Depth is More Powerful than Width with Prediction Concatenation in Deep Forest

Shen-Huan Lyu, Yi-Xiao He, Zhi-Hua Zhou
National Key Laboratory for Novel Software Technology,
Nanjing University, Nanjing, 210023, China.
{lvsh,heyx,zhouzh}@lamda.nju.edu.cn

Abstract

Random Forest (RF) is an ensemble learning algorithm proposed by Breiman [1] that constructs a large number of randomized decision trees individually and aggregates their predictions by naive averaging. Zhou and Feng [2] further propose Deep Forest (DF) algorithm with multi-layer feature transformation, which significantly outperforms random forest in various application fields. The prediction concatenation (PreConc) operation is crucial for the multi-layer feature transformation in deep forest, though little has been known about its theoretical property. In this paper, we analyze the influence of Preconc on the consistency of deep forest. Especially when the individual tree is inconsistent (as in practice, the individual tree is often set to be fully grown, i.e., there is only one sample at each leaf node), we find that the convergence rate of two-layer DF *w.r.t.* the number of trees M can reach $\mathcal{O}(1/M^2)$ under some mild conditions, while the convergence rate of RF is $\mathcal{O}(1/M)$. Therefore, with the help of PreConc, DF with deeper layer will be more powerful than the shallower layer. Experiments confirm theoretical advantages.

1 Introduction

Random forest (RF) [1] is a state-of-art classification and regression algorithm that constructs a number of randomized individual decision trees during a parallel training phase and predicts by naive averaging the results. Recently, Zhou and Feng [2] propose Deep Forest (DF) algorithm with multi-layer feature transformation to investigate the possibility of tree-based representation learning. Benefiting from the multi-layer feature transformation in the cascade structure with prediction concatenation (PreConc), DF outperforms various tree-based algorithms [1, 3–5] in empirical study, and have been involved in real applications such as medicine [6], computer vision [7], remote sensing [8], financial fraud detection [9], etc. Numerous variants have been extended to various tasks and meet with remarkable success in [10–13]. There are also variants [14, 15] aiming at improving performance and reducing computational cost.

Empirical successes have attracted attention to the theoretical analysis of DF. Since its PreConc operation transforms features layer by layer, and each layer consists of complex non-parametric random forest estimators, analyzing the impact of the new features extracted from these estimators on generalization performance is important. To start theoretical analysis on DF, Lyu et al. [16] prove a margin-based generalization bound for additive new features in DF. Then, Pang et al. [14] prove an upper bound on the sample efficiency and inspire an efficient improvement for reducing computational cost of DF. In the direction of consistency, Arnould et al. [17] provide tight lower and upper bounds on the excess risk of a shallow centered random tree network, which leverages the new features to improve the performance of a single centered random tree. Assuming that the original and new features are separated and independently used in two stages, Lyu et al. [18] prove that the new features based on prediction are easy to cause overfitting risk, and propose to use the interaction of decision rules to alleviate it.

36th Conference on Neural Information Processing Systems (NeurIPS 2022).

However, previous theoretical studies often choose to ignore the impact of the new features on generating forests. For example, recent works [16, 18, 14] assume that the forest estimator in each layer is a black-box model and Arnould et al. [17] assume that there is only a single centered random tree in each layer. Such strong assumptions are in favor of theoretical analysis, but widen the gap between the theoretical results and practical deep forest architecture. To properly analyze DF, we must notice that, among the complex architecture in DF, both prediction concatenation (PreConc) [16] and the classification and regression trees (CART)-split criterion [19] play critical roles. Therefore, in order to open the black box of DF, we need to study the subtle combination of different techniques, i.e., analyze the properties of the new feature and study how CART generates and utilizes it.

Contributions. We compare the advantages of depth (cascade structure with PreConc in deep forest) over width (bagging of trees in random forest) in the scope of $\mathbb{L}_2^2$-consistency, based on the assumptions of the additive regression functions and uniform distribution over input space. The main contributions can be summarized as follows:

- We are the first to establish a consistency analysis of the prediction concatenation (PreConc) operation which is crucial for multi-layer feature transformation in deep forest, though based on a simplified version.
- We prove the universal consistency of deep forest when the total number of leaves of individual trees is chosen properly.
- In the practical setting, when the individual trees are fully grown, we prove that the convergence rate of two-layer deep forest reach $\mathcal{O}(1/M^2)$ *w.r.t.* the increasing number of trees M, while that of random forest is $\mathcal{O}(1/M)$. This result reflects that deep forest with deeper layer will be more powerful than shallower layer.

Organization. The rest of this work is organized as follows: Section 3 shows the setting and notations related to tree-based estimators in this work. Section 4 recalls the original deep forest architecture and simplifies it into a two-layer deep forest. Section 5 contains the properties of cascade structure with prediction concatenation. Section 6 proves the main result that depth is more powerful than width in deep forest architecture in the scope of consistency. Section 7 is devoted to the empirical studies of deep forest by verifying the theoretical results above. Section 8 concludes with future work. More experimental results and detailed proofs for theorems and propositions are given in the supplementary material due to page limitation.

2 Related work

Despite the widespread use and remarkable success of random forest in real world applications, the theoretical properties of it are still not fully understood [20, 21]. Breiman [1] offers an upper bound on the generalization error of random forest in terms of correlation and accuracy of the individual trees. Lin and Jeon [22] establish a connection between random forest and a particular class of nearest neighbor predictors, which are further studied by Biau and Devroye [23]. Meinshausen [24] studies the consistency of the quantile random forest for regression. In recent years, various theoretical works [20, 25–28] have been performed, analyzing the consistency of various simplified forests, and moving ever closer to practice. Denil et al. [21] narrows the gap between theory and practice of random forests for regression. Scornet et al. [29] prove the first $\mathbb{L}_2^2$-consistency of Breiman's original random forest with CART-split criterion based on the assumptions of the additive regression functions and uniform distribution over input space. Scornet [30] prove that infinite forest consistency implies finite forest consistency. Gao and Zhou [31] then present the convergence rate of purely randomized trees and a simplified variant of Breiman's original CART trees. Gao et al. [32] further expand it to multi-class setting. In addition, another research route is the theory analysis of feature importance [33–35]. Recently, Li et al. [36] derive non-asymptotic lower and upper bounds on the expected bias of MDI importance for random forests. Sutera et al. [37] establish a connection between MDI importance of pure random forest and Shapley values.

However, while deep forests further improve generalization performance, there is no theory to prove the advantages brought by depth. Therefore, studying the influence of depth is the theoretical cornerstone for distinguishing deep forests from random forests. For example, in the well-known deep neural networks (DNNs), there are a lot of theoretical works to study the effect of depth and width on its representation ability and generalization performance, which show the theoretical advantages of deep neural networks over shallow neural networks [38–42]. These works all contribute to the understanding of deep learning and provide insight for designing algorithms.

