PAPER REVIEW REPORT

Title: Gradient Scarcity with Bilevel Optimization for Graph Learning

Authors: Hashem Ghanem, Samuel Vaiter, and Nicolas Keriven

Reviewed by Johanna N Kakoto, 200813854

1. INTRODUCTION

Many practical machine learning applications deal with data where only a tiny fraction is labeled (e.g., in medical diagnosis, social networks, or scientific research). This setting is known as semi-supervised learning (SSL), where a model must learn from a combination of labeled and unlabeled data. According to (Y C A Padmanabha Reddy, 2018) the main objective of SSL is to overcome the drawbacks of both supervised and unsupervised learning. Supervised learning requires huge amount of training data to classify the test data, which is cost effective and time-consuming process (Y C A Padmanabha Reddy, 2018). Graphs are a way to represent a combination of labeled and unlabeled data points and edges represent relationships. Graphs allows one to show a mix of labeled and unlabeled data points and edges indicate relationships. A major approach in SSL is to build or improve such graphs during training to make label predictions more accurate.

2. PROBLEM STATEMENT

A basic issue in graph -based semi-supervised learning (SSL) is the gradient scarcity phenomena, which is addressed in the work “Gradient Scarcity with Bilevel Optimization for Graph Learning”. According to (Hashem Ghanem, 2023) In SSL, models are trained using a small subset of labeled data points and a large amount of unlabeled data. Graphs are commonly used to model such data, where each node represents a data point and edges represent relationships or similarities between them. The central problem arises during the graph learning process, where the goal is to adjust the graph structure to improve model performance on the labeled nodes. However, when the model is optimized by minimizing a loss function over only a few labeled nodes, the edges between unlabeled nodes that are far from the labeled ones receive little to no gradient updates. This means that these parts of the graph structure are not learned or improved during training. This issue, known as gradient scarcity, leads to suboptimal graph structures, poor generalization to unlabeled nodes, and potential overfitting to the small labeled subset.

While this problem has been previously observed in joint optimization settings (where the graph and the Graph Neural Network (GNN) parameters are learned together), the paper explores a more advanced training framework: bilevel optimization. In this setting, the graph is optimized in an outer loop, while the model parameters are optimized in an inner loop. The key question the paper addresses is:

Does gradient scarcity still occur in bilevel optimization, and if so, how does it manifest?

The authors also investigate whether this problem is limited to GNNs which have a finite receptive field or whether it also occurs in Laplacian-based regularization models, which have an infinite receptive field.

3. MAIN CONTRIBUTION OF THE WORK

The main contributions of the paper are:

- Theoretical Characterization of Gradient Scarcity

  The paper mathematically proves that gradient scarcity exisit in bilevel optimization settings. It shows that in GNNs, nodes and edges that are more than *k* hops away from labeled nodes receive zero gradients. This result holds regardless of whether the optimization is joint, alternating, or bilevel.

- Extension to Laplacian Regularization Models

  The authors demonstrated that even Laplacian regularization, which theoretically uses information from the entire graph, suffers from gradient scarcity. In these models, gradient magnitudes decay exponentially with the graph distance from labeled nodes, indicating a subtler but still significant form of the same issue.

The paper also proposed solutions to mitigate gradient scarcity by introducing the following methods:

a. Latent Graph Learning (G2G models)**:** Train a model to generate the graph based on node features.
b. Graph Regularization: Impose structural priors to guide learning towards well-formed graphs.
c. Generalized Edge Refinement: Expand the graph by adding multi-hop edges to reduce the effective distance between nodes.

- Empirical Validation

  Through experiments on synthetic data, a custom “cheaters” dataset, and the Cora dataset, the paper validates:

a. The existence and impact of gradient scarcity in practice.
b. That the proposed solutions help reduce the scarcity.
c. That reducing gradient scarcity does not automatically improve generalization**,** highlighting a complex relationship with overfitting.

4. LITERATURE REVIEW

Several machine learning scholars have researched on the issue of learning efficient graph structures for semi-supervised machine learning. The foundational and recent works that set the scene for the current paper are reviewed below, with special attention paid to gradient flow in graph models, bilevel optimization, graph neural networks (GNNs), and semi-supervised learning (SSL).

Semi-supervised learning and label propagation

Zhu et al. (2003, 2005) introduced label propagation methods, which were one of the first and most significant approaches to semi-supervised learning using graphs. These algorithms disperse labels from a small number of labeled nodes throughout the graph by taking advantage of the homophiles assumption, which states that neighboring nodes in a graph are probably going to have the same label. They enforce smoothness over the node labels using the graph Laplacian. The foundation for numerous later graph-based SSL techniques was established by these techniques.

Graph Neural Networks and finite receptive fields

Graph Neural Networks (GNNs) emerged as a leading framework for learning on graph-structured data with the advent of deep learning. For SSL on citation networks and social graphs, models like the Graph Convolutional Network (GCN) by Kipf and Welling (2017) and the Graph Attention Networks (GAT) by Veličković et al. (2018) gained popularity.

However, because each layer aggregates data from nearby neighbors, GNNs usually have a small receptive field. The information flow across the graph is restricted by this architectural constraint, particularly in deeper layers where over-smoothing may take place. In situations involving sparse labeling, this limitation becomes especially troublesome.

