When Narrower is Better

The Narrow Width Limit of Bayesian Parallel Branching Neural Networks

Problem Setup

Branching Neural Networks

- Architecture: Sum of L independent branches
$$f(x;\Theta) = \sum_l \frac{1}{\sqrt{NL}} a_l^\top \phi_l(x;W_l)$$
- Examples: **GCN:** Branch l uses convolutions of graph A^l; **Residual-MLP:** ϕ_0 linear, ϕ_1 ReLU.

Bayesian Regression

- Posterior $P(\Theta|Y) \propto \exp\big(-\frac{1}{2T}\sum_{\mu=1}^{P}(f^\mu - y^\mu)^2 - \frac{1}{2\sigma_w^2}\Theta^T\Theta\big)$.
- Likelihood (Loss) + Prior (L_2 Reg. σ_w^2). T: Temperature.
- Regime: Overparameterized $P, N \to \infty$, $\alpha = P/N$ finite.
- Equilibrium $\approx$ an ensemble of trained NNs over random initializations

Method: Kernel Renormalization

Partition function $Z = \int e^{-E(\Theta)/T}\, d\Theta$ contains all statistics. Integrating out weights a_l, then W_l leads to an effective theory described by order parameters u_l:

- **GP Limit ($\alpha \to 0$, infinite width):**
  - $u_l \to \sigma_w^2$ (Symmetric branches).
  - Kernel $K_{GP} = \sum_l \frac{\sigma_w^2}{L} K_l$ (Task-independent, NNGP).
- **Finite α (Feature Learning / Narrow Width):**
  - u_l depend on data Y via saddle-point equations: $N(1 - u_l/\sigma_w^2) = -r_l + \mathrm{Tr}_l$.
  - Kernel $K = \sum_l \frac{u_l}{L} K_l$ is **renormalized** by data.
  - Leads to **symmetry breaking** (u_l differ based on relevance).

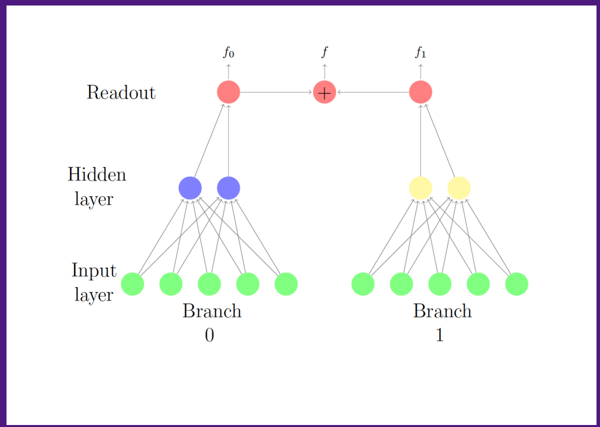
Main Theorem: The Narrow Width Limit

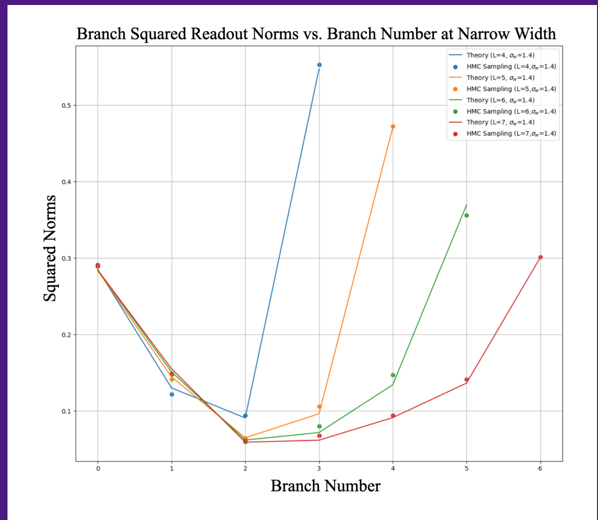
Student network (width N, prior σ_w^2) learns from the teacher network (branch norms $\beta_l^2\sigma_*^2$). Student's order parameters u_l converge to match teacher's scaled norms at the narrow limit ($P/N = \alpha \to \infty$):

$$u_l\sigma_w^2 \xrightarrow{\alpha\to\infty} \beta_l^2\sigma_*^2$$

$(\implies \langle\|a_l\|^2\rangle\sigma_w^2/N \to \|a_l^*\|^2\rangle\sigma_*^2/N)$.

Heard of the infinite width limit? We discovered the ***narrow width limit***, where the **narrower** the better. We prove a general theorem and demonstrate for GCN and residual-MLP.





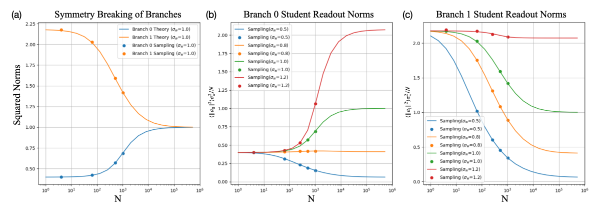
FIRST AUTHOR ZECHEN ZHANG

SECOND AUTHOR HAIM SOMPOLINSKY

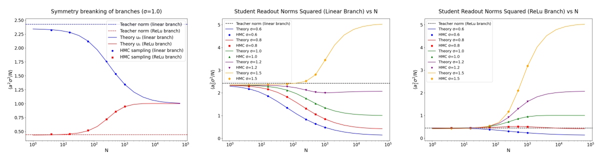
Main Results

Symmetry Breaking & Robust Learning

GCN

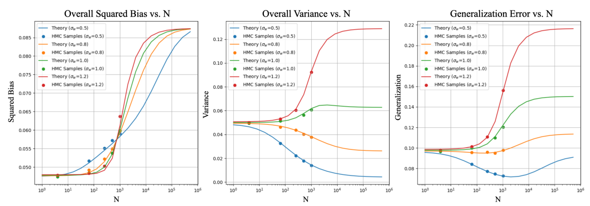


Residual-MLP

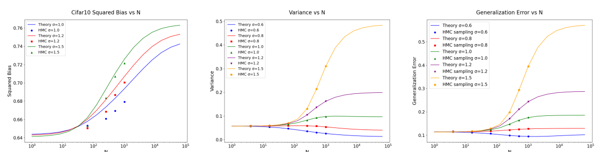


Generalization vs. Width

GCN

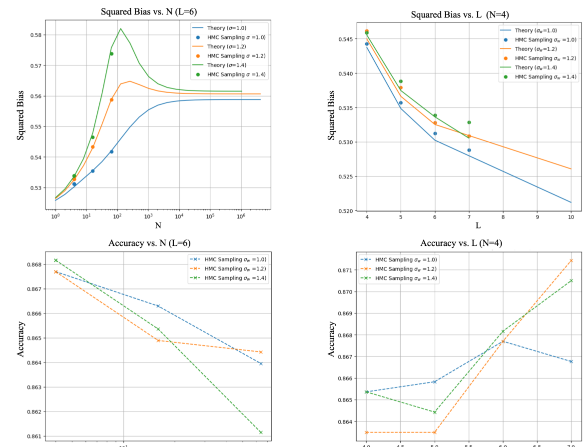


Residual-MLP



Real Dataset Results (Cora & Cifar10)

Cora



Cifar 10