3 Setting and notations

We first describe the setting and notations related to tree-based estimators in this work. For the sake of conciseness, we consider the regression setting.

Setting. We consider a regression framework, where the training set $S_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$ consists of $[0,1]^d \times \mathbb{R}$-valued independent random variables distributed as the prototype sample $(\boldsymbol{x}, y)$ with $\mathbb{E}[y^2] < \infty$. This underlying distribution, characterized by the marginal distribution $\mathcal{D}_{\mathcal{X}}$ on $[0,1]^d$ and by the conditional distribution $\mathcal{D}_{\mathcal{Y}|\mathcal{X}}$, can be written as

$$y = f(\boldsymbol{x}) + \epsilon \,, \tag{1}$$

where $f(\boldsymbol{x}) = \mathbb{E}[y|\boldsymbol{x}]$ is the conditional expectation of y given $\boldsymbol{x}$, and ϵ is a noise satisfying $\mathbb{E}[\epsilon] = 0$ and $\mathrm{Var}[\epsilon] < \sigma^2$. The task considered in this paper is to output a randomized estimator $h_n(\cdot, \Theta, S_n)\colon [0,1]^d \to \mathbb{R}$ where Θ is a random variable that accounts for the randomization procedure and independent of the training set S_n. To simplify notation, we denote $h_n(\boldsymbol{x}, \Theta) = h_n(\boldsymbol{x}, \Theta, S_n)$. The quality of a randomized estimator h_n is measured by its $\mathbb{L}_2$ risk

$$R(h_n) = \mathbb{E}\left[\left(h_n(\boldsymbol{x}, \Theta) - f(\boldsymbol{x})\right)^2\right] \,, \tag{2}$$

where the expectation is taken with respect to $(\boldsymbol{x}, \Theta)$, conditionally on S_n. As the training data size n increases, we get a sequence of estimators $\{h_1, h_2, \ldots, h_n, \ldots\}$. A sequence of estimators $\{h_n\}_{n=1}^{\infty}$ is said to be consistent if $R(h_n) \to 0$ as $n \to \infty$.

Trees and forests. A random forest estimator $h_{M,n}(\boldsymbol{x}, \Theta)$ outputs the average prediction over M individual randomized trees $h_n(\boldsymbol{x}, \Theta_j)$, $\forall j \in [M]$. Here, $[M] = \{1, 2, \ldots, M\}$ denotes the indexes of all individual randomized trees, where $\Theta_1, \ldots, \Theta_M$ are distributed identically and independently and denoted by a generic random variable Θ. The random variable Θ can be used to sample the training set and select the candidate dimensions and positions for splitting. Specifically, a recursive partition Π of $[0,1]^d$ is built by performing successive axis-aligned splits according to Θ:

$$h_n(\boldsymbol{x}, \Theta) = \sum_{i=1}^{n} \frac{y_i \cdot \mathbb{1}(\boldsymbol{x}_i \in C_{\Pi,n}(\boldsymbol{x}))}{N_n(C_{\Pi,n}(\boldsymbol{x}))} \,, \tag{3}$$

where $C_{\Pi,n}(\boldsymbol{x})$ is the cell of the tree partition containing $\boldsymbol{x}$ and $N_n(C_{\Pi,n}(\boldsymbol{x}))$ is the number of training samples falling into $C_{\Pi,n}(\boldsymbol{x})$ with convention that the estimation equals to zero if the cell $C_{\Pi,n}(\boldsymbol{x})$ is empty. These trees are combined to form a finite forest estimation:

$$h_{M,n}(\boldsymbol{x}, \Theta) = h_{M,n}(\boldsymbol{x}, \Theta_1, \ldots, \Theta_M) = \frac{1}{M} \sum_{i=1}^{M} h_n(\boldsymbol{x}, \Theta_i) \,. \tag{4}$$

By the law of large numbers, for any fixed $\boldsymbol{x}$, conditionally on $\mathcal{D}_n$, the finite forest estimation converges to the infinite forest estimation:

$$h_{\infty,n}(\boldsymbol{x}) = \lim_{M \to \infty} h_{M,n}(\boldsymbol{x}, \Theta) = \mathbb{E}_{\Theta}[h_n(\boldsymbol{x}, \Theta)] \,. \tag{5}$$

When the number of samples tends to ∞, we denote by $h_{\infty}(\boldsymbol{x}, \Theta)$ the randomized tree with infinite samples and $h_{M,\infty}(\boldsymbol{x}, \Theta)$ the random forest with M trees and infinite samples.

4 Deep forest

We recall the original deep forest algorithm in Section 4.1, and describe the simplified two-layer deep forest algorithm in Section 4.2.

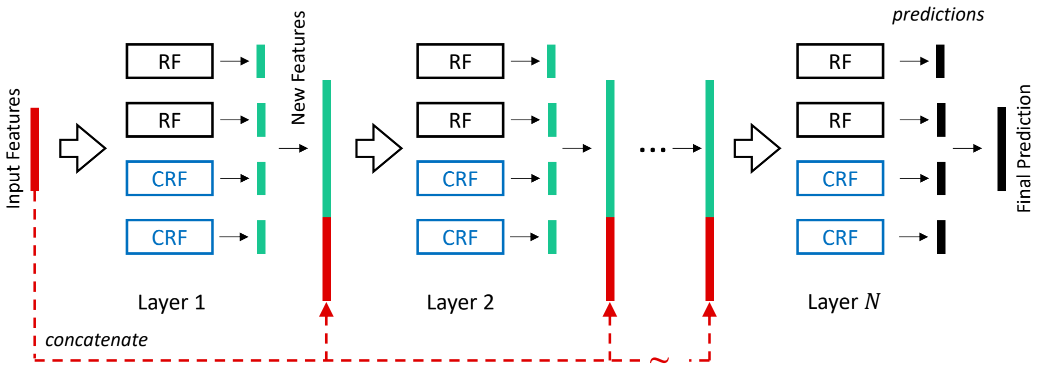


Figure 1: Deep forest architecture (the scheme is taken from Zhou and Feng [2]).

4.1 Original deep forest

Deep forest [2] is a tree-based deep model made up of non-differentiable forest modules without back-propagation. Each layer of DF is composed of two Breiman's random forest (RF) and two Completely-Random Forest (CRF). The RF is composed of CARTs and CRF is composed of pure random trees. In the regression setting, each forest of each layer outputs a prediction for any query point $\boldsymbol{x}$, corresponding to the average response in the leaf node (cell) containing $\boldsymbol{x}$. At a given layer, the predictions of all forests of this layer are concatenated together with raw features. This prediction concatenation process is called PreConc, which is repeated for each layer until the best layer and construct a deep forest. For an overview of the architecture of deep forest, we refer readers to the work [43, 2].

