Analyzing Tree Architectures in Ensembles via Neural Tangent Kernel

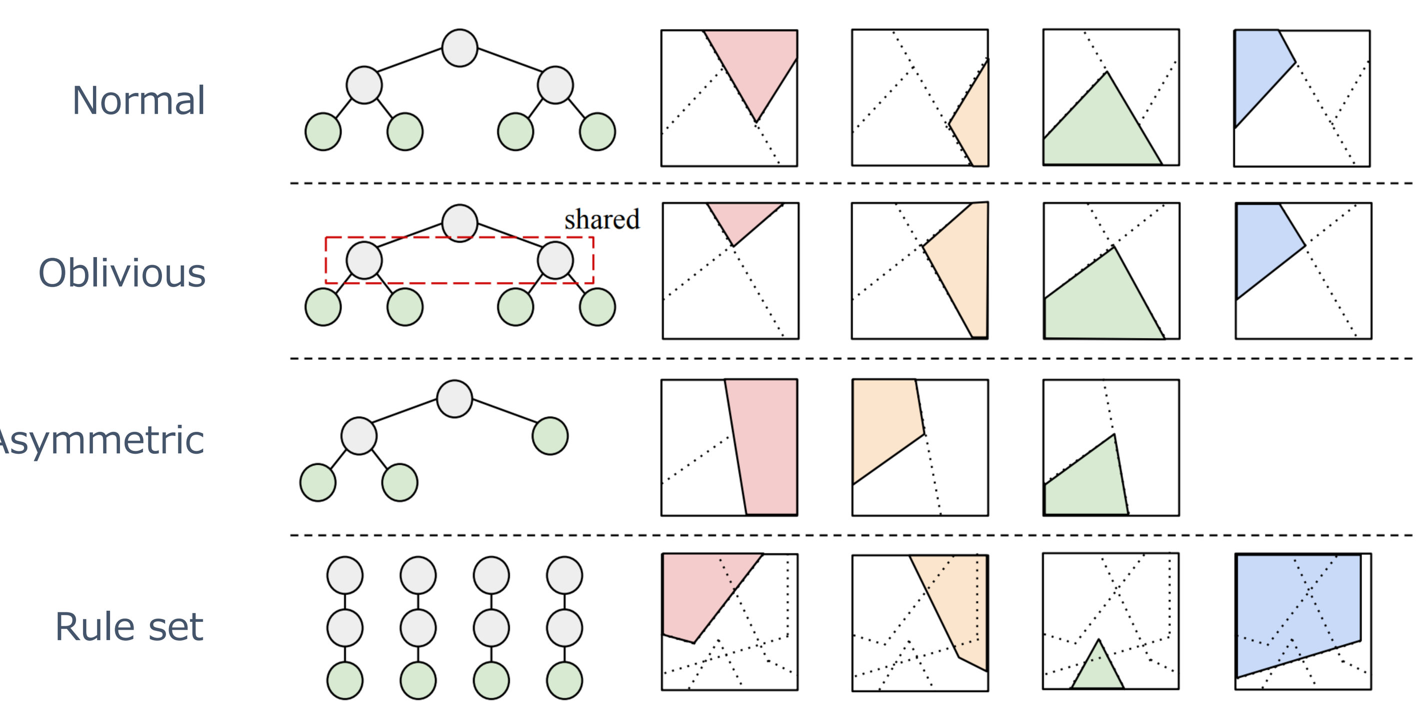
Ryuichi Kanoh[1,2], Mahito Sugiyama[1,2]

[1]National Institute of Informatics, [2]The Graduate University for Advanced Studies, SOKENDAI

The 11th International Conference on Learning Representations (ICLR 2023), May 1-5, 2023

