Robust Learning in Bayesian Parallel Branching Graph Neural Networks: The Narrow Width Limit

Zechen Zhang
Department of Physics, Harvard University
Center for Brain Science, Harvard University
zechen_zhang@g.harvard.edu

Haim Sompolinsky
Center for Brain Science, Harvard University
Kempner Institute for the Study of Natural and Artificial Intelligence, Harvard University
Edmond and Lily Safra Center for Brain Sciences, Hebrew University of Jerusalem
hsompolinsky@mcb.harvard.edu

Abstract

The infinite width limit of random neural networks is known to result in Neural Networks as Gaussian Process (NNGP) (Lee et al. [2018]), characterized by task-independent kernels. It is widely accepted that larger network widths contribute to improved generalization (Park et al. [2019]). However, this work challenges this notion by investigating the narrow width limit of the Bayesian Parallel Branching Graph Neural Network (BPB-GNN), an architecture that resembles residual networks. We demonstrate that when the width of a BPB-GNN is significantly smaller compared to the number of training examples, each branch exhibits more robust learning due to a symmetry breaking of branches in kernel renormalization. Surprisingly, the performance of a BPB-GNN in the narrow width limit is generally superior or comparable to that achieved in the wide width limit in bias-limited scenarios. Furthermore, the readout norms of each branch in the narrow width limit are mostly independent of the architectural hyperparameters but generally reflective of the nature of the data. Our results characterize a newly defined narrow-width regime for parallel branching networks in general.

1 Introduction

The study of neural network architectures has seen substantial growth, particularly in understanding how network width impacts learning and generalization. It is generally believed that wider networks generally perform better (Allen-Zhu et al. [2019], Jacot et al. [2018], Gao et al. [2024]). However, this work challenges the prevailing assumption by exploring the narrow width limit of Bayesian Parallel Branching Graph Neural Networks (BPB-GNNs), an architecture inspired by residual GCN networks (Chen et al. [2020a, 2022]). We show theoretically and empirically that narrow-width networks can perform better than their wider counterparts due to a symmetry-breaking effect in kernel renormalization, in bias-limited scenarios. This paper presents a detailed theoretical analysis of BPB-GNNs in the narrow-width regime, highlighting realistic conditions under which these networks demonstrate robust learning and comparable generalization.

Contributions :

1. We introduce a novel yet simple GCN architecture with parallel independent branches, and derive the exact generalization error for node regression in the statistical limit as the sample

38th Conference on Neural Information Processing Systems (NeurIPS 2024).

size $P \to \infty$ and network width $N \to \infty$, with their ratio a finite number $\alpha = P/N$, in the over-parametrized regime.

2. We show that in the Bayesian setting, the bias will decrease and saturate at narrow hidden layer width, a surprising phenomenon due to kernel renormalization. We demonstrate that this can be understood as a robust learning effect of each branch in the student-teacher task, where each student branch is learning the teacher's branch.
3. We demonstrate this narrow-width limit in real-world dataset Cora and understand each branch's importance as a nature of the dataset.

2 BPB-GNN

We are motivated to study the parallel branching networks as they resemble residual blocks in commonly used architectures and tractable to study analytically with our Bayesian framework. Given graph $G = (A, X)$, where A is the adjacency matrix and X the node feature matrix, the final readout for node μ is a scalar $f^\mu(G;\Theta)$ which depends on the graph and network parameters Θ.

2.1 Parallel branching GNN architecture

Concretely, the overall readout $f^\mu(G;\Theta)$ for node μ is a sum of the branch readouts

$$f^\mu(G;\Theta) = \sum_{l=0}^{L-1} f_l^\mu(G;\Theta_l = \{W^{(l)}, a^{(l)}\}), \tag{1}$$

where

$$f_l^\mu(G,\Theta_l) = \frac{1}{\sqrt{L}} \sum_{i=1}^{N} \frac{1}{\sqrt{N}} a_i^{(l)} \sum_{j=1}^{N_0} \frac{1}{\sqrt{N_0}} W_{ij}^{(l)} \sum_{\nu=1}^{n} (A^l)_{\mu\nu} x_j^\nu \tag{2}$$

Note that when $L = 2$, the BPB-GNN reduces exactly to a 2-layer residual GCN (Chen et al. [2020b]).

2.2 Bayesian node regression

We consider a Bayesian semi-supervised node regression problem, for which the posterior probability for the weight parameters is given by

$$P(\Theta) = \frac{1}{Z} e^{-E(\Theta;G,Y)/T} = \frac{1}{Z} \exp(-\frac{1}{2T} \sum_{\mu=1}^{P} (f^\mu(G,\Theta) - y^\mu)^2 - \frac{1}{2\sigma_w^2} \Theta^T \Theta), \tag{3}$$

where the first term in the exponent corresponds to the likelihood term induced by learning P node labels y_μ with squared loss and the second term corresponds to the Gaussian prior with variance σ_w^2. $Z = \int e^{-E(\Theta)/T}\, d\Theta$ is the normalization constant. In the following theoretical derivations, our working regime is in the overparametrizing high dimensional limit (Li and Sompolinsky [2021], Montanari and Subag [2023], Bordelon and Pehlevan [2022], Howard et al. [2024]): $P, N, N_0 \to \infty$, $\frac{P}{N} = \alpha$ finite, and the capacity $\alpha_0 = \frac{P}{LN_0}$<1. As we will show later, this limit is practically true even with P, N not so large (our smallest N is 4). We will also use near 0 temperature in which case the training error will be near 0 and the prior L_2 regularization has an inductive bias on the solution space that will influence the generalization properties.

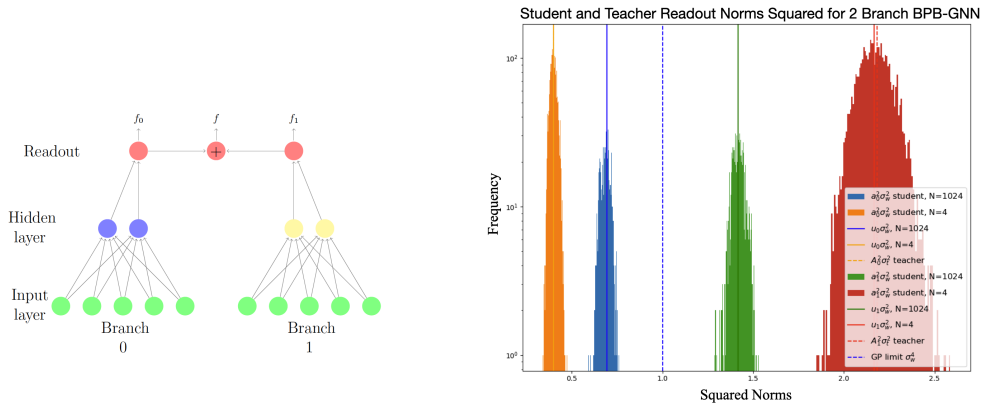
2.3 Kernel renormalization and order parameters