4.2 Simplified two-layer deep forest

In order to narrow the gap between theoretical and empirical studies of deep forest, we retain CART-split criterion and PreConc when simplifying the model. We define a simplified two-layer deep forest, whose each layer is composed of one Breiman's random forest. We denote by $h^{(1)}_{M,n}(\boldsymbol{x}, \Theta, S_n)$ the first-layer forest estimator and $h^{(2)}_{M,n}([\boldsymbol{x}, h], \Theta, S_n)$ the second-layer forest estimator, where $[\boldsymbol{x}, h]$ is the concatenation of the raw features $\boldsymbol{x}$ and the new feature h. Then we can define the two-layer deep forest as follows:

$$\bar{h}_{2M,n}(\boldsymbol{x}, \Theta, S_n) = h^{(2)}_{M,n} \circ h^{(1)}_{M,n}(\boldsymbol{x}, \Theta, S_n)\ . \tag{6}$$

The complete algorithm is shown in Algorithm 1. This algorithm has three parameters: $m_{\text{try}} \in \{1, \ldots, d\}$ is the number of pre-selected directions for splitting, $a_n \in \{1, \ldots, n\}$ is the number of samples in each tree, $t_n \in \{1, \ldots, a_n\}$ is the number of leaves in each tree. In the default procedure, the parameters are set as follows: m_{try} is set to d. a_n is set to n/k, where k is the k-fold cross-validation in Zhou and Feng [2]'s deep forest. $t_n = a_n$ means we use fully grown CART. Notice that k controls the subsampling rate $a_n/n = 1/k$, which is proved by Scornet et al. [29] to be the key component for imposing tree diversity.

Given the above parameters, the most basic part of Algorithm 1 is the training process of each CART. Let C be a cell and $N_n(C)$ be the number of data points falling in C. A split in C is a pair (j, z), where j is a dimension in $\{1, \ldots, d\}$ and z is the position of the split along the j-th dimension in cell C. Let $\mathcal{S}_C$ be the set of all possible splits in C. The CART-split criterion [19] takes the form

$$\begin{aligned}\hat{L}_n(j, z) = &\frac{1}{N_n(C)} \sum_{i:\boldsymbol{x}_i \in C} (y_i - \mu_n(C))^2 \\ &- \frac{1}{N_n(C_L)} \sum_{i:\boldsymbol{x}_i \in C_L} (y_i - \mu_n(C_L))^2 - \frac{1}{N_n(C_R)} \sum_{i:\boldsymbol{x}_i \in C_R} (y_i - \mu_n(C_R))^2\ ,\end{aligned} \tag{7}$$

where $C_L = \{\boldsymbol{x} \in C \colon \boldsymbol{x}^{(j)} < z\}$, $C_R = \{\boldsymbol{x} \in C \colon \boldsymbol{x}^{(j)} \geq z\}$, and $\mu_n(C) = \frac{1}{N_n(C)} \sum_{i:\boldsymbol{x} \in A} y_i$ denotes the average response in any cell C (resp. $\mu_n(C_L)$, $\mu_n(C_R)$), with the convention $0/0 = 0$.

Algorithm 1: A simplified variant of Zhou's original deep forest

```
Require: A training set $S_n = \{(\boldsymbol{x}_1, y_1), \ldots, (\boldsymbol{x}_n, y_n)\}$, the number of trees $2M$,
    $m_{\text{try}} \in \{1, \ldots, d\}$, $a_n \in \{1, \ldots, n\}$, $t_n \in \{1, \ldots, a_n\}$ and the query point $\boldsymbol{x} \in [0,1]^d$.
Ensure: Two-layer deep forest $\bar{h}_{2M,n}(\cdot) = h^{(2)}_{M,n} \circ h^{(1)}_{M,n}(\cdot)$ and the prediction of $\boldsymbol{x}$.
for layer $\ell$ in $\{1, 2\}$ do
  if $\ell = 2$ then
    $S_n \leftarrow \{([\boldsymbol{x}_1, h^{(1)}_{M,n}(\boldsymbol{x}_1)], y_1), \ldots, ([\boldsymbol{x}_n, h^{(1)}_{M,n}(\boldsymbol{x}_n)], y_n)\}$.    ▷ Prediction concatenation.
  end if
  for tree $j \in \{1, 2, \ldots, M\}$ do
    Select $a_n$ data points, without replacement, uniformly in $S_n$.
    Set $\Pi_0 = \{[0,1]^d\}$ the partition associated with the root of the tree.
    For all $1 \leq i \leq a_n$, set $\Pi_i = \emptyset$.
    Set $n_{\text{nodes}} = 1$ and level $= 0$.
    while $n_{\text{nodes}} < t_n$ do
      if $\Pi_{\text{level}} = \emptyset$ then
        level $=$ level $+ 1$.
      else
        Let $C$ be the first element in $\Pi_{\text{level}}$.
        if $C$ contains exactly one point then
          $\Pi_{\text{level}} \leftarrow \Pi_{\text{level}} \setminus \{C\}$.
          $\Pi_{\text{level}+1} \leftarrow \Pi_{\text{level}+1} \cup \{C\}$.
        else
          Select the best split $(j_n^*, z_n^*)$ in $C$ by optimizing the CART-split criterion along the
          dimension $D$ in $\{1, \ldots, d\}$.    ▷ See details in Eq. (7).
          Split cell $C$ along $D$ according to the best split $(j_n^*, z_n^*)$. Call $C_L$ and $C_R$.
          $\Pi_{\text{level}} \leftarrow \Pi_{\text{level}} \setminus \{C\}$.
          $\Pi_{\text{level}+1} \leftarrow \Pi_{\text{level}+1} \cup \{C_L\} \cup \{C_R\}$.
          $n_{\text{nodes}} \leftarrow n_{\text{nodes}} + 1$.
        end if
      end if
    end while
    Compute the predicted value $h^{(\ell)}_n(\boldsymbol{x}, \Theta_j, S_n)$ at the query point $\boldsymbol{x}$ equaling the average of
    the $Y_i$'s falling in the cell of $\boldsymbol{x}$ in partition $\Pi_{\text{level}} \cup \Pi_{\text{level}+1}$.
  end for
  Compute the random forest estimation $h^{(\ell)}_{M,n}(\boldsymbol{x}, \Theta, S_n)$ at the query point $\boldsymbol{x}$ according to
  Eq. (4).
end for
Compute the two-layer deep forest estimation $\bar{h}_{2M,n}(\boldsymbol{x}, \Theta, S_n)$ at the query point $\boldsymbol{x}$ according
to Eq. (6).
```

At each cell C, the best split (j_n^*, z_n^*) is selected by maximizing $\hat{L}_n(j, z)$ over $\mathcal{M}_{\text{try}}$ and $\mathcal{S}_C$, that is,

$$(j_n^*, z_n^*) \in \mathop{\arg\max}_{\substack{(j^*, z^*) \in \mathcal{S}_C \\ j \in \mathcal{M}_{try}}} \hat{L}_n(j, z) \, . \tag{8}$$

5 Properties of prediction concatenation

In this section we show that the properties of the simplified deep forest enable us to analyze the influence of the concatenated new feature in deep forest and the local variation related to the empirical CART-split criterion.