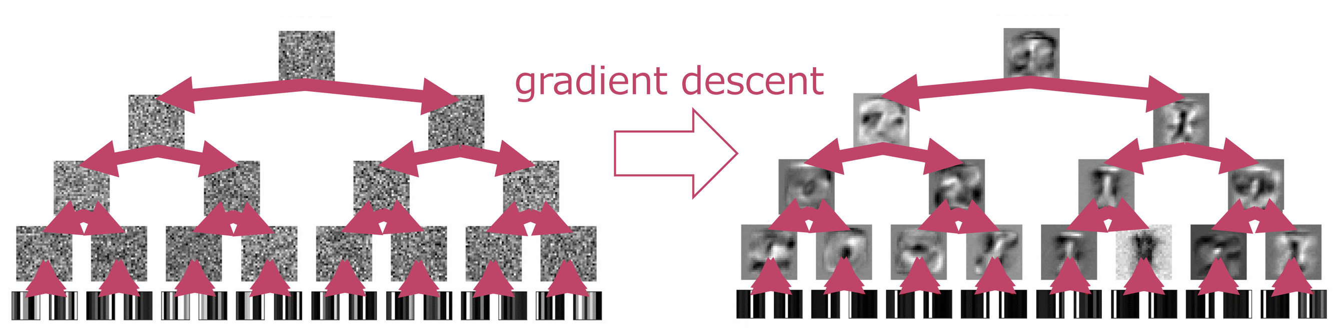
Contribution

- **We formulate and analyze the Neural Tangent Kernel (NTK) induced by soft tree ensembles for arbitrary tree architectures**



Soft Tree Ensemble

- **A variant of trees that inherits characteristics of neural networks**
  - Splitting rules and leaf values are updated with gradient descent
  - Unlike typical decision trees, feature engineering is included in training



- **The NTK for ensembles of perfect binary trees is known**
  - It converges to a closed kernel when we consider infinite trees (M→∞)

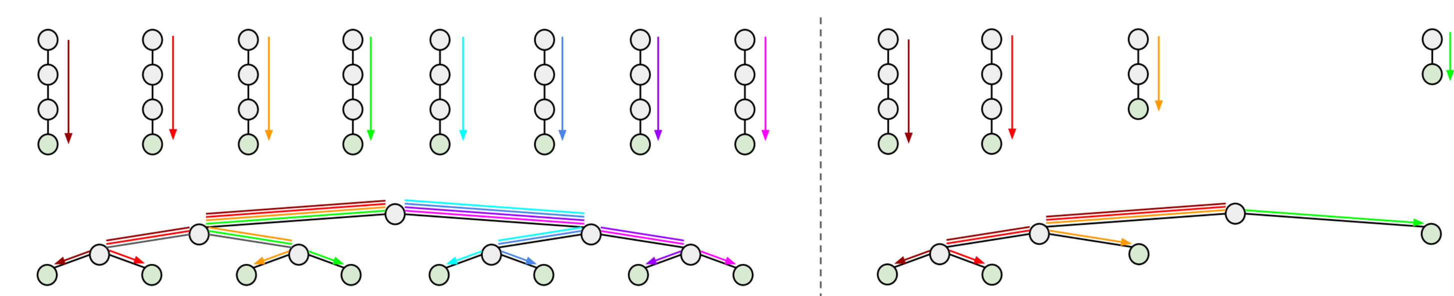
$$\Theta^{(D,\mathrm{PB})}(\boldsymbol{x}_i,\boldsymbol{x}_j) \coloneqq \lim_{M\to\infty} \widehat{\Theta}_0^{(D,\mathrm{PB})}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \underbrace{2^D D\,\Sigma(\boldsymbol{x}_i,\boldsymbol{x}_j)(\mathcal{T}(\boldsymbol{x}_i,\boldsymbol{x}_j))^{D-1}\dot{\mathcal{T}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}_{\text{contribution from internal nodes}} + \underbrace{(2\mathcal{T}(\boldsymbol{x}_i,\boldsymbol{x}_j))^D}_{\text{contribution from leaves}}$$