As shown in Appendix B.2, we can integrate out the weights in the partition function and get $Z = \int Du e^{-H(u)}$ described by a final effective Hamiltonian independent of weights

$$H(u) = S(u) + E(u), \tag{4}$$

where we call $S(u)$ the entropic term

$$S(u) = -\sum_l \frac{N}{2} \log u_l + \sum_l \frac{N}{2\sigma_w^2} u_l \tag{5}$$



(a) (b)

Figure 1: Overview of the main takeaway: BPB-GNN learns robust representations for each branch at narrow width. (a) The parallel branching GNN architecture, with 2 branches. The independent branches have non-sharing weights and produce the final output f as a sum of branch-level readouts f_l. (b) Student and teacher readout norms squared for wide and narrow student BPB-GNN networks. The student network with width N is trained with the teacher network's output. Histograms correspond to the samples from Hamiltonian Monte Carlo simulations and solid lines correspond to the order parameters calculated theoretically. $\sigma_t = \sigma_w = 1$. At $N = 4$, the HMC samples of branch readout norms squared (orange and red histograms) for the student network $\frac{\|a_l\|^2}{N}\sigma_w^2$ concentrate at their respective theoretical values $u_l\sigma_w^2$ and overlap with the teacher's readout norms squared $\frac{\|A_l\|^2}{N}\sigma_t^2$ (orange and red dashed lines) for corresponding branches. At $N = 1024$ the samples for the student network (blue and green histograms) concentrate at their respective theoretical values but remain far from the teacher's values, instead approaching the GP limit σ_w^4 (blue dashed line).

and $E(u)$ the energetic term

$$E(u) = \frac{N\alpha}{2P} Y^T (\sum_l \frac{1}{L} u_l K_l + TI)^{-1} Y + \frac{N\alpha}{2P} \log\det(\sum_l \frac{1}{L} u_l K_l + TI), \tag{6}$$

where $K_l = \frac{\sigma^2}{N_0}[A^l X X^T A^l]|_P$ is the $(P \times P)$ input node feature kernel.

Therefore, the final effective Hamiltonian has the overall kernel

$$K = \sum_l \frac{1}{L} u_l K_l, \tag{7}$$

where u_l's are order parameters which is the minimum of the effective Hamiltonian Eq. 4 by saddle point methods, which correspond to the statistical average of each branch's readout norm squared (Appendix B.4)

$$u_l = \langle \|a_l\|^2 \rangle / N \tag{8}$$

3 The narrow width limit experiments

As we discussed briefly in 2.3, the kernel becomes highly renormalized at narrow width. In fact, in the other extreme scenario when $N/P \to 0$, the energetic term in the Hamiltonian completely dominates, and we would expect that the generalization performance saturates as the order parameters in the energetic terms become independent of width N. Therefore, just as infinitely wide networks correspond to the GP limit, we propose that there exists a ***narrow width limit*** when the network width is extremely small compared to the number of training samples. We briefly showcase the narrow width limit for the student-teacher and Cora experiments, with more details discussed in the appendix.

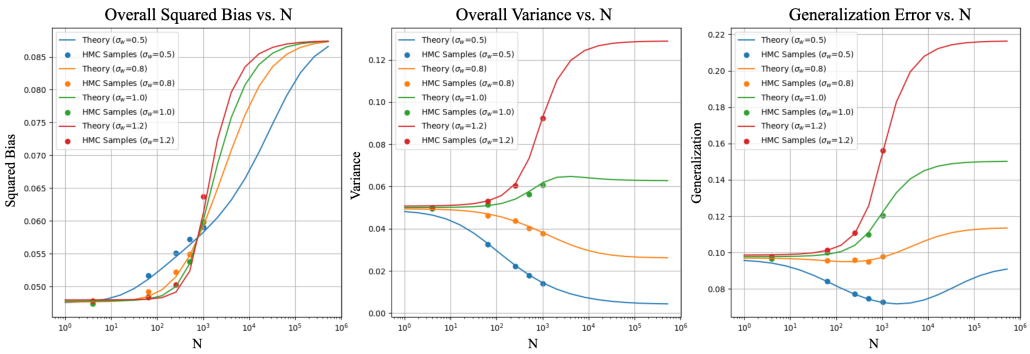


Figure 2: Student network generalization performance as a function of network width N and regularization strength σ_w. Generalization is normalized over the average true readout labels.

3.1 Student-Teacher experiment on robust branch learning

We demonstrate this robust learning phenomenon and provide a first evidence of the equipartition conjecture with the student-teacher experiment, where a student BPB-GNN network with varying hidden layer width N learns from the lables generated by a teacher network with fixed layer width (Appendix B.5,D.1).

As shown in Figure 1(b), an extremely narrow student network learns the teacher's branch readout norms very robustly, despite having a large variance; on the other hand, a much wider network fails to learn the teachers' norms and approaches the GP limit, while having a smaller variance.

Using the mean predictor and variance from the theory (Appendix B.3), we can determine the generalization error of the student network as a function of network width N, as shown in Figure 2. At narrow width, we expect individual branch to learn the teacher's branch output y_l independently, causing the bias to increase with network width. This is observed for both branches, with a transition from the narrow-width regime to the GP regime. The regularization strength σ^2 controls the transition window, with larger σ's leading to sharper transitions. This aligns with our analysis of the entropic and energetic contributions, where larger σ amplifies the distinction between the two terms. In contrast, the variance decreases with network width for small σ_w's, resulting in a trade-off between the contributions of bias and variance to overall generalization performance.

4 BPB-GNN on Cora

We also perform experiments on the Cora benchmark dataset by training the BPB-GNN with binary node regression, for a range of L, N, σ_w values (Appendix D.2). We observe a similar narrow-to-wide width transition for the bias term. As shown in Figure 3, the bias increases with network width, transitioning to the GP regime, and we observe the trend extending to a potential narrow width limit.[1] Additionally, it is demonstrated that using more branches that involve higher-order convolutions improves performance.

5 Conclusion

In conclusion, this paper introduces and investigates the concept of narrow width limits in Bayesian Parallel Branching Graph Neural Networks. Contrary to the common belief that wider networks inherently generalize better, our results indicate that BPB-GNNs with significantly narrower widths can achieve better or competitive performance. This is attributed to effective symmetry breaking and kernel renormalization in the narrow-width limit, which lead to robust learning.