We consider an additive regression model satisfying the following assumption:

Assumption 1. *The response y follows*

$$y = \sum_{j=1}^{d} f_j(x^{(j)}) + \epsilon \, , \tag{9}$$

where $\boldsymbol{x} = \left(x^{(1)}, \ldots, x^{(d)}\right)$ is uniformly distributed over $[0,1]^d$, ϵ is an independent centered Gaussian noise with finite variance $\sigma^2 > 0$ and each component f_j is continuous.

Stone [44], Hastie and Tibshirani [45] popularize these models, which decompose the regression function as a sum of univariate functions. Especially, Scornet et al. [29] prove the consistency of Breiman's original random forest under this assumption. On this basis, we study the impact of the new feature generated in the deep forest.

To start with, we analyze the priority of the new features generated by the previous layer in the selection of splitting features under the CART-split criterion, in both infinite sample regime and finite sample regime respectively.

Proposition 1 (Priority of the new feature)**.** *Assume the data set follows Assumption 1 and the first-layer forest is consistent. The following results hold for any CART in the second-layer forest.*

1. *In the infinite sample regime ($n = \infty$), we consider a single second-layer CART $h_\infty^{(2)}(\cdot, \Theta)$. All splits in this CART are performed along the new feature only.*

2. *In the finite sample regime ($n < \infty$), we consider a single second-layer CART $h_n^{(2)}(\cdot, \Theta)$. Fix $k \in \mathbb{N}^*$ and $\xi, \rho > 0$. Then, with probability $\geq 1 - \rho$, for all n large enough, we have, the error of the first-layer forest is bound by ξ. As a consequence, the first k splits ($j_{q,n}(\boldsymbol{x}), 1 \leq q \leq k$) in this CART are performed along the new feature only.*

Proof sketch. **(P1.1).** In the infinite sample regime, the random forest estimation of the first layer has zero error. Therefore, the new feature of the second layer is the target function $h_{M,\infty}^{(1)}(\boldsymbol{x}) = f(\boldsymbol{x})$. Obviously, the new feature is the most informative dimension. Therefore, when splitting in any cell, CART algorithm will select the new feature as the splitting dimension. **(P1.2).** In the finite sample regime, the random forest estimation of the first layer is not precise. Therefore, there is an error related to n between the new feature and the target function. Firstly, we prove that the distance between the theoretical ($n = \infty$) and empirical ($n < \infty$) first k splits of the CART algorithm is bounded by $c\xi$ with probability $\geq 1 - \rho$, when n is large enough. Connecting with the result of theoretical split above, the proof is completed. □

Remark 1. Proposition 1 shows that the trees in the second layer primarily choose the new feature to split and the degree of this priority depends on the error of the first-layer forest estimator. This also reveals that the advantages of deep forest depend on the performance of the first layer. If the forest of the first layer does not return an estimation with noise reduction, the performance of deep forest cannot be further improved through PreConc operation. This is consistent with the empirical results in previous work [2, 17].

In order to control the risk of deep forest, we need study the local variation property of the empirical CART-split criterion. For any cell C, the variation of regression function $f(\boldsymbol{x})$ within C is defined as

$$\Delta(f, C) = \sup_{\boldsymbol{x}, \boldsymbol{x}' \in C} |f(\boldsymbol{x}) - f(\boldsymbol{x}')| \,. \tag{10}$$

Proposition 2 (Variation of f in the empirical cell)**.** *Assume that Assumption 1 holds and the first-layer forest is consistent. The following results hold for any CART in the second-layer forest $h_n^{(2)}(\cdot, \Theta)$. After splitting along the new feature, the CART will estimate the residual of the first-layer forest estimation. Then for all $\rho, \xi > 0$, there exists $N \in \mathbb{N}^*$ such that, for all $n > N$,*

$$\Pr\left[\Delta\left(f, C_{\Pi,n}(\boldsymbol{x}, \Theta)\right) \leq \xi\right] \geq 1 - \rho \,. \tag{11}$$

Proof sketch. Firstly, we prove the variation of $f(\boldsymbol{x})$ within the cell obtained by the theoretical CART-split criterion converges to zero. Secondly, we prove that the distance between the theoretical and empirical first k splits of the CART convergence to zero. Finally, we prove that the variation of $f(\boldsymbol{x})$ within the empirical cell is close to the theoretical cell. □

Proposition 2 shows that the variation of the regression function $f(\boldsymbol{x})$ within a cell of a random tree $C_{\Pi,n}$ is small provided n is large enough, thereby forcing the approximation error of DF to asymptotically approach zero.

6 Depth is more powerful than width

Our first result considers the $t_n < a_n$ regime, where the number of samples in each leaf node tends to ∞ as $t_n \to \infty$ and $a_n \to \infty$. We prove that controlling the depth of the trees through the number of leaves t_n is sufficient to achieve consistency of deep forest.

Theorem 3 (Universal consistency)**.** *Let $M \geq 1$. Consider two-layer deep forest $\bar{h}_{2M,n}$ given by Eq. (6) and Breiman's random forest $h_{M,n}$ given by Eq. (4) for the random CARTs satisfying $a_n \to \infty, t_n \to \infty$ and $t_n(\log a_n)^9/a_n \to 0$. Then under the setting described in Section 3 and assume the data set follows Assumption 1,*

1. *[29, Theorem 1] the Breiman's random forest $h_{2M,n}$ is consistent for any $M \geq 1$,*
2. *the two-layer deep forest $\bar{h}_{M,n}$ is consistent for any $M \geq 1$.*

Proof sketch. **(T3.1).** The universal consistency of Breiman's random forest is proved by Scornet et al. [29]. **(T3.2).** Firstly, we already know that the variation of the target regression function $f(\boldsymbol{x})$ within a cell of a randomized CART in the second-layer forest is small when n is large enough via Proposition 2. Similar as Scornet et al. [29], we utilize Proposition 2 to control the approximation error of the two-layer deep forest. And the parameter t_n allows us to control the size of the leaves of CART, which allows us to have enough samples in each leaf node to smooth the impact of noise, so as to control the estimation error. Connecting the approximation and estimation error, the consistency of a CART of second-layer deep forest is proved. The universal consistency can be proved via [25, Proposition 1], which guarantees that the error of forest estimator is no more than twice that of individual randomized CART. □

Remark 2. Notice that under this setting, random forest and even deep forest have no obvious advantages over single CART in theory. When we use forests in practice, we do not choose to control the depth of the trees. Empirical studies in Section 7.2 show that the forest with $t_n = a_n$ always outperforms the forest with $t_n < a_n$. Actually, the fully grown trees $t_n = a_n$ is the setting close to practical forest algorithm.

