $N(v)$ inside we have:

$$y_v^1 = \frac{\beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_{N(v)} X_{N(v)}}{|N(v)|} \tag{2}$$

Now why did we do this? If you recall Theorem 2, if the β were all equal to a constant c, the theorem tells us this converges to $c\mu$. This motivates the following conjecture:

Conjecture 1. *If the distribution $D(n)$ and the random graph model are such that $D(n) \times |N(v)| \to c$ almost surely, then y_v^1 tends to $c\mu$ almost surely as n tends to infinity and the GNN satisfies a zero one law.*

Let us not go into this from a theoretical point of view, but rather look at some experiments. You can find the experiments on this address: https://github.com/1069907/Geometric-Deep-Learning.

6 Conclusion

In this report, I tried to illustrate the connection between one of the most fundamental results in probability theory, the strong law of large numbers, and the zero one laws of graph neural networks. Using this connection, we developed intuition about why the zero one laws exist, proposed simple proofs for the one layer case, and obtained a stronger type of convergence while requiring weaker assumptions compared to the paper. The SLLN inspired us to explore the connection between SUMGNN and random walks, where we extended the results of the paper and formulated an interesting conjecture. Lastly, we discussed the behaviour of more general models and stated a conjecture about such models. We then experimentally verified our hypothesis.

Let us now discuss some shortcomings of our approach. One problem is that it is not straightforward to argue about higher layers of the GNN using our method. This is obviously a big problem and we cannot pretend that this can be overlooked in any way. On the other hand, I believe it is the first layer that makes or breaks the zero one law[6]. Initially, we start with very little information - we only have independent node features sampled from a distribution D and all we know is that it has a mean. After the aggregation of the first layer, we already know that each node converges to the mean, not only in probability, but almost surely. This means that the input to the second layer is much less random than in the first layer and obtaining convergence should now be "easier". However, it is difficult to turn the above argument into a rigorous proof and in the end, and it is probably easier to argue as was done in the paper. The conclusion is that in this case, our method is inferior.

[5] in case of Erdos-Renyi, this is $Bin(n, r)$

[6] This is a bit of an overstatement, but consider looking at the proof of the second layer which is given in the paper, to understand what I mean.

Finally, I believe that we achieved the goals that we defined in the beginning. The point of this report was to look at the zero one laws in a different light and see if this can bring any fruit. Not only did we manage to extend some of the results, but we also achieved a better intuitive understanding of the phenomenon. To conclude, I believe that the methods presented in this report deserve to be studied further, but their limitations must be taken into consideration.

7 Appendix

Definition 5. *A **random variable** $X : \Omega \to \mathbb{R}$ is a measurable function from a probability space (Ω, Σ, P) to $\mathbb{R}$ equipped with the Borel sigma algebra. We say X is **(square) integrable** if $\int |X| dP$ (resp. $\int X^2 dP$) is finite.*

Definition 6. *Let (E_n) be a sequence of events. We define the event $\{E_n$ infinitely often $\}$ to be the event $\cap_{n=1} \cup_{N \geq n} E_N$.*

Definition 7. *Let (E_n) be a sequence of events. We define the event $\{E_n$ eventually $\}$ to be the event $\cup_{n=1} \cap_{N \geq n} E_N$.*

Definition 8. *We say a sequence of random variables defined on the same probability space, $(X_n)_{n \in \mathbb{N}}$ converges to a random variable X **almost surely** (a.s), if*

$$\forall \epsilon > 0 \mathbb{P}(\{\omega \in \Omega | X_n(\omega) - X(\omega)| < \epsilon \text{ eventually}\}) = 1.$$

Almost sure convergence is just standard pointwise convergence. Thus means we can utilize all kinds of theorems from measure theory (e.g dominated convergence). However, we will not really need this. What is important for us is that a.s convergence gives us algebra of limits.

Definition 9. *We say a sequence of random variables (defined on the same probability space) $(X_n)_{n \in \mathbb{N}}$ converges to a random variable X **in probability**, if $\forall \epsilon > 0, lim_{N \to \infty} \mathbb{P}(|X_n - X| \geq \epsilon) = 0$.*

Almost sure convergence is a stronger type of convergence than convergence in probability. In fact, a.s convergence implies convergence in probability.

Lemma 10. *Borel Cantelli Lemma Let (E_n) be a sequence of events. Then*

$$\sum_{n=1}^{\infty} \mathbb{P}(E_n) < \infty \implies \mathbb{P}(E_n \text{ infinitely often}) = 0$$

Theorem 11. *Strong Law of Large Numbers Let (X_n) be a sequence of pairwise independent integrable random variables with mean μ. Then*

$$\frac{X_1 + X_2 + \ldots X_n}{n} \to \mu \text{ a. s}$$