[1] In this case, the narrow width limit is hard to demonstrate as the transition window is below realistic minimum of network width.

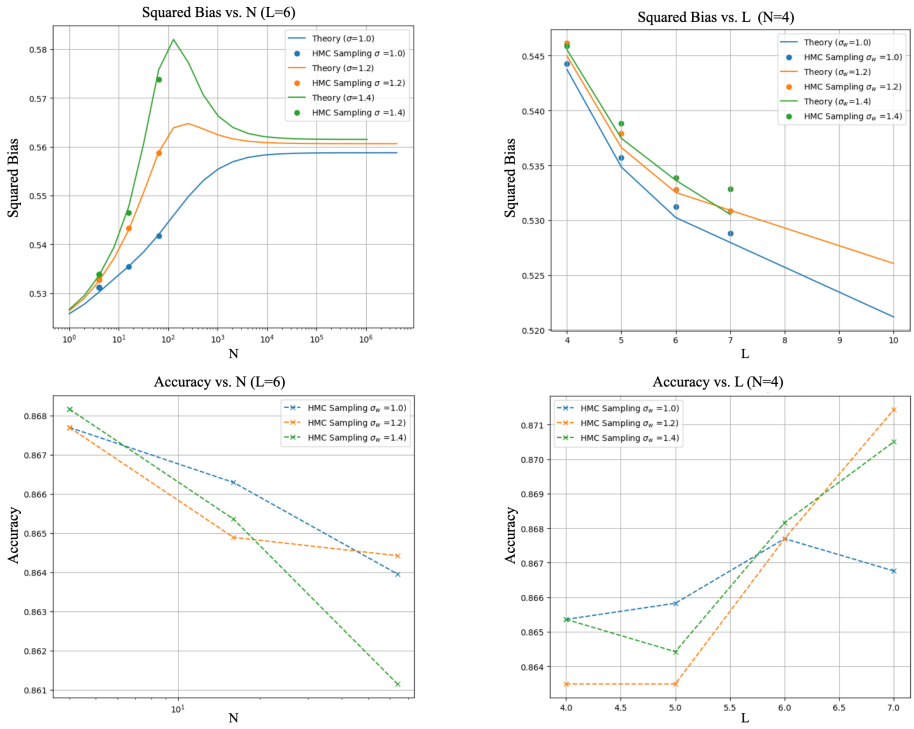


Figure 3: Cora generalization performance vs. network width N and branch number L, for various regularization strength σ_w's. The accuracy is computed by turning the mean predictor from HMC samples into a class label using its sign.

Acknowledgments and Disclosure of Funding

We acknowledge support of the Swartz Foundation, the Kempner Institute for the Study of Natural and Artificial Intelligence at Harvard University and the Gatsby Charitable Foundation. This material is partially based upon work supported by the Center for Brains, Minds and Machines (CBMM), funded by NSF STC award CCF-1231216. We have benefitted from helpful discussions with Alexander van Meegen, Lorenzo Tiberi and Qianyi Li.

References

Zeyuan Allen-Zhu, Yuanzhi Li, and Yingyu Liang. Learning and generalization in overparameterized neural networks, going beyond two layers. *Advances in neural information processing systems*, 32, 2019.

Gholamali Aminian, Yixuan He, Gesine Reinert, Łukasz Szpruch, and Samuel N Cohen. Generalization error of graph neural networks in the mean-field regime. *arXiv preprint arXiv:2402.07025*, 2024.

Alexander Atanasov, Blake Bordelon, and Cengiz Pehlevan. Neural networks as kernel learners: The silent alignment effect. *arXiv preprint arXiv:2111.00034*, 2021.

Yehonatan Avidan, Qianyi Li, and Haim Sompolinsky. Connecting ntk and nngp: A unified theoretical framework for neural network learning dynamics in the kernel regime. *arXiv preprint arXiv:2309.04522*, 2023.

Yasaman Bahri, Boris Hanin, Antonin Brossollet, Vittorio Erba, Christian Keup, Rosalba Pacelli, and James B. Simon. Les houches lectures on deep learning at large and infinite width, 2024.

Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. *Advances in Neural Information Processing Systems*, 35:32240–32256, 2022.

Lei Chen, Le Wu, Richang Hong, Kun Zhang, and Meng Wang. Revisiting graph based collaborative filtering: A linear residual graph convolutional network approach. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 27–34, 2020a.

Ming Chen, Zhewei Wei, Zengfeng Huang, Bolin Ding, and Yaliang Li. Simple and deep graph convolutional networks. In Hal Daumé III and Aarti Singh, editors, *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pages 1725–1735. PMLR, 13–18 Jul 2020b. URL `https://proceedings.mlr.press/v119/chen20v.html`.

Rong Chen, Li Guanghui, and Chenglong Dai. Feature fusion via deep residual graph convolutional network for hyperspectral image classification. *IEEE Geoscience and Remote Sensing Letters*, 19: 1–5, 2022.

Weilin Cong, Morteza Ramezani, and Mehrdad Mahdavi. On provable benefits of depth in training graph convolutional networks. *Advances in Neural Information Processing Systems*, 34:9936–9949, 2021.

Yash Deshpande, Subhabrata Sen, Andrea Montanari, and Elchanan Mossel. Contextual stochastic block models. *Advances in Neural Information Processing Systems*, 31, 2018.

Simon S Du, Kangcheng Hou, Russ R Salakhutdinov, Barnabas Poczos, Ruosong Wang, and Keyulu Xu. Graph neural tangent kernel: Fusing graph neural networks with graph kernels. *Advances in neural information processing systems*, 32, 2019.

Tianxiang Gao, Xiaokai Huo, Hailiang Liu, and Hongyang Gao. Wide neural networks as gaussian processes: Lessons from deep equilibrium models. *Advances in Neural Information Processing Systems*, 36, 2024.

Vikas Garg, Stefanie Jegelka, and Tommi Jaakkola. Generalization and representational limits of graph neural networks. In *International Conference on Machine Learning*, pages 3419–3430. PMLR, 2020.

Floris Geerts and Juan L Reutter. Expressiveness and approximation properties of graph neural networks. *arXiv preprint arXiv:2204.04661*, 2022.

Jessica N Howard, Ro Jefferson, Anindita Maiti, and Zohar Ringel. Wilsonian renormalization of neural network gaussian processes. *arXiv preprint arXiv:2405.06008*, 2024.

Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

Jaehoon Lee, Yasaman Bahri, Roman Novak, Samuel S. Schoenholz, Jeffrey Pennington, and Jascha Sohl-Dickstein. Deep neural networks as gaussian processes. 2018.

Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Phys. Rev. X*, 11:031059, Sep 2021. doi: 10.1103/PhysRevX.11.031059. URL `https://link.aps.org/doi/10.1103/PhysRevX.11.031059`.

Andrea Montanari and Eliran Subag. Solving overparametrized systems of random equations: I. model and algorithms for approximate solutions. 2023.

Daniel Park, Jascha Sohl-Dickstein, Quoc Le, and Samuel Smith. The effect of network width on stochastic gradient descent and generalization: an empirical study. In *International Conference on Machine Learning*, pages 5042–5051. PMLR, 2019.

Huayi Tang and Yong Liu. Towards understanding generalization of graph neural networks. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett, editors, *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pages 33674–33719. PMLR, 23–29 Jul 2023. URL `https://proceedings.mlr.press/v202/tang23f.html`.

Ian Walker and Ben Glocker. Graph convolutional gaussian processes. In *International Conference on Machine Learning*, pages 6495–6504. PMLR, 2019.

Zihan Wang and Arthur Jacot. Implicit bias of sgd in l_{2}-regularized linear dnns: One-way jumps from high to low rank. *arXiv preprint arXiv:2305.16038*, 2023.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? *arXiv preprint arXiv:1810.00826*, 2018.

A Related works

Infinitely wide neural networks: Our work follows a long tradition of mathematical analysis of infinitely-wide neural networks (**?**Jacot et al. [2018], Lee et al. [2018], Bahri et al. [2024]), resulting in NTK or NNGP kernels. Recently, such analysis has been extended to structured neural networks, including GNNs (Du et al. [2019], Walker and Glocker [2019]). However, they do not provide an analysis of feature learning in which the kernel depends on the tasks.

Kernel renormalization and feature learning: There has been progress in understanding simple MLPs in the feature-learning regime as the shape of the kernel changes with task or time (Li and Sompolinsky [2021], Atanasov et al. [2021], Avidan et al. [2023], Wang and Jacot [2023]). We develop such understanding in graph-based networks.

Theoretical analysis of GCN: There is a long line of works that theoretically analyze the expressiveness (Xu et al. [2018], Geerts and Reutter [2022]) and generalization performance (Tang and Liu [2023], Garg et al. [2020], Aminian et al. [2024]) of GNN. However, it is challenging to calculate the dependence of generalization errors on tasks. To our knowledge, our work is first to provide a tight bound of the generalization error for GNN with residual-like structures. The architecture closest to our linear BPB-GCN is the linearly decoupled GCN proposed by Cong et al. [2021]; however, the overall readout vector is shared for all branches, which will not result in kernel renormalization for different branches.

B Details on Theory of BPB-GNN

B.1 Summary of Notations

Hyperparameters and Dimensions	
P	Number of training nodes
n	Total number of nodes for a graph
N_0	Input node feature dimension
N	BPB-GNN hidden layer width
L	Total number of branches
σ_w	L_2 prior regularization strength
T	Temperature
$\alpha_0 = \dfrac{P}{LN_0}$	Network capacity
$\alpha = \dfrac{P}{N}$	Width ratio
Network Architecture and Input/Output	

$\hat{A} \in \mathbb{R}^{n \times n}$	Adjacency matrix
$A \in \mathbb{R}^{n \times n}$	Normalized adjacency matrix by its degree matrix D
$X \in \mathbb{R}^{n \times N_0}$	Input node feature matrix
$G = (X, A)$	Graph
$W^{(l)} \in \mathbb{R}^{N_0 \times N}$	Hidden layer weight for branch l
$a^{(l)}, a_l \in \mathbb{R}^N$	Readout vector for branch l
$\Theta_l = (W_l, a_l)$	Collection of parameters for branch l
$h_l^\mu \in \mathbb{R}^N$	Activation vector for branch l and node μ
$f_l^\mu \in \mathbb{R}$	Readout prediction for branch l and node μ
$f^\mu \in \mathbb{R}$	Overall readout prediction for node μ
$y^\mu \in \mathbb{R}$	Overall node label for node μ
$H_l(W_l) \in \mathbb{R}^{P \times N}$	Activation feature matrix for branch l
$Y \in \mathbb{R}^P$	Training node labels
$F \in \mathbb{R}^P$	Readout predictions

Statistical Theory

$E(\Theta; G, Y)$	Energy loss function
Z	Partition function
$I \in \mathbb{R}^{P \times P}$	Identity matrix
$H(W)$	Hamiltonian after integrating out readout a_l's
$u \in \mathbb{R}^L$	Order parameters as saddle point solution
$u_l \in \mathbb{R}$	Order parameter for branch l
$H(u)$	Hamiltonian as a function of order parameters
$S(u)$	Entropy as a function of order parameters
$E(u)$	Energy as a function of order parameters
r_l	Mean squared readout
Tr_l	Variance-related of readout

Kernels

$K(W) \in \mathbb{R}^{P \times P}$	Hidden layer weight dependent overall kernel
K_l	Branch l kernel
$K \in \mathbb{R}^{P \times P}$	Overall kernel averaged over W's, $K = \sum_l \frac{u_l K_l}{L}$
$k_l^\nu \in \mathbb{R}^{P \times 1}$	Branch l kernel column for the P training nodes against test node ν
$k^\nu \in \mathbb{R}^{P \times 1}$	Overall kernel column for the P training nodes against test node ν
$\vert_P$	Kernel restricted to the P training nodes against P training nodes
$\vert_{(P,\nu)}$	Kernel restricted to P training nodes against the test node ν
$\vert_{(\nu,\nu)}$	Kernel restricted to the test node ν against test node ν

Student-teacher Setup

$Y_l^* \in \mathbb{R}^P$	Teacher network readout prediction for branch l for P nodes
$Y^* \in \mathbb{R}^P$	Teacher network overall readout for P nodes
$W_l^* \in \mathbb{R}^{N_0 \times N}$	Teacher hidden layer weight for branch l
$A_l^* \in \mathbb{R}^N$	Teacher readout vector for branch l
β_l^2	Teacher readout variance
σ_t^2	Teacher hidden layer weight variance

B.2 Kernel renormalization

Following similar derivations as the first kernel renormalization work Li and Sompolinsky [2021], we will integrate out the weights in the partition function $Z = \int d\theta \exp(-E(\Theta)/T)$, from the readout layer weights a_l's to the hidden layer weights W_l's and arrive at an effective Hamiltonian shown in the main text.