Motivation

- **Previous studies could handle only perfect binary trees**
- **Theoretical understanding of other types of widely used soft trees (e.g., asymmetric tree, rule set) has not been developed yet**

NTK for Arbitrary Tree Architectures

- **The NTK induced by infinite ensembles of arbitrary trees can be decomposed by the NTKs induced by rule ensembles**
  - Characterized only by the number of tree leaves per depth



NTK for infinite ensembles of rule sets

$$\Theta^{(D,\mathrm{Rule})}(\boldsymbol{x}_i,\boldsymbol{x}_j) \coloneqq \lim_{M\to\infty} \widehat{\Theta}_0^{(D,\mathrm{Rule})}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \underbrace{D\,\Sigma(\boldsymbol{x}_i,\boldsymbol{x}_j)(\mathcal{T}(\boldsymbol{x}_i,\boldsymbol{x}_j))^{D-1}\dot{\mathcal{T}}(\boldsymbol{x}_i,\boldsymbol{x}_j)}_{\text{contribution from internal nodes}} + \underbrace{(\mathcal{T}(\boldsymbol{x}_i,\boldsymbol{x}_j))^D}_{\text{contribution from leaves}}$$

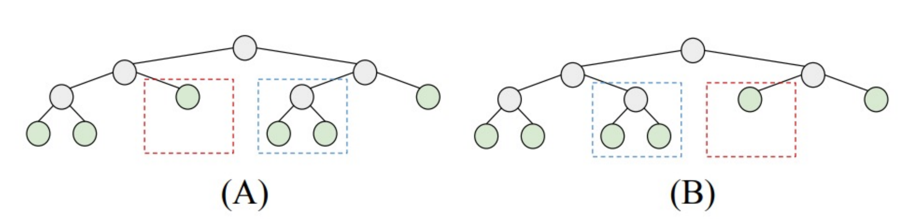
NTK for infinite ensembles of arbitrary trees

$$\Theta^{(\mathrm{ArbitraryTree})}(\boldsymbol{x}_i,\boldsymbol{x}_j) \coloneqq \lim_{M\to\infty} \widehat{\Theta}_0^{(\mathrm{ArbitraryTree})}(\boldsymbol{x}_i,\boldsymbol{x}_j) = \sum_{d=1}^{D} \mathcal{Q}(d)\Theta^{(d,\,\mathrm{Rule})}(\boldsymbol{x}_i,\boldsymbol{x}_j)$$

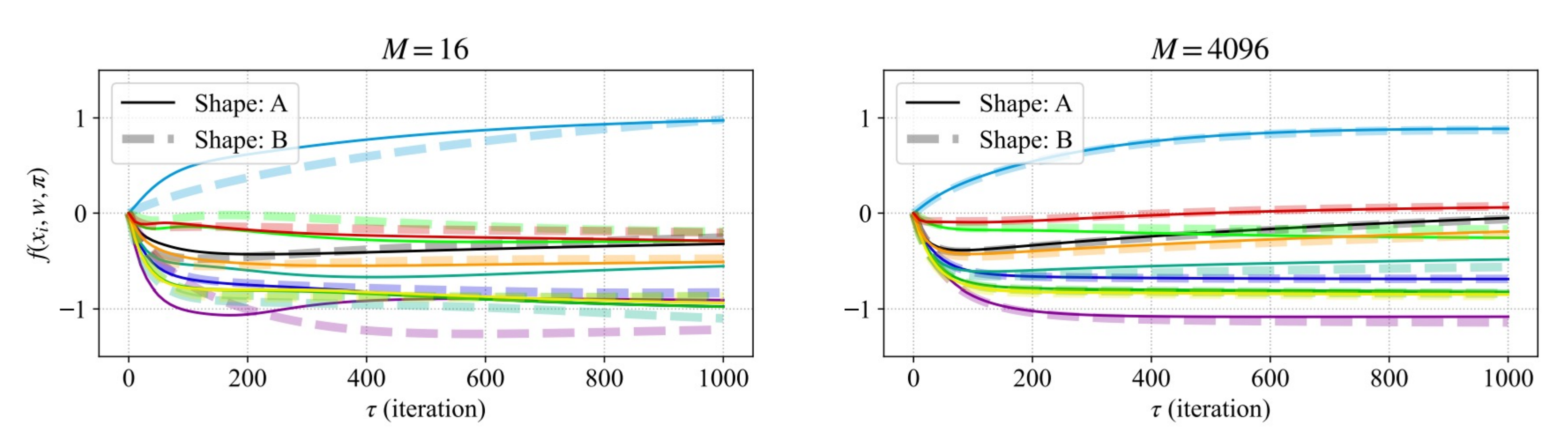
Q(d): The number of leaves per depth at d

