ICML
International Conference On Machine Learning

Project Page

Nonparametric Teaching for Graph Property Learners

Chen Zhang[1]*, **Weixin Bu**[2]*, **Zeyi Ren**[1], **Zhengwu Liu**[1], **Yik-Chung Wu**[1], **Ngai Wong**[1]

[1]The University of Hong Kong

[2]Reversible Inc

香 港 大 學
THE UNIVERSITY OF HONG KONG

REVERSIBLE

May 19, 2025

Content by: Chen Zhang.

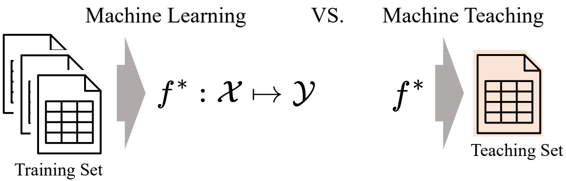
Overview

Nonparametric Teaching

What is Nonparametric Teaching?

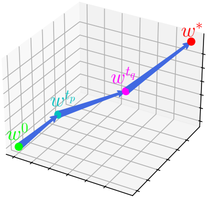
Nonparametric Teaching builds on the idea of *machine teaching* [14, 15]–involving designing a training set (dubbed the teaching set) to help the learner rapidly converge to the target functions–but relaxes the assumption of target functions being parametric [8, 9], allowing for the teaching of nonparametric (viz. non-closed-form) target functions, with a focus on function space.

Machine teaching can be considered as an inverse problem of machine learning, where machine learning aims to learn a model from a dataset, while machine teaching aims to find a minimal dataset from the target model.

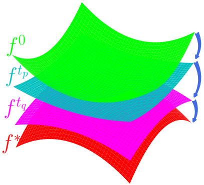


“Parametric” VS. “Nonparametric”

The parametric case [8, 9] assumes that f can be represented by a set of parameters $\boldsymbol{w}$, *e.g.*, $f(\boldsymbol{x}) = \langle \boldsymbol{w}, \boldsymbol{x} \rangle$ with input $\boldsymbol{x}$[1].



(a) Parametric IMT



(b) Nonparametric IMT

Parametric assumption results in difficulty when the target models are defined to be functions without dependency on parameters (viz. non-closed-form functions). Such a limitation is addressed by **Nonparametric Teaching** [11, 12, 13], which generalizes model space from a finite dimensional one to an infinite dimensional one.

[1] The loss $\mathcal{L}$ can be general for different tasks, *e.g.*, square loss for regression and hinge loss for classification.

Graph Neural Teaching (GraNT)

Graph Property Learning

Graph-structured data, commonly referred to as graphs, are typically represented by vertices and edges. The vertices, or nodes, contain individual features, while the edges link these nodes and capture the structural information, collectively forming a complete graph.

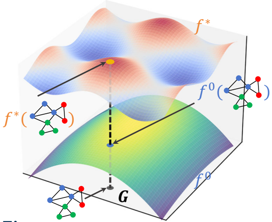


Figure: The implicit mapping.

Graph properties can be categorized as either node-level or graph-level. For example, the node category is a node-level property in social network graphs [3], while the solubility of molecules is a graph-level property in molecular graphs [10]. Inferring these graph properties essentially involves learning the implicit mapping from graphs to these properties [4].

Graph Convolutional Network (GCN)

We introduce the concatenation operation $\bigoplus$ and define

$$\boldsymbol{A}^{[\kappa]} := \bigoplus_{i=0}^{\kappa-1} \boldsymbol{A}^{i} = [\boldsymbol{I}\ \boldsymbol{A}\ \cdots\ \boldsymbol{A}^{\kappa-1}],$$

an $n \times \kappa n$ matrix. By unfolding the aggregated features at different orders and assigning them distinct weights [6], the **flexible GCN** can be expressed as

$$\boldsymbol{X}^{(\ell)} = \sigma\left(\boldsymbol{A}^{[\kappa_\ell]}\mathrm{diag}(\boldsymbol{X}^{(\ell-1)}; \kappa_\ell) \cdot \boldsymbol{W}^{(\ell)}\right), \ell \in \mathbb{N}_{L-1}$$
$$\boldsymbol{X}^{(L)} = \rho\left(\boldsymbol{A}^{[\kappa_L]}\mathrm{diag}(\boldsymbol{X}^{(L-1)}; \kappa_L) \cdot \boldsymbol{W}^{(L)}\right). \quad (1)$$