First, we linearize the energy in terms of a_l's by introducing the auxiliary variables $t^\mu, \mu = 1, \ldots, P$.

$$Z = \int d\Theta \int \prod_{\mu=1}^{P} dt_\mu \exp\left[-\frac{1}{2\sigma_w^2}\Theta^\top \Theta - \sum_{\mu=1}^{P} i t_\mu \left(\frac{1}{\sqrt{LN}} \sum_{i=1}^{N} \sum_{l=0}^{L-1} a_i^{(l)} h_i^\mu(G) - Y^\mu \right)^2 - \frac{T}{2} t^\top t \right] \tag{9}$$

Now we can integrate out a_l's as they are linearized and the partition function becomes

$$Z = \int DW e^{-H(W)}, \tag{10}$$

with effective Hamiltonian

$$H(W) = \frac{1}{2\sigma_w^2} \sum_{l=0}^{L-1} \mathrm{Tr} W_l^T W_l + \frac{1}{2} Y^T (K(W) + TI)^{-1} Y + \frac{1}{2} \log\det(K(W) + TI), \tag{11}$$

where

$$K(W) = \frac{1}{L} \sum_l \frac{\sigma_w^2}{N} (H_l(W_l) H_l(W_l)^T)|_P \tag{12}$$

is the $P \times P$ kernel matrix dependent on the observed P nodes with node features $H_l = A^l X W_l$ and denote $|_P$ as the matrix restricting to the elements generated by the training nodes.

Now we perform the integration on W_l's, and get a Fourier representation of Z with h_l,u_l as auxiliary variables after inserting t:

$$\begin{aligned} Z &= \int \prod_{l=0}^{L} dh_l du_l dt \exp\Big(i t^T Y - \sum_l \frac{N}{2} \log(1 + h_l) + \sum_l \frac{N}{2\sigma_w^2} u_l h_l - \frac{1}{2} t^T \Big(\sum_l \frac{1}{L} u_l K_l + TI \Big) t \Big) \\ &= \int \prod_{l=0}^{L} dh_l du_l \exp\Big(- \sum_l \frac{N}{2} \log(1 + h_l) + \sum_l \frac{N}{2\sigma_w^2} u_l h_l \frac{1}{2} Y^T \Big(\sum_l \frac{1}{L} u_l K_l + TI \Big)^{-1} Y \Big) \end{aligned} \tag{13}$$

where

$$K_l = \frac{\sigma_w^2}{N_0} [A^l X X^T A^{lT}]|_P \tag{14}$$

is the input kernel for branch l. Now as $N \to \infty$ and $\alpha = \frac{P}{N}$ fixed, we can perform the saddle point approximation and get the saddle points for h_l as

$$1 + h_l = \frac{\sigma_w^2}{u_l} \tag{15}$$

Plugging this back to the equation, we get

$$Z = \int \Pi_l \, du_l e^{-H_{eff}(u)}, \tag{16}$$

with the effective Hamiltonian

$$H_{eff}(u) = S(u) + E(u), \tag{17}$$

where we call $S(u)$ the entropic term

$$S(u) = -\sum_l \frac{N}{2} \log u_l + \sum_l \frac{N}{2\sigma_w^2} u_l \tag{18}$$

and $E(u)$ the energetic term

$$E(u) = \frac{1}{2} Y^T (\sum_l \frac{1}{L} u_l K_l + TI)^{-1} Y + \frac{1}{2} \log\det(\sum_l \frac{1}{L} u_l K_l + TI) \tag{19}$$

Therefore, after integrating out W_l, the effective kernel is given by

$$K = \sum_l \frac{1}{L} u_l K_l, \tag{20}$$

where K_l is

$$K_l = \frac{\sigma_w^2}{N_0} [A^l X X^T A^l]|_P \tag{21}$$

And the saddle point equations for u_l's are determined by

$$N(1 - \frac{u_l}{\sigma_w^2}) = -Y^T (K + TI)^{-1} \frac{u_l K_l}{L} (K + TI)^{-1} Y + \mathrm{Tr}[K^{-1} \frac{u_l K_l}{L}], \tag{22}$$

where we call

$$r_l = Y^T (K + TI)^{-1} \frac{u_l K_l}{L} (K + TI)^{-1} Y \tag{23}$$

and

$$\mathrm{Tr}_l = \mathrm{Tr}[K^{-1} \frac{u_l K_l}{L}] \tag{24}$$

As we will show later, these represent the mean and variance of the readout norm squared respectively. In the $T = 0$ case, the saddle point equation becomes

$$N(1 - \frac{u_l}{\sigma_w^2}) = -Y^T K^{-1} \frac{u_l K_l}{L} K^{-1} Y + \mathrm{Tr}[K^{-1} \frac{u_l K_l}{L}] \tag{25}$$

B.3 Predictor statistics and generalization

We can get the predictor statistics of each branch readout $y_l^\nu(G)$ on a new test node ν by considering the generating function:

$$\begin{aligned} Z(\eta_1, \ldots, \eta_L) = \int D\Theta \exp\Bigg\{ & -\frac{\beta}{2} \sum_\mu (f^\mu(G;\Theta) - y^\mu)^2 \\ & + \sum_l i\eta_l \frac{1}{\sqrt{NL}} \sum_i a_i^{(l)} h_i^{(l),\nu}(G, W_l) - \frac{T}{2\sigma_w^2} \Theta^T \Theta \Bigg\} \end{aligned} \tag{26}$$

Therefore, by taking the derivative with respect to each η_l, we arrive at the statistics for $y_l(x)$ as:

$$\langle f_l^\nu(G) \rangle = \partial_{i\eta_l} \log Z \big|_{\vec{\eta}=0} \tag{27}$$

$$\langle \delta f_{l,\nu}^2(G) \rangle = \partial_{i\eta_l}^2 \log Z \big|_{\vec{\eta}=0} \tag{28}$$

After integrating out the weights Θ layer by layer, we have:

$$Z(\eta_1,\ldots,\eta_L) = \int \Pi_l du_l \, \exp\left\{\sum_l \left(\frac{N}{2}\log u_l - \frac{N}{2\sigma_w^2}u_l\right) + \frac{1}{2}(iY + \sum_l \frac{1}{L}\eta_l u_l k_l^\nu)^T (\sum_l \frac{1}{L}u_l K_l + TI)^{-1}(iY + \sum_l \eta_l \frac{1}{L}u_l k_l^\nu - \frac{1}{2}\log\det(\sum_l \frac{1}{L}u_l K_l + TI) - \frac{1}{2}\sum_l \eta_l^2 \frac{1}{L}K_l^{\nu,\nu}\right\}. \tag{29}$$