In order to deal with the $t_n = a_n$ regime, we need to introduce an assumption first proposed by Scornet et al. [29]. We denote by $Z_i = \mathbb{1}(\boldsymbol{x}_i \in C_{\Pi,n}(\boldsymbol{x}))$ the indicator that $\boldsymbol{x}_i$ falls into the same cell as $\boldsymbol{x}$ in the random tree designed with $\mathcal{D}_n$ and the random parameter Θ. $Z'_j = \mathbb{1}(\boldsymbol{x}_i \in C_{\Pi'}(\boldsymbol{x}))$ is another independent indicator. Then, we define the correlation between these two indicators conditionally on y_i, y_j or not, respectively

$$\phi_{i,j}(y_i, y_j) = \mathbb{E}[Z_i, Z'_j | \boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n, y_i, y_j] \quad \text{and} \quad \phi_{i,j} = \mathbb{E}[Z_i, Z'_j | \boldsymbol{x}, \boldsymbol{x}_1, \ldots, \boldsymbol{x}_n] \, . \tag{12}$$

Assumption 2. *Let $Z_{i,j} = (Z_i, Z_j)$. Then one of the following two conditions holds:*

1. *One has*
$$\lim_{n\to\infty} (\log a_n)^{2d-2} (\log n)^2 \mathbb{E}\left[\max_{i \neq j} |\phi_{i,j}(y_i, y_j) - \phi_{i,j}|\right]^2 = 0 \, . \tag{13}$$

2. *There exists a constant $C > 0$ and a sequence $(\gamma_n)_n \to 0$ such that, almost surely,*
$$\max_{\ell_1, \ell_2 = 0,1} \frac{|\mathrm{Cor}\,[(y_i - f(\boldsymbol{x}_i)), \mathbb{1}(Z_{i,j} = (\ell_1, \ell_2) | \boldsymbol{x}_i, \boldsymbol{x}_j, y_j)]|}{\mathrm{Pr}^{1/2}[Z_{i,j} = (\ell_1, \ell_2) | \boldsymbol{x}_i, \boldsymbol{x}_j, y_j]} \leq \gamma_n \, , \tag{14}$$
and
$$\max_{\ell_1 = 0,1} \frac{\left|\mathrm{Cor}\left[(y_i - f(\boldsymbol{x}_i))^2, \mathbb{1}(Z_i = \ell_1 | \boldsymbol{x}_i)\right]\right|}{\mathrm{Pr}^{1/2}[Z_i = \ell_1 | \boldsymbol{x}_i]} \leq C \, . \tag{15}$$

(A2.1.) means that the influence of two labels y_i, y_j on the probability of connection of two couples of random points converge to zero as $n \to \infty$. **(A2.2.)** means that the correlation between the noise and the probability of connection of two couples of random points vanishes quickly enough, as $n \to \infty$. However, this assumption is too strong for the Breiman's original random forest [1]. Scornet et al. [29] emphasize that they cannot know whether or not Assumption 2 is satisfied in random forest. In this paper, we recall this assumption and state that this assumption is mild for the second-layer forest estimation in deep forest algorithm.

Since deep forest concatenates the prediction with the raw features as the input for the next layer, the label information is encoded into the new feature of second layer. In this way, the influence of two labels on the connection probability of this pair of samples tends to zero, so **(A2.1.)** is mild for the second-layer forest estimation. From another point of view, Proposition 1, the priority of new feature also shows that the partition can be independent of labels, and the information of new feature is enough to obtain appropriate partition results, so **(A2.2.)** is mild too.

Theorem 4 (Depth is more powerful than width)**.** *Let $M \geq 1$. Consider two-layer deep forest $\bar{h}_{2M,n}$ given by Eq.* (6) *and Breiman's random forest $h_{M,n}$ given by Eq.* (4) *for the random CARTs satisfying $a_n \to \infty, t_n \to \infty$, $t_n = a_n$ and $a_n \log n/n \to 0$. Then under the setting described in Section 3 and assume the data set follows Assumption 1 and 2, the following results hold*

1. *[29, Theorem 2] [30, Theorem 3] The Breiman's random forest $h_{\infty,n}$ is consistent, and for all $M, n \in \mathbb{N}$,*
$$0 \leq R(h_{2M,n}) - R(h_{\infty,n}) \leq \frac{8\|f\|_\infty^2 + 8\sigma^2(1 + 4\log n)}{M} \ . \tag{16}$$

2. *The two-layer deep forest $\bar{h}_{\infty,n}$ is consistent, and for all $M, n \in \mathbb{N}$, if $\Delta(f, C_{\Pi,n}(\boldsymbol{x}, \Theta))$ is small enough, then*
$$0 \leq R(\bar{h}_{2M,n}) - R(\bar{h}_{\infty,n}) \leq \frac{64\|f\|_\infty^2 + 64\sigma^2(1 + 4\log n)}{M^2} \ . \tag{17}$$

Proof sketch. **(T4.1).** The consistency of the infinite Breiman's random forest is proved by Scornet et al. [29]. And the convergence rate of the finite random forest with the number of trees M is proved by Scornet [30]. **(T4.2).** Similar to Scornet et al. [29], we recall Proposition 2 to control the approximation error of the two-layer deep forest. Then the estimation error is controlled by forcing the subsampling rate a_n/n to be $o(1/\log n)$. Different from the bagging-style mechanism in random forest, the residual-style mechanism shown in Proposition 2 makes the second-layer forest in DF can reuse the first-layer estimation and focus on the residual learning. □

Remark 3. In the $t_n = a_n$ setting, Scornet et al. [29] show that the sub-sampling rate a_n/n is the key component in random forest. Because the small rate ensures that query point $\boldsymbol{x}$ is connected with enough different data points through different trees, the convergence rate of RF is $\mathcal{O}(1/M)$ *w.r.t.* the number of trees M. Theorem 4 proves that, if the first layer forest can encode the regression function $f(\boldsymbol{x})$ into the new feature with noise reduction, the cascade structure with PreConc in DF can further accelerate the convergence. Because the second layer forest estimates the residual of the first layer, the trees in each layer of forest are more different. As a result, the convergence rate of deep forest will be improved to $\mathcal{O}(1/M^2)$. This result reflects that deep forest with deeper layer will be more powerful than shallower layer.