Tree Equivalence

- **Training behavior can be equivalent even for non-isomorphic trees**
  - The NTK depends only on the number of leaves per depth

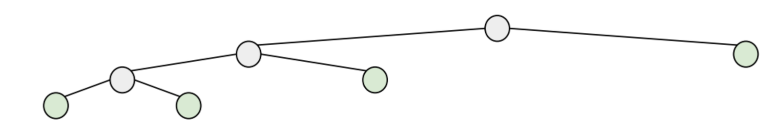


Non-isomorphic tree architectures used in ensembles that induce the same limiting NTK



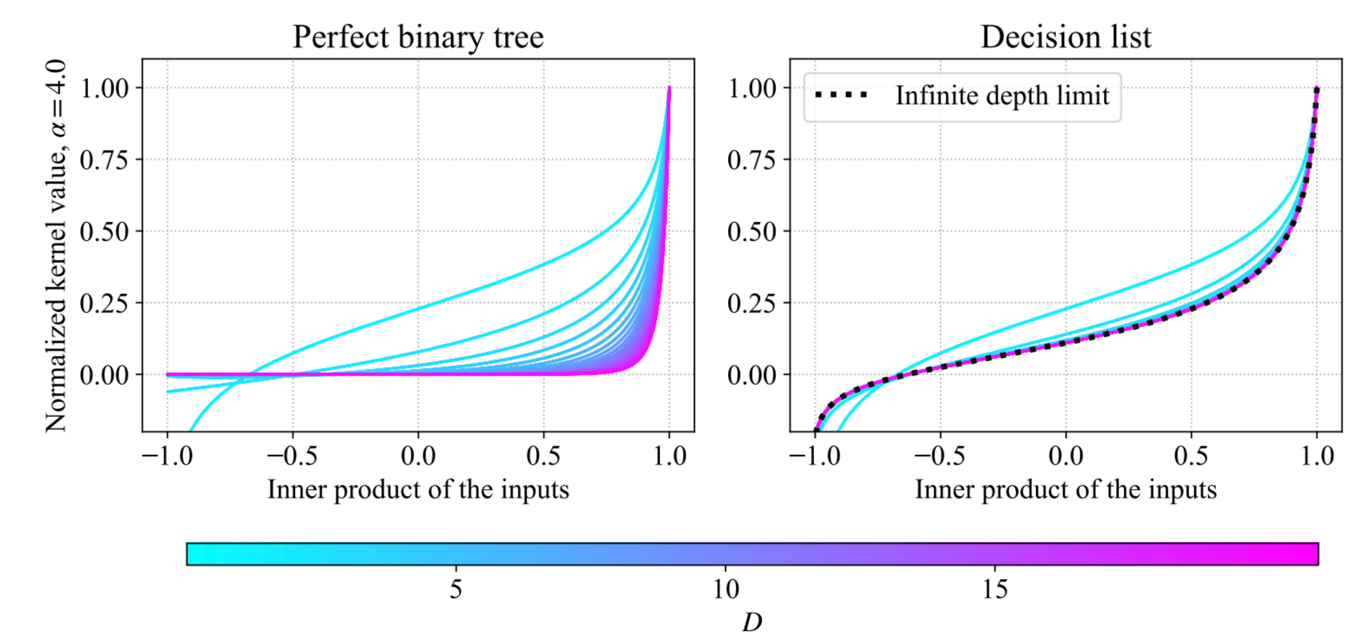
Output dynamics for test data points. Each line color corresponds to each data point