Here

$$k_l^\nu = \frac{\sigma_w^2}{N_0}[A^l X X^T A^l]|_{(P,\nu)} \tag{30}$$

is the $P \times 1$ column kernel matrix for test node ν and all training nodes, and

$$K_l^{\nu,\nu} = \frac{\sigma_w^2}{N_0}[A^l X X^T A^l]|_{(\nu,\nu)} \tag{31}$$

is the single matrix element for the test node. Therefore, eventually, we have:

$$\langle f_l^\nu \rangle = \frac{u_l k_{l,\nu}^T}{L}(K + TI)^{-1} Y \tag{32}$$

and

$$\langle \delta f_{l,\nu}^2 \rangle = \frac{u_l K_l^{\nu,\nu}}{L} - \frac{u_l k_{l,\nu}^T}{L}(K + TI)^{-1}\frac{u_l k_{l,\nu}}{L} \tag{33}$$

The predictor statistics of the overall readout $f = \sum_l f_l$ is given by:

$$\langle f^\nu(G)\rangle = \sum_l \frac{u_l k_{l,\nu}^T}{L}(K + TI)^{-1} Y = k_\nu^T (K + TI)^{-1} Y \tag{34}$$

$$\langle \delta f(G)_\nu^2 \rangle = \sum_l u_l K_l^{\nu,\nu} - \sum_{l,l'} u_l k_{l,\nu}^T (K + TI)^{-1} u_{l'} k_{l',\nu} = K_{\nu,\nu} - k_\nu^T (K + TI)^{-1} k_\nu \tag{35}$$

B.4 Statistics of branch readout norms

From the partition function Eq.9, we can relate the mean of readout weights a_l to the auxiliary variable t by

$$\langle a_l \rangle_W = -i\frac{\sigma_w^2}{\sqrt{N}}\Phi_l^T \langle t\rangle = -\frac{\sigma_w^2}{\sqrt{NL}}\Phi_l^T (K + TI)^{-1} Y, \tag{36}$$

where Φ_l is the node feature matrix for the hidden layer nodes. We have

$$\langle a_l^T \rangle \langle a_l \rangle = \sigma_w^2 Y^T (K + TI)^{-1}\frac{u_l K_l}{L}(K + TI)^{-1} Y = r_l \sigma_w^2 \tag{37}$$

We can calculate the second-order statistics of a_l: the variance is

$$\langle \delta a_l^T \delta a_l \rangle = \sigma_w^2 \mathrm{Tr}(I + \frac{\sigma_w^2 \beta}{NL}\Phi_l \Phi_l^T)^{-1} = \sigma_w^2 (N - \mathrm{Tr}(K + TI)^{-1}\frac{u_l K_l}{L}) = \sigma_w^2 (N - \mathrm{Tr}_l) \tag{38}$$

Therefore,

$$\langle a_l^2 \rangle = \langle \delta a_l^T \delta a_l \rangle + \langle \delta a_l^T \delta a_l \rangle = N\sigma_w^2 + \sigma_w^2 r_l + 1 - \sigma_w^2 \mathrm{Tr}_l = N u_l \tag{39}$$

Therefore, we have proved the main text claim that the order parameter u_l's are really the mean squared readout norms of the branches.

B.5 Robust learning of branches: the equipartition conjecture

What happens in the narrow width limit? In the following, we demonstrate that each branch will learn robustly at narrow width.

The equipartition conjecture Consider a student-teacher network setup, where the teacher network is given by

$$f^*(G;\Theta^*) = \sum_l f_l^*(G;W_l^*) = \sum_l \frac{1}{\sqrt{N_t L}} \sum_i A_{i,l} h_i^l(G;W_l^*). \tag{40}$$

$W_{ij}^* \sim \mathcal{N}(0,\sigma_t^2)$ and $A_{i,l} \sim \mathcal{N}(0,\beta_l^2)$, where β_l^2 is the variance assigned to the readout weight for the teacher branch l and N_t is the width of the hidden layer. Similarly, the student network is given by the same architecture, with layer width N and learns from P node lables from the teacher $Y_{\mu=1}^{*P} = f^*(G,\Theta^*)_\mu$ in the Bayesian regression setup of Eq. 3 with prior variance σ_w^2. We conjecture that as $\alpha = P/N \to \infty$ the posterior distribution of the student network readout vector a_l satisfies

$$\sigma_w^2 \langle \|a_l\|^2 \rangle = \sigma_t^2 \|A_l\|^2. \tag{41}$$

Sketch of proof:

At narrow width, the saddle equation 22 becomes $r_l = \mathrm{Tr}_l$. Now consider $(Y^* Y^{*T})_{\mu,\nu} = \sum_{i,j,l_1,l_2} \frac{1}{N_t L} A_{i,l_1} A_{j,l_2} h_i^{l_1,\mu}(G) h_j^{l_2,\nu}(G)$; we conjecture that this quantity concentrates at its expectation value $Y^* Y^{*T} \approx \mathbb{E}_{a^*,W^*}(YY^T) = \sum_l \beta_l^2 K_l/L$. Given this assumption, r_l becomes

$$r_l = Y^{*T} K^{-1} \frac{u_l K_l}{L} K^{-1} Y^* = \mathrm{Tr}(K^{-1} \frac{u_l K_l}{L} Y^* Y^{*T}) \approx \mathrm{Tr}_l(K^{-1} \sum_l \beta_l^2 \frac{K_l}{L}) \tag{42}$$

The solution that satisfies the saddle point equations is

$$u_l \sigma_w^2 = \beta_l^2 \sigma_t^2. \tag{43}$$

Therefore, by Eq.8, we proved the conjecture given N_t is also large enough. We call this equipartition conjecture, as the mean-squared readout and the variance (B.4) have to exactly balance each other, which contribute to the energy term in the Hamiltonian. This is only a conjecture, as it relies on the concentration equality assumption made in the proof. However, as we demonstrate in the experiments, the equipartition conjecture holds empirically.

Furthermore, writing $u_l K_l/L \approx Y_l^* Y^{*T}$, at narrow width, the predictor statistics for each stream for the training data from the teacher becomes

$$\langle f_l(X) \rangle = \frac{u_l K_l}{L} (K + TI)^{-1} Y^* \approx Y_l^*, \tag{44}$$

ie. not only do we recover the statistics for the teacher A_l's, we have also recovered the feature learned by each branch.

