Project Proposal

During the lecture on 28th of February, we discussed about Johnson and Lindenstrauss lemma. In particular, we elaborate that thanks to this lemma we are able to project n points of a dimension $\mathbb{R}^k$ to a dimension $\mathbb{R}^m$, where $m < k$ while at the same time we preserve the euclidean distances between the n points on the lower dimension representation with a small deviation of ϵ, in which $\epsilon \in (0, \frac{1}{2})$. As we explained the proof, it was clear that the intermediate matrix A needed for the matrix multiplication was build thanks to a Gaussian distribution and multiplied with a normalization constant factor. So a plethora of question arises: "What if the matrix content is based on other distributions? Is the euclidean distance between the points still preserved? Can we experimentally compare different distribution's results?". Microsoft has already tried to work with a different flavor of matrix A, whose values are probabilistically selected among values of set $\{0, -1, 1\}$ in this paper here. Moreover, MIT researchers tried two different sparse parameterized approaches for JL lemma in this paper. Also, alternative solutions have been described by Aarhus and Stanford University researchers respectively. At this project, my initial milestone goals are the following:

- Implement the original Johnson and Lindenstrauss lemma algorithm and experimentally justify that it follows the theoretical bound.
- Reproduce the paper alternative approaches described above, utilize their parameters and check if they suffice the theoretical bound. If they suffice the theoretical bound, is on average the euclidean distance of the projected vectors close to the inequality margins or not?
- Utilize statistical difference metrics to evaluate the differences between different approaches. For example, the standard deviation or t-score metrics can be utilized. Moreover, as discussed with professor system flavor metrics should be interesting such as execution time, memory utilization since some alternative methods require sparse matrix representation rising the question of how it affects machines resources.

At this section, I will answer the questions asked on the proposal document.

- **Why do you find this project interesting?**
  As a PhD student working with Graph Machine Learning, dimentionality reduction techniques are a useful tool for preprocessing the graph data I am using for downstream machine learning tasks. Moreover, storage efficient techniques allow increasing the volume of graph datasets, that Graph ML can be applied upon.
- **What is already known about the problem?**
  My project is relied upon the Johnson and Lindenstrauss lemma that was theoretically proven in course's lecture. Moreover, a plethora of papers provided above analyze different flavors of JL lemma. My goal is to prove them experimentally by tuning their hyperparameters and compare the different approaches with tools of statistical significance and system specific metrics such as memory utilization.
- **How is what you are doing different from what was previously achieved?**
  As far as I am concerned, with the literature study I made, there is no experimental comparison between different flavors of JL lemma. Most of relevant publication only rely on theoretical proofs that their flavor is between the theoretical bounds required without actual comparison between the original or other pappers.

- **What tools are you going to use?**

  My analysis will be based on python language with the help of matrix related libraries such as numpy and scipy.