Gradient Scarcity and Supervision Starvation

The phenomenon of gradient scarcity, in which some graph regions receive little to no gradient updates during training, was first thoroughly examined by Fatemi et al. (2021). They dubbed the issue "supervision starvation" and demonstrated that when a GNN and the graph structure are optimized together, edges between unlabeled nodes that are distant from labeled ones receive zero gradients. This served as the direct inspiration for the current study, which explores the same phenomenon in bilevel optimization settings.

Bilevel Optimization in Graph Learning

Bilevel optimization became a popular powerful training framework, especially for problems involving hyperparameter tuning and meta-learning. In the context of graph learning, Franceschi et al. (2019) proposed using bilevel optimization to learn discrete graph structures by optimizing graph edges in the outer loop and model parameters in the inner loop. This framework allows learning task-specific graphs, but also makes gradient computation more difficult, especially for unlabeled parts of the graph. The paper in review builds on this approach but introduces a new theoretical concern: even with the added expressiveness of bilevel optimization, gradient scarcity still exists. The authors also prove that this occurs in models with infinite receptive fields, such as those using Laplacian regularization.

Graph Regularization and Structure Learning

In order to learn graphs from data, Kalofolias (2016) suggested optimizing for signal smoothness across the graph. This method influenced the creation of regularization-based methods, which use a penalty to promote significant graph structures. In order to address gradient scarcity, the current paper uses a similar approach as one of its suggested solutions, which is graph regularization.

Latent Graph Learning and G2G Models

Instead of depending on a noisy observed graph, Jiang et al. (2019) proposed the Graph-to-Graph (G2G) model, which suggests learning a latent graph from node features. Limitations in observed connectivity may be addressed by this model, which views the graph as a function of node attributes. By ensuring that distant nodes can still be efficiently connected and updated during training, the current study adapts this concept to suggest latent graph learning as a way to mitigate gradient scarcity.

Comparison with Graph Agreement Models (GAM)

The authors use the Graph Agreement Model (GAM), which ensures consistency between node label similarity and graph structure, as a benchmark for their work. Although GAM is good at making graphs better, it doesn't solve the problems with gradient flow that this paper examines. However, when it comes to test accuracy on benchmark datasets, it is still competitive.

Although it doesn't address bilevel-specific gradient issues, GAM is cited in the paper as a baseline graph refinement method that produces good results.

Although bilevel optimization, semi-supervised learning, and graph construction have all been the subject of numerous earlier studies, no one has formally defined or addressed gradient scarcity in bilevel settings, particularly across finite and infinite receptive field models. This study closes that gap by demonstrating that gradient scarcity is a widespread occurrence in graph learning, supplying theoretical limits for models based on Laplacian, and presenting empirical fixes that enhance gradient flow directly.

5. LIMITATIONS OF THE PAPER

Despite the contributions made by the paper, the paper also has its limitations as follows:

- Scalability: Experiments are limited to small or medium-sized datasets. It remains unclear how the solutions scale to large graphs (e.g., with millions of nodes).
- Loose Bounds: The theoretical results for Laplacian models rely on upper bounds that may not tightly reflect real-world decay rates.
- Not Always Better than state of the art (SOTA): The proposed solutions improve training accuracy and gradient coverage but do not always beat state-of-the-art models like GAM in test accuracy.
- Distinction Between Scarcity and Overfitting: While the paper clearly separates the two concepts, it does not fully explore how to address both simultaneously.

REFERENCES

1. Hashem Ghanem, S.V. (2023).Gradient Scarcity with Bilevel Optimization.*Gradient Scarcity with Bilevel Optimization.*
2. Y C A Padmanabha Reddy,P.V. (2018).Semi-supervised learning:a brief review.*Semi-supervised learning:a brief review,2*
3. Zhu, X. (2005). *Semi-Supervised Learning with Graphs*. Carnegie Mellon University Technical Report CMU-LTI-05-192.
4. Kipf, T. N., & Welling, M. (2017). Semi-supervised classification with graph convolutional networks. *International Conference on Learning Representations (ICLR)*.

5. Veličković, P., Cucurull, G., Casanova, A., Romero, A., Lio, P., & Bengio, Y. (2018). Graph Attention Networks. *International Conference on Learning Representations (ICLR)*.
6. Fatemi, E., Li, Y., Bui, T., Barati Farimani, A., & Zemel, R. (2021). SLAPS: Self-supervision improves structure learning for graph neural networks. *Advances in Neural Information Processing Systems (NeurIPS)*, 34.
7. Franceschi, L., Niepert, M., Pontil, M., & He, X. (2019). Learning discrete structures for graph neural networks. *International Conference on Machine Learning (ICML)*, 1972–1982.
8. Kalofolias, V. (2016). How to learn a graph from smooth signals. *Proceedings of the International Conference on Artificial Intelligence and Statistics (AISTATS)*, 920–929.
9. Jiang, B., Zhang, C., Ming, D., Tang, J., & Lu, Z. (2019). Semi-supervised learning with graph learning-convolutional networks. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, 11313–11320.
10. Ghanem, H., Vaiter, S., & Keriven, N. (2024). Gradient scarcity with bilevel optimization for graph learning. *Transactions on Machine Learning Research*. HAL Id: hal-04041721.
11. Wu, Z., Pan, S., Chen, F., Long, G., Zhang, C., & Yu, P. S. (2021). A comprehensive survey on graph neural networks. *IEEE Transactions on Neural Networks and Learning Systems*, 32(1), 4–24.
12. Liu, Y., Lin, Z., Li, Q., Zhou, J., & Li, H. (2022). Towards unsupervised graph structure learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(6), 6585–6593.