C Further results

C.1 Symmetry Breaking and Convergence of Branches

We perform theoretical calculations and HMC sampling for the student branch squared readout norms as we vary the student network width N and the prior regularization strength σ_w. Figure 4(a) shows the statistical average of the student branch squared readout norms, ie. $\langle \|a_l\|^2 \rangle \sigma_w^2/N$ as a function of the network width N, where the branch norms split as the network width gets smaller, which we call symmetry breaking. The symmetry breaking of branch norms from the GP limit to the narrow width limit accompanies the convergence to learning teacher's norms at narrow width for different σ_w's as shown in Figure 4(b)(c), supporting Eq. 41.

In addition, we show the bias and variance for individual branches of the student network in Figure 5.

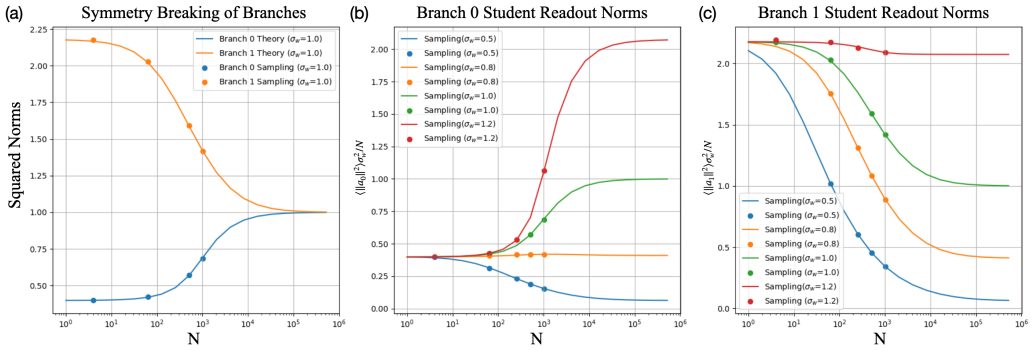


Figure 4: Statistical average of student readout norms squared as a function of network width from theory and HMC sampling, for student-teacher tasks described in Section 3.1. (a): $\langle \|a_l\|^2 \rangle \sigma_w^2 / N$ as a function of network width N for a fixed σ_w. The branch norms break the GP symmetry as it goes to the narrow width limit. (b)(c): Branch 0 and branch 1 readout norm squared respectively for a range of σ_w regularization values. The student branch norms with different regularization strengths all converge to the same teacher readout norm values at narrow width.

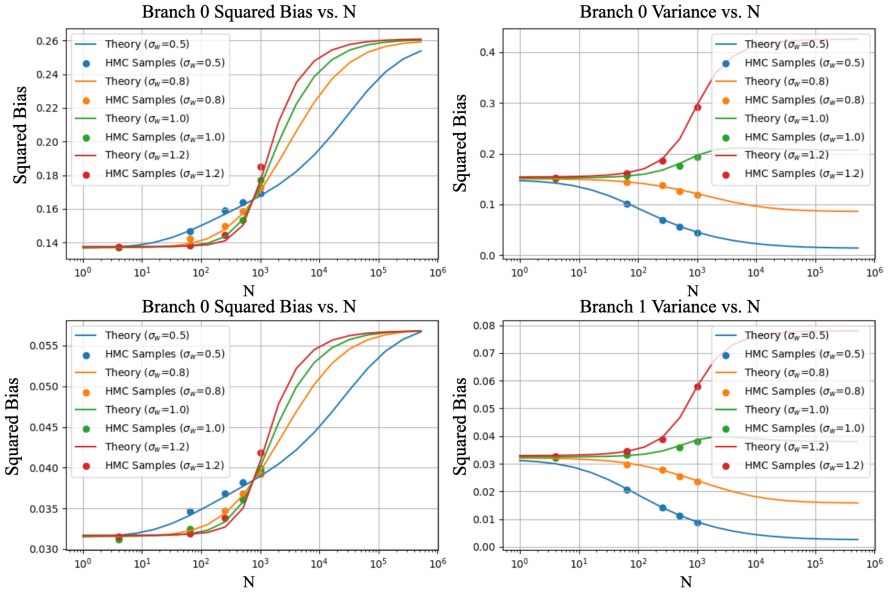


Figure 5: Student network squared bias and variance for individual branches as a function of network width N and regularization strength σ_w. The mean and variance of branch l readout f_l^μ for node μ is calculated in B.3. Generalization values are normalized over the average true readout labels.

C.2 Convergence of Branch Importance at Narrow Width

An interesting aspect of the BPB-GNN network is that the branch readout norms converge at the narrow width for different hyperparameters σ_w and L, reflecting the natural branch importance for the task.

As shown in Figure 6, the BPB-GNN with branches $L = 6$ robustly learns the readout norms at narrow width independently of σ_w' s, consistent with the student-teacher results. This suggests that we can recast the data as generated from a ground-truth teacher network even for real-world datasets. The last branch of the BPB-GNN network has a larger contribution, reflecting the presence of higher-order convolutions in the Cora dataset. From a kernel perspective, increasing branches better

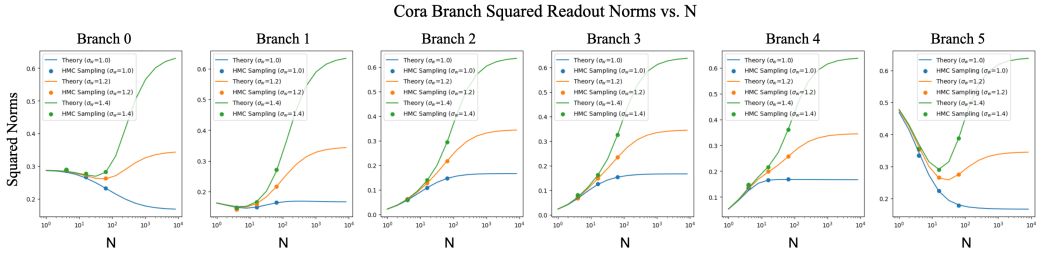


Figure 6: Cora experiment: statistical average of squared readout norms $\langle \|a_l\|^2 \rangle \sigma_w^2 / N$ for each branch l as a function of the network width N, and regularization strength σ_w. The BPB-GNN has $L = 6$ branches.