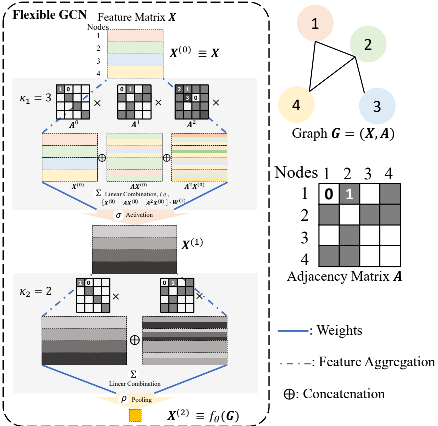


Figure: A workflow illustration of a two-layer flexible GCN with a four-node graph G as input.

Motivation

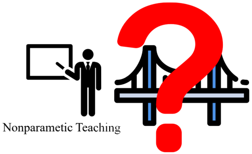
The motivation comes from two folds:

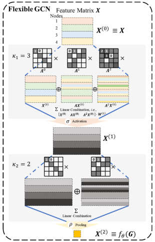
- Lower the training cost and enhance the training efficiency of GCN, which is urgently needed when dealing with large-scale graphs. For example, learning node-level properties in real-world e-commerce relational networks involves millions of nodes.
- Expand the applicability of nonparametric teaching in the context of graph property learning. "Nonparametric" is a quite abstract concept, which may be of interest for theoretical analysis but less practical.

Cont.

- If we can connect nonparametric teaching to GCN training, both problems including training efficiency and applicability are addressed.
- Unfortunately, the evolution of an GCN is typically achieved by gradient descent on its parameters, whereas nonparametric teaching involves functional gradient descent as the means of function evolution.

Bridging this (theoretical + practical) gap is of great value and calls for more examination prior to the application of nonparametric teaching algorithms in the context of graph property learning. ***Can we do that***?

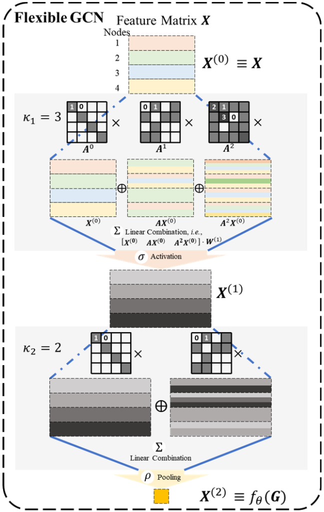




Cont.

Graph Neural Tangent Kernel

Nonparametic Teaching



Graph Neural Tangent Kernel

Graph Neural Tangent Kernel [5, 7, 1, 2] is a symmetric and positive definite kernel function, which is derived from the analysis of the evolution of a GCN.

$$K_{\theta^t}(\boldsymbol{G}_i, \cdot) := \left\langle \frac{\partial f_{\theta^t}(\boldsymbol{G}_i)}{\partial \theta^t}, \frac{\partial f_{\theta^t}(\cdot)}{\partial \theta^t} \right\rangle$$

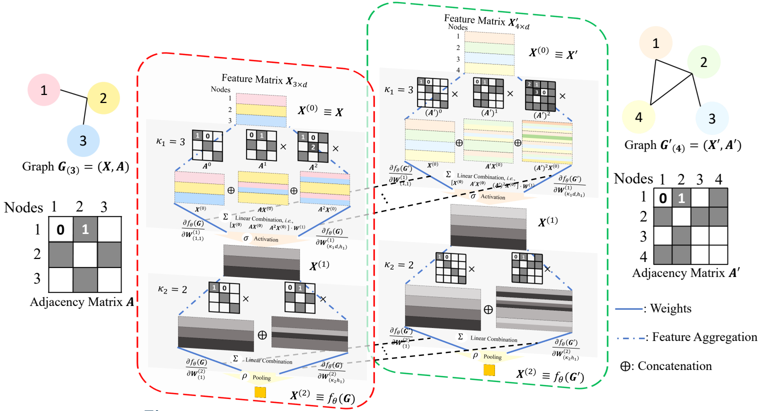


Figure: Graphical illustration of GNTK computation.

GraNT Algorithm

Algorithm 1 GraNT Algorithm

Input: Target mapping f^* realized by a dense set of graph-property pairs, initial GCN f_{θ^0}, the size of selected training set $m \leq N$, small constant $\epsilon > 0$ and maximal iteration number T.