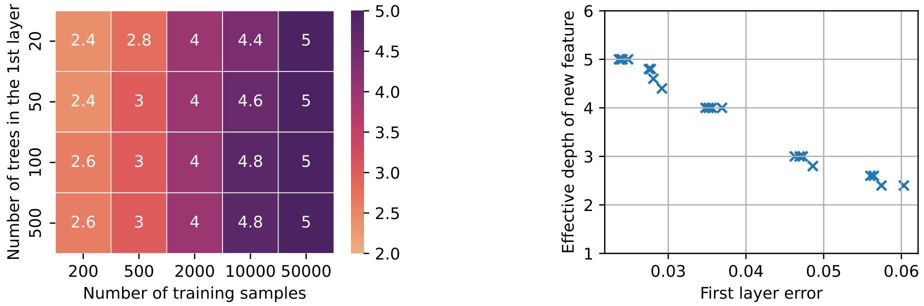
7 Simulation experiments

7.1 Priority of the new feature

This experiment aims to verify the priority of the new feature in choosing which feature to split as suggested in Proposition 1. We focus on a second layer decision tree built upon the first layer random forest. Since a regression forest has only one output dimension, there is only one new feature for the second layer tree. More specifically, we count the maximum consecutive levels from root node that use the new feature to split, which we call *effective depth* for short.

The synthetic data set is generated as $y = f(\boldsymbol{x}) + \epsilon$, where
$$f(\boldsymbol{x}) = \frac{1}{5} \sum_{1 \leq j \leq 5} x_j \ , \tag{18}$$

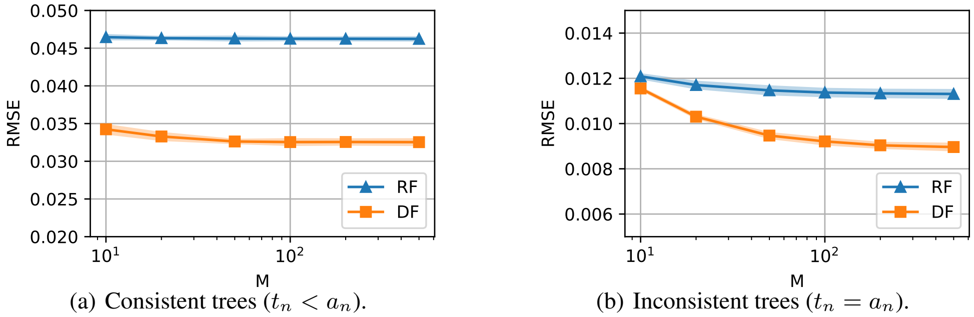
$\boldsymbol{x}$ is uniformly distributed over $[0,1]^5$, $\epsilon \sim \mathcal{N}\left(0, \sigma^2\right)$, where $\sigma = 0.02$. We vary the number of training samples and the number of trees in the first layer forest, and report the average effective depth of 5 runs in Figure 2(a). It is easy to observe that no matter what the first layer's setting is, the effective depth of new feature is at least 2.4. That is to say, at the beginning of the growing of the second layer tree, CART will always choose the new feature to split. And we can see that with



(a) Heatmap of the effective depth of new feature under different settings of the first layer forest.

(b) Effective depth of new feature against the predictive error of the first layer.

Figure 2: Illustrations of the effective depth of the new feature (the consecutive levels from root node that split on the new feature only). The larger the effective depth, the higher priority the new feature takes in being chosen as the split feature under the CART-split criterion.



(a) Consistent trees ($t_n < a_n$).

(b) Inconsistent trees ($t_n = a_n$).

Figure 3: Root mean square error with the increasing of trees.

the increase of trees in the first layer and training samples, the CART's preference of new feature extend to deeper layers. Figure 2(b) further plots the effective depth against the predictive error of the first layer measured by root mean square error (RMSE), showing that the better the prediction performance of the first layer, the more favored the new feature is for splitting.

7.2 Convergence rate *w.r.t.* the number of trees

To study the convergence rate of predictive error with respect to the number of trees in the whole model, we fix the number of training samples to be 50,000. We compare a 2-layer deep forest (DF) with M trees each layer to a random forest (RF) with $2M$ trees, and we adopt 3-fold split in training both DF and RF to ensure fairness. The performance is measured by RMSE with respect to the noise-free version of data generating function. Since the training data are with noise, a fully grown tree is inconsistent. In Figure 3(a), we set the minimum leaf size of the trees to be $\sqrt{n}$, i.e., 233 in the case when training data size is 50,000. In Figure 3(b) the trees are fully grown with only one sample in each leaf. We plot the average RMSE of 5 runs against the increasing number of trees, with the colored band indicating the standard deviation.

Theorem 3 reveals that when the component trees are consistent, random forest and deep forest are both consistent. However, the consistency analysis result cannot guarantee the finite sample performance in practice. Comparing Figure 3(a) to Figure 3(b), we can see that even though the training set is as large as 50,000, the performances of RF and DF using consistent trees are still much worse than using inconsistent trees. This observation confirms the efficacy of the default experimental setting that uses fully grown trees in RF and DF. Figure 3(b) shows that DF enjoys a faster improvement in RMSE with the increasing of M. More specifically, DF with $M = 20$ outperforms RF with $M = 500$. This experimental result matches our theoretical result in Theorem 4 that DF has a faster convergence rate *w.r.t.* the number of trees M.

8 Conclusion

In this paper, we prove that a two-layer deep forest has a faster convergence rate *w.r.t.* the number of component trees M than random forest. This work provides the first theoretical analysis of the prediction concatenation (PreConc) operation which is crucial for feature transformation in deep forests, although based on a very simplified structure where the concatenation of multiple random forests' predictions and completely-random forests' predictions in each layer of deep forest has not been taken into account.

On the one hand, this paper focuses on the asymptotic consistency of deep forests, so the result is strictly true only when the number of samples tends to infinity. As for the generalization analysis of deep forests with finite samples, we leave it to future work. On the other hand, the two assumptions used in this paper have certain limitations. Experiments on simulation and real-world data sets show that our theoretical results are valid in many objective function classes other than Assumption 1. How to further relax the conditions in Assumption 1 will be an interesting problem. As for Assumption 2, it is still not strictly verified. However, quantifying the trade-off between label information and partition randomness will be a very important topic in future work.

Acknowledgements

This research was supported by the NSFC (61921006) and the Collaborative Innovation Center of Novel Software Technology and Industrialization. The authors would like to thank the anonymous reviewers for constructive suggestions, as well as Jin-Hui Wu, Qin-Cheng Zheng and Peng Tan for helpful discussions.

References

[1] Leo Breiman. Random forests. *Machine learning*, 45(1):5–32, 2001.

[2] Zhi-Hua Zhou and Ji Feng. Deep forest. *National Science Review*, 6(1):74–86, 2019.

[3] Jerome H. Friedman. Greedy function approximation: A gradient boosting machine. *The Annals of Statistics*, 29(5):1189–1232, 2001.

[4] Pierre Geurts, Damien Ernst, and Louis Wehenkel. Extremely randomized trees. *Machine Learning*, 63(1):3–42, 2006.

[5] Tianqi Chen and Carlos Guestrin. XGboost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 785–794, 2016.

[6] Ran Su, Xinyi Liu, Leyi Wei, and Quan Zou. Deep-resp-forest: A deep forest model to predict anti-cancer drug response. *Methods*, 166:91–102, 2019.