Case Study: Decision List

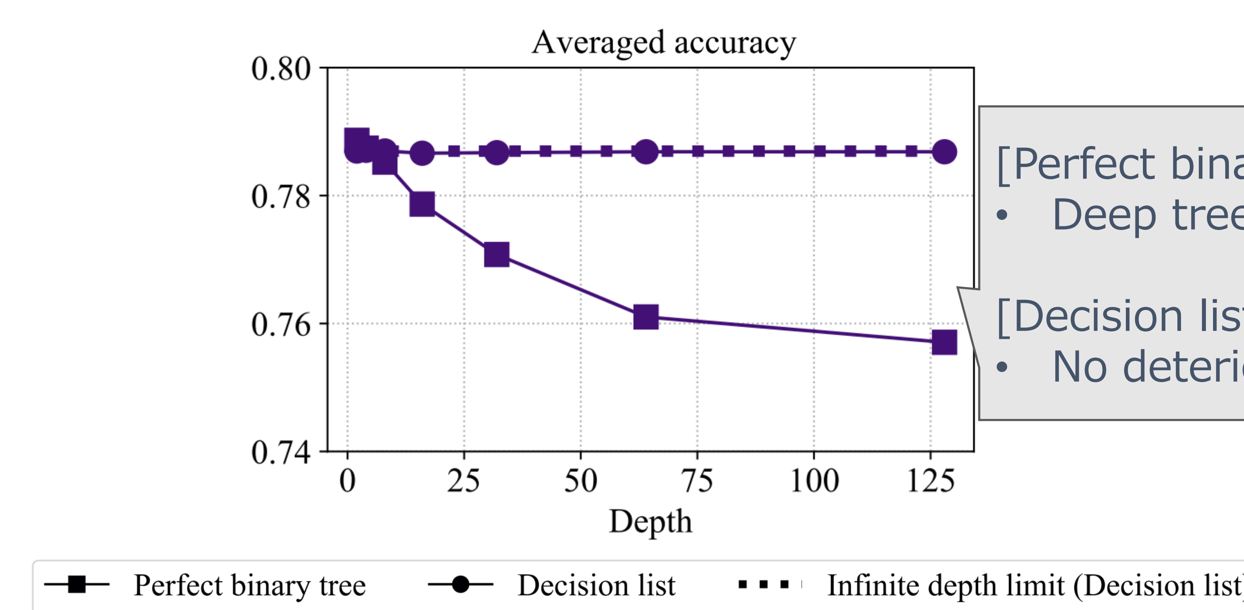


Decision list: a binary tree that grows in only one direction

- **The NTK for infinitely deep decision lists does not degenerate**
  - This is in contrast to the perfect binary trees



- **The degeneracy leads to worse generalization performance**
  - Unable to distinguish between a 90-degree and a 180-degree differences



[Perfect binary tree]
- Deep trees do not generalize well

[Decision list]
- No deterioration

Reference

[Neural Tangent Kernel]
- Jacot et al. (NeurIPS 2018), Neural Tangent Kernel: Convergence and Generalization in Neural Networks
- Kanoh&Sugiyama (ICLR 2022), A Neural Tangent Kernel Perspective of Infinite Tree Ensembles

[Soft Tree]
- Kontschieder et al. (ICCV 2015), Deep Neural Decision Forests
- Popov et al. (ICLR 2020), Neural Oblivious Decision Ensembles for Deep Learning on Tabular Data
- Hazimeh et al. (ICML 2020), The Tree Ensemble Layer: Differentiability meets Conditional Computation