Set $f_{\theta^t} \leftarrow f_{\theta^0}, t = 0$.

while $t \leq T$ and $\|[f_{\theta^t}(\boldsymbol{G}_i) - f^*(\boldsymbol{G}_i)]_N\|_2 \geq \epsilon$ **do**

The teacher selects m teaching graphs:

`/* (Graph-level) Graphs corresponding to the` m `largest` $|f_{\theta^t}(\boldsymbol{G}_i) - f^*(\boldsymbol{G}_i)|$. `*/`

$$\{\boldsymbol{G}_i\}_m{}^* = \underset{\{\boldsymbol{G}_i\}_m \subseteq \{\boldsymbol{G}_i\}_N}{\arg\max} \|[f_{\theta^t}(\boldsymbol{G}_i) - f^*(\boldsymbol{G}_i)]_m\|_2.$$

`/* (Node-level) Graphs associated with the` m `largest` $\frac{\|f_{\theta^t}(\boldsymbol{G}_i) - f^*(\boldsymbol{G}_i)\|_2}{n_i}$. `*/`

$$\{\boldsymbol{G}_i\}_m{}^* = \underset{\{\boldsymbol{G}_i\}_m \subseteq \{\boldsymbol{G}_i\}_N}{\arg\max} \left\| \left[\frac{f_{\theta^t}(\boldsymbol{G}_i) - f^*(\boldsymbol{G}_i)}{n_i} \right]_m \right\|_{\mathcal{F}},$$

with Frobenius norm $\|\cdot\|_{\mathcal{F}}$.

Provide $\{\boldsymbol{G}_i\}_m{}^*$ to the GCN learner.

The learner updates f_{θ^t} based on received $\{\boldsymbol{G}_i\}_m{}^*$:

`// Parameter-based gradient descent.`

$$\theta^t \leftarrow \theta^t - \frac{\eta}{m} \sum_{\boldsymbol{G}_i \in \{\boldsymbol{G}_i\}_m{}^*} \nabla_\theta \mathcal{L}(f_{\theta^t}(\boldsymbol{G}_i), f^*(\boldsymbol{G}_i)).$$

Set $t \leftarrow t + 1$.

end

By comparing the disparity between the property true values and the GCN outputs, the nonparametric teacher selectively chooses examples (graphs) of the greatest disparity, instead of using all, to feed to the GCN learner who undergoes learning (*i.e.*, training).

Experiments and Results

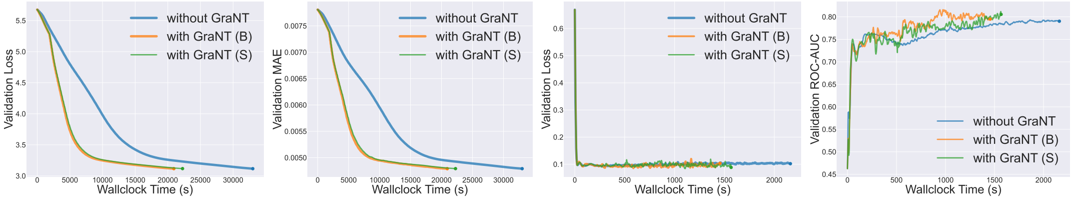
ICML International Conference On Machine Learning

We conduct extensive experiments to validate the effectiveness of GraNT.

GraNT		Dataset	Time (s)	Loss ↓	MAE ↓	ROC-AUC ↑	AP ↑
✗		QM9	9654.81	2.0444	0.0051±0.0009	-	-
		ZINC	33033.82	3.1160	0.0048±0.0004	-	-
		ogbg-molhiv	2163.50	0.1266	-	0.7572±0.0005	-
		ogbg-molpcba	130191.26	0.0577	-	-	0.3270±0.0000
		gen-reg	3344.78	0.0086	0.0007±0.0001	-	-
		gen-cls	11662.25	0.1314	-	0.9150±0.0024	-
✓	B	QM9	6392.26 (-33.79%)	2.0436	0.0051±0.0009	-	-
	B	ZINC	20935.24 (-36.62%)	3.1165	0.0048±0.0004	-	-
	B	ogbg-molhiv	1457.39 (-32.64%)	0.1238	-	0.7676±0.0036	-
	B	ogbg-molpcba	80465.06 (-38.19%)	0.0577	-	-	**0.3358±0.0001**
	B	gen-reg	2308.97 (-30.97%)	0.0086	0.0007±0.0001	-	-
	B	gen-cls	6145.72 (-47.30%)	0.1314	-	**0.9157±0.0013**	-
	S	QM9	7076.37 (-26.71%)	2.0443	0.0051±0.0009	-	-
	S	ZINC	22265.83 (-32.60%)	3.1170	0.0048±0.0004	-	-
	S	ogbg-molhiv	1597.69 (-26.15%)	0.1421	-	**0.7705±0.0027**	-
	S	ogbg-molpcba	89858.65 (-30.98%)	0.0575	-	-	0.3351±0.0025
	S	gen-reg	2337.46 (-30.12%)	0.0086	0.0007±0.0001	-	-
	S	gen-cls	8171.21 (-29.93%)	0.1313	-	**0.9157±0.0014**	-