[7] Meng Zhou, Xianhua Zeng, and Aozhu Chen. Deep forest hashing for image retrieval. *Pattern Recognition*, 95:114–127, 2019.

[8] Yaakoub Boualleg, Mohamed Farah, and Imed Riadh Farah. Remote sensing scene classification using convolutional features and deep forest classifier. *IEEE Geoscience and Remote Sensing Letters*, 16(12):1944–1948, 2019.

[9] Ya-Lin Zhang, Jun Zhou, Wenhao Zheng, Ji Feng, Longfei Li, Ziqi Liu, Ming Li, Zhiqiang Zhang, Chaochao Chen, Xiaolong Li, Yuan (Alan) Qi, and Zhi-Hua Zhou. Distributed deep forest and its application to automatic detection of cash-out fraud. *ACM Transactions on Intelligent Systems and Technology*, 10(5):55:1–55:19, 2019.

[10] Lev V. Utkin and Mikhail A. Ryabinin. A siamese deep forest. *Knowledge-Based Systems*, 139: 13–22, 2018.

[11] Lev V. Utkin and Mikhail A. Ryabinin. Discriminative metric learning with deep forest. *International Journal on Artificial Intelligence Tools*, 28(2):1950007:1–1950007:19, 2019.

[12] Liang Yang, Xi-Zhu Wu, Yuan Jiang, and Zhi-Hua Zhou. Multi-label learning with deep forest. In *Proceedings of the 24th European Conference on Artificial Intelligence*, volume 325, pages 1634–1641, 2020.

[13] Qian-Wei Wang, Liang Yang, and Yu-Feng Li. Learning from weak-label data: A deep forest expedition. In *Proceedings of the 34th AAAI Conference on Artificial Intelligence*, volume 34, pages 6251–6258, 2020.

[14] Ming Pang, Kai-Ming Ting, Peng Zhao, and Zhi-Hua Zhou. Improving deep forest by confidence screening. *IEEE Transactions on Knowledge and Data Engineering*, 34(9):4298–4312, 2022.

[15] Yi-He Chen, Shen-Huan Lyu, and Yuan Jiang. Improving deep forest by exploiting high-order interactions. In *IEEE International Conference on Data Mining*, pages 1030–1035, 2021.

[16] Shen-Huan Lyu, Liang Yang, and Zhi-Hua Zhou. A refined margin distribution analysis for forest representation learning. In *Advances in Neural Information Processing Systems 32*, pages 5531–5541, 2019.

[17] Ludovic Arnould, Claire Boyer, and Erwan Scornet. Analyzing the tree-layer structure of deep forests. In *Proccedings of the 37th International Conference on Machine Learning*, pages 342–350, 2021.

[18] Shen-Huan Lyu, Yi-He Chen, and Zhi-Hua Zhou. A region-based analysis for the feature concatenation in deep forests. *Chinese Journal of Electronics*, 2022.

[19] Leo Breiman, Jerome Friedman, Charles J Stone, and Richard A Olshen. *Classification and Regression Trees*. CRC press, 1984.

[20] Gérard Biau. Analysis of a random forests model. *The Journal of Machine Learning Research*, 13(1):1063–1095, 2012.

[21] Misha Denil, David Matheson, and Nando De Freitas. Narrowing the gap: Random forests in theory and in practice. In *Proccedings of the 30th International Conference on Machine Learning*, pages 665–673, 2014.

[22] Yi Lin and Yongho Jeon. Random forests and adaptive nearest neighbors. *Journal of the American Statistical Association*, 101(474):578–590, 2006.

[23] Gérard Biau and Luc Devroye. On the layered nearest neighbour estimate, the bagged nearest neighbour estimate and the random forest method in regression and classification. *Journal of Multivariate Analysis*, 101(10):2499–2518, 2010.

[24] Nicolai Meinshausen. Quantile regression forests. *Journal of Machine Learning Research*, 7: 983–999, 2006.

[25] Gérard Biau, Luc Devroye, and Gäbor Lugosi. Consistency of random forests and other averaging classifiers. *Journal of Machine Learning Research*, 9(9), 2008.

[26] Hemant Ishwaran and Udaya B Kogalur. Consistency of random survival forests. *Statistics & Probability Letters*, 80(13-14):1056–1064, 2010.

[27] Robin Genuer. Variance reduction in purely random forests. *Journal of Nonparametric Statistics*, 24(3):543–562, 2012.

[28] Ruoqing Zhu, Donglin Zeng, and Michael R Kosorok. Reinforcement learning trees. *Journal of the American Statistical Association*, 110(512):1770–1784, 2015.

[29] Erwan Scornet, Gérard Biau, and Jean-Philippe Vert. Consistency of random forests. *The Annals of Statistics*, 43(4):1716–1741, 2015.

[30] Erwan Scornet. On the asymptotics of random forests. *Journal of Multivariate Analysis*, 146: 72–83, 2016.

[31] Wei Gao and Zhi-Hua Zhou. Towards convergence rate analysis of random forests for classification. In *Advances in Neural Information Processing Systems 33*, pages 9300–9311, 2020.

[32] Wei Gao, Fan Xu, and Zhi-Hua Zhou. Towards convergence rate analysis of random forests for classification. *Artificial Intelligence*, 313:103788, 2022. doi: https://doi.org/10.1016/j.artint.2022.103788.

[33] Carolin Strobl, Anne-Laure Boulesteix, Achim Zeileis, and Torsten Hothorn. Bias in random forest variable importance measures: Illustrations, sources and a solution. *BMC Bioinformation*, 8, 2007.

[34] Gilles Louppe, Louis Wehenkel, Antonio Sutera, and Pierre Geurts. Understanding variable importances in forests of randomized trees. In *Advances in Neural Information Processing Systems 26*, pages 431–439, 2013.

[35] Ludovic Arnould, Claire Boyer, and Erwan Scornet. Is interpolation benign for random forests? *arXiv preprint arXiv:2202.03688*, 2022.

[36] Xiao Li, Yu Wang, Sumanta Basu, Karl Kumbier, and Bin Yu. A debiased MDI feature importance measure for random forests. In *Advances in Neural Information Processing Systems 32*, pages 8047–8057, 2019.

[37] Antonio Sutera, Gilles Louppe, Vân Anh Huynh-Thu, Louis Wehenkel, and Pierre Geurts. From global to local MDI variable importances for random forests and when they are shapley values. In *Advances in Neural Information Processing Systems 34*, pages 3533–3543, 2021.

[38] Ronen Eldan and Ohad Shamir. The power of depth for feedforward neural networks. In *Proceedings of the 29th Conference on Learning Theory*, pages 907–940, 2016.

[39] Matus Telgarsky. Benefits of depth in neural networks. In *Proceedings of the 29th Conference on Learning Theory*, pages 1517–1539, 2016.