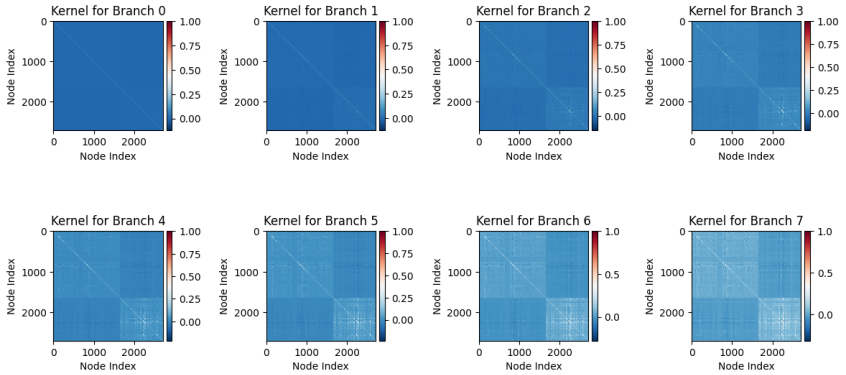


Figure 7: Kernel $K_l = \frac{\sigma_w^2}{N_0} A^l X_0 X_0^T A^l$ for each branch l shown for the first 8 branches on Cora dataset, sorted by node labels. A is the normalized adjacency matrix of the Cora graph, X_0 the node feature matrix and N_0 the node feature dimension. Initialization variance $\sigma_w^2 = 1$ and total node number $n = 2708$.

distinguish the nodes, as shown in Figure 7. This could explain the selective turn-off of intermediate branches and the increased contribution of the last branch.

Furthermore, the first two branches are learned most robustly at narrow width, as shown in Figure 8, where the branch norms converge for the first two branches even for BPB-GNNs with different L. This suggests that the branch importance, as reflected by the norms learned at narrow width, indicates the contribution of the bare data and the first convolution layer.

D Experimental details

D.1 Student-teacher CSBM

For the student-teacher task, we use the contextual stochastic block model introduced by Deshpande et al. [2018] to generate the graph G. The adjacency matrix is given by

$$A_{ij} = \begin{cases} 1 & \text{with probability } p = c_{in}/n, \text{ if } i,j \leq n/2 \\ 1 & \text{with probability } p = c_{in}/n, \text{ if } i,j \geq n/2 \\ 1 & \text{with probability } q = c_{out}/n, \text{ otherwise} \end{cases} \tag{45}$$

where

$$c_{in,out} = d \pm \sqrt{d}\lambda \tag{46}$$

d is the average degree and λ the homophily factor.

The feature vector $\vec{x}^\mu$ for a particular node μ is given by

$$\vec{x}_\mu = \sqrt{\frac{\mu}{n}} y^\mu \vec{u} + \vec{\xi}_\mu, \tag{47}$$

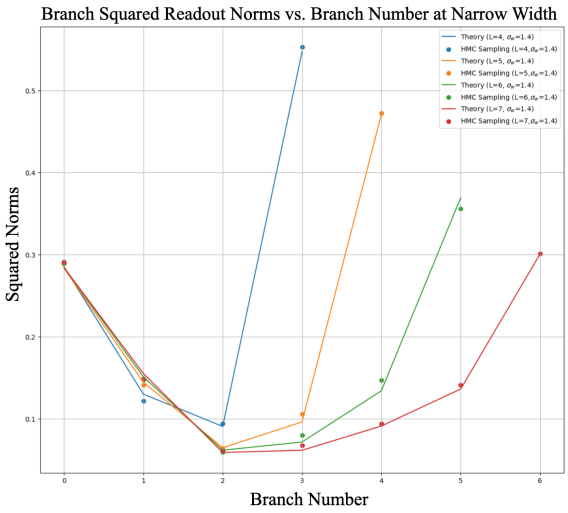


Figure 8: Branch Importance vs. branch number l on Cora. Legends represent different total number of branches L. The branch importance is defined as the statistical average of branch squared readout norms $\langle \|a_l\|^2 \rangle \sigma_w^2 / N$ at the narrow width limit; here we take the empirical branch norm values at $N = 4$ and fixed $\sigma_w = 1.4$.

where

$$\vec{u} \sim \mathcal{N}(0, I_{N_0}), \vec{\xi}_\mu \sim \mathcal{N}(0, I_{N_0}) \tag{48}$$

In the experiment, we use $N_0 = 950$,$d = 20$,$\lambda = 4$ and $\mu = 4$. The teacher network parameters are variance $\sigma_t^2 = 1$, width $N_t = 1024$, branch norms variance $\beta_0^2 = 0.4$, $\beta_1^2 = 2$ for individual element of the readout vector a_l. Temperature $T = 0.0005\sigma_w^2$ for each σ_w value.

D.2 Cora

For the Cora dataset, we use a random split of the data into 21% as training set and 79% as test set. We group the classes $(1, 2, 4)$ into one group and the rest for the other group for binary node regression, with labels as ± 1's. The Bayesian theory and HMC sampling follows the same design as in the student-teacher setup. We use temperature $T = 0.01$ for both theory and sampling as the sampling becomes more difficult for smaller temperature. This explains the discrepancy of the GP limit bias for different σ_w values.

D.3 Hamiltonian Monte Carlo

The sampling experiments in the paper are all done with Hamiltonian Monte Carlo simulations, a popular method for sampling a probability distribution. HMC has faster convergence to the posterior distribution compared to Langevin dynamics. We used Numpyrho to set up chains and run the simulations on the GPU cluster. Due to memory constraint, we only sampled up to $N = 1024$ hidden layer width for the student-teacher CSBM experiment and $N = 64$ for the Cora experiment. Since we mainly aim to demonstrate the narrow width effect in this paper, this suffices the purpose.

E Debunking Challenge Submission

E.1 What commonly-held position or belief are you challenging?

Provide a short summary of the body of work challenged by your results. Good summaries should outline the state of the literature and be reasonable, e.g. the people working in this area will agree with your overview. You can cite sources beside published work (e.g., blogs, talks, etc).

There is a common heuristics/belief that wider network generally does better both in the Bayesian setting and in DNNs trained with SGD ((Jacot et al. [2018], Lee et al. [2018], Bahri et al. [2024]).

E.2 How are your results in tension with this commonly-held position?

Detail how your submission challenges the belief described in (1). You may cite or synthesize results (e.g. figures, derivations, etc) from the main body of your submission and/or the literature.

We characterize a new regime that we call the narrow width limit in the Bayesian setting, such that when each kernel from parallel branches is sufficiently different, the kernel is re-normalized by an order parameter u that depends on the data and task. We demonstrate empirically that the overall bias is smaller at the narrow width limit (Figure 23), an "inverse scaling law" if you will.

E.3 How do you expect your submission to affect future work?

We expect and hope more work to explore the narrow width limit; in particular, the attention network resembles the BPB-GNN with different attention heads as the parallel branches. There will be a narrow width limit for the attention network in the Bayesian setting, but more interestingly we would like to see if they hold with SGD or Adam training.