Table 1: Training time and testing results across different benchmarks. GraNT (B) and GraNT (S) demonstrate similar testing performance while significantly reducing training time compared to the "without GraNT", across graph-level (QM9, ZINC, ogbg-molhiv, ogbg-molpcba) and node-level (gen-reg, gen-cls) datasets, for both regression and classification tasks.

Cont.



(a) ZINC Loss. (b) ZINC MAE. (c) ogbg-molhiv Loss. (d) ogbg-molhiv ROC-AUC.

Figure: Validation set performance for graph-level tasks: ZINC (regression) and ogbg-molhiv (classification).

Contribution Summary

Contributions Summary

Main Contribution:

- We propose **Gra**ph **N**eural **T**eaching (GraNT) that interprets graph property learning within the theoretical context of nonparametric teaching. This enables the use of greedy algorithms from the latter to effectively enhance the learning efficiency of the graph property learner, GCN.
- We unveil a strong link between the evolution of a GCN using gradient descent on its parameters and that of a function using functional gradient descent in nonparametric teaching. These connect nonparametric teaching theory to graph property learning, thus expanding the applicability of nonparametric teaching in the context of graph property learning.
- We demonstrate the effectiveness of GraNT through extensive experiments in graph property learning. Specifically, GraNT saves training time for graph-level regression (-36.62%), graph-level classification (-38.19%), node-level regression (-30.97%) and node-level classification (-47.30%), while upkeeping its generalization performance.

Thank you for listening!

References I

[1] Alberto Bietti and Julien Mairal. On the inductive bias of neural tangent kernels. In NeurIPS, 2019.

[2] Xialiang Dou and Tengyuan Liang. Training neural networks as learning data-adaptive kernels: Provable representation and approximation benefits. Journal of the American Statistical Association, 116(535):1507–1520, 2021.

[3] Wenqi Fan, Yao Ma, Qing Li, Yuan He, Eric Zhao, Jiliang Tang, and Dawei Yin. Graph neural networks for social recommendation. In WWW, 2019.

[4] William L. Hamilton, Rex Ying, and Jure Leskovec. Representation learning on graphs: Methods and applications. IEEE Data Engineering Bulletin, 40:52–74, 2017.

[5] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In NeurIPS, 2018.

[6] Sanjukta Krishnagopal and Luana Ruiz. Graph neural tangent kernel: Convergence on large graphs. In ICML, 2023.

[7] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. In NeurIPS, 2019.

References II

[8] Weiyang Liu, Bo Dai, Ahmad Humayun, Charlene Tay, Chen Yu, Linda B Smith, James M Rehg, and Le Song. Iterative machine teaching. In ICML, 2017.

[9] Weiyang Liu, Bo Dai, Xingguo Li, Zhen Liu, James Rehg, and Le Song. Towards black-box iterative machine teaching. In ICML, 2018.

[10] Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. Scientific data, 1(1):1–7, 2014.

[11] Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric iterative machine teaching. In ICML, 2023.

[12] Chen Zhang, Xiaofeng Cao, Weiyang Liu, Ivor Tsang, and James Kwok. Nonparametric teaching for multiple learners. In NeurIPS, 2023.

[13] Chen Zhang, Steven Tin Sui Luo, Jason Chun Lok Li, Yik-Chung Wu, and Ngai Wong. Nonparametric teaching of implicit neural representations. In ICML, 2024.

[14] Xiaojin Zhu. Machine teaching: An inverse problem to machine learning and an approach toward optimal education. In AAAI, 2015.

[15] Xiaojin Zhu, Adish Singla, Sandra Zilles, and Anna N Rafferty. An overview of machine teaching. arXiv preprint arXiv:1801.05927, 2018.