[40] Itay Safran and Ohad Shamir. Depth-width tradeoffs in approximating natural functions with neural networks. In *Proceedings of the 34th International Conference on Machine Learning*, pages 2979–2987, 2017.

[41] Eran Malach and Shai Shalev-Shwartz. Is deeper better only when shallow is good? In *Advances in Neural Information Processing Systems 32*, pages 6426–6435, 2019.

[42] Amit Daniely and Eran Malach. Learning parities with neural networks. In *Advances in Neural Information Processing Systems 33*, pages 20356–20365, 2020.

[43] Zhi-Hua Zhou and Ji Feng. Deep forest: Towards an alternative to deep neural networks. In *Proceedings of the 26th International Joint Conference on Artificial Intelligence*, pages 3553–3559, 2017.

[44] Charles J Stone. Additive regression and other nonparametric models. *The Annals of Statistics*, 13(2):689–705, 1985.

[45] Trevor Hastie and Robert Tibshirani. Generalized additive models: some applications. *Journal of the American Statistical Association*, 82(398):371–386, 1987.

[46] László Györfi, Michael Kohler, Adam Krzyzak, Harro Walk, et al. *A Distribution-Free Theory of Nonparametric Regression*, volume 1. Springer, 2002.

[47] Pengfei Ma, Youxi Wu, Yan Li, Lei Guo, He Jiang, Xingquan Zhu, and Xindong Wu. HW-Forest: Deep forest with hashing screening and window screening. *ACM Transactions on Knowledge Discovery from Data*, 2022. doi: 10.1145/3532193.

[48] Misha Denil, David Matheson, and Nando Freitas. Consistency of online random forests. In *Proccedings of the 31st International Conference on Machine Learning*, pages 1256–1264, 2013.

[49] Shai Shalev-Shwartz and Shai Ben-David. *Understanding Machine Learning: From Theory to Algorithms*. Cambridge university press, 2014.

[50] Luc Devroye, László Györfi, and Gábor Lugosi. *A Probabilistic Theory of Pattern Recognition*. Springer, 1996.

[51] Luc Devroye. A note on the height of binary search trees. *Journal of the ACM*, 33(3):489–498, 1986.

[52] Sumanta Basu, Karl Kumbier, James B. Brown, and Bin Yu. Iterative random forests to discover predictive and stable high-order interactions. *Proceedings of the National Academy of Sciences*, 115(8):1943–1948, 2018.

[53] Zhi-Hua Zhou. Why over-parameterization of deep neural networks does not overfit? *Science China Information Sciences*, 64(1):1–3, 2021.

[54] Gérard Biau and Erwan Scornet. A random forest guided tour. *Test*, 25(2):197–227, 2016.

Checklist

The checklist follows the references. Please read the checklist guidelines carefully for information on how to answer these questions. For each question, change the default **[TODO]** to [Yes] , [No] , or [N/A] . You are strongly encouraged to include a **justification to your answer**, either by referencing the appropriate section of your paper or providing a brief inline description. For example:

- Did you include the license to the code and datasets? [Yes] See Section 7.
- Did you include the license to the code and datasets? [No] The code and the data are proprietary.
- Did you include the license to the code and datasets? [N/A]

Please do not modify the questions and only use the provided macros for your answers. Note that the Checklist section does not count towards the page limit. In your paper, please delete this instructions block and only keep the Checklist section heading above along with the questions/answers below.

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes]
   (b) Did you describe the limitations of your work? [Yes]
   (c) Did you discuss any potential negative societal impacts of your work? [No] This is a theoretical work.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes]
   (b) Did you include complete proofs of all theoretical results? [Yes]
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes]
   (b) Did you specify all the training details (e.g., data splits, hyper-parameters, how they were chosen)? [Yes]
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes]
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [Yes]
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [Yes]
   (b) Did you mention the license of the assets? [Yes]
   (c) Did you include any new assets either in the supplemental material or as a URL? [No]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [No]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [No]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [No]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [No]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [No]

Supplementary Materials for Depth is More Powerful than Width in Deep Forest

A Table of notations

Category	Sign	Description
Setting	$\mathcal{D}$	The unknown distribution
	$f(\boldsymbol{x})$	The target regression function
	ϵ	The Gaussian noise
	S_n	Data set with n samples
	$(\boldsymbol{x}, y)$	A sample drawn from $\mathcal{D}$
	$R(h)$	$\mathbb{L}_2$ risk of the estimator h
Tree and forest	Θ	A random variable that accounts for the randomization procedure and independent of S_n
	$h_n(\boldsymbol{x}, \Theta, S_n)$	A randomized decision tree
	Π	The partition built by performing successive axis-aligned splits according to Θ
	$C_{\Pi,n}(\boldsymbol{x})$	The cell of the tree partition containing $\boldsymbol{x}$
	$N_n(C_{\Pi,n}(\boldsymbol{x}))$	The number of training samples falling into $C_{\Pi,n}(\boldsymbol{x})$
	$h_{M,n}(\boldsymbol{x}, \Theta, S_n)$	A random forest
	$W_{ni}(\boldsymbol{x})$	The probability that $\boldsymbol{x}$ and $\boldsymbol{x}_i$ fall into a same cell.
	$\mathbb{1}_{\boldsymbol{x} \overset{\Theta}{\leftrightarrow} \boldsymbol{x}_i}$	The indicator where $\boldsymbol{x}$ and $\boldsymbol{x}_i$ are connected.
Deep forest	$[\boldsymbol{x}, h]$	PreConc: The concatenation of the raw features $\boldsymbol{x}$ and the new feature (prediction) h
	$\bar{h}_{2M,n}(\boldsymbol{x}, \Theta, S_n)$	A two-layer deep forest with $2M$ trees
	$h^{(\ell)}_{M,n}(\boldsymbol{x}, \Theta, S_n)$	The ℓ-th layer forest of two-layer deep forest with M trees, $\ell \in \{1, 2\}$
CART-split	$t = (j, z)$	A split where j is a dimension in $\{1, \ldots, d\}$ and z is the position of the split along the j-th dimension
	C, C_L and C_R	Any cell C, its left node and right node
	$\mu_n(\cdot)$	The average response in any cell
	$\hat{L}_n(j, z)$	The empirical version of CART-split criterion function
	(j_n^*, z_n^*)	The best split selected by maximizing $\hat{L}_n(j, z)$
	$L^*(j, z)$	The theoretical version of CART-split criterion function
	(j^*, z^*)	The best split selected by maximizing $L_n^*(j, z)$
	$\boldsymbol{t}_k = (t_1, \ldots, t_k)$	A sequence of first k splits
	$\mathcal{T}_k(\boldsymbol{x})$	The collection of all possible $k \geq 1$ consecutive splits used to build the cell containing $\boldsymbol{x}$
	$d_\infty(\boldsymbol{t}_k, \mathcal{T})$	The infinite distance d_∞ between $\boldsymbol{t}_k \in \mathcal{T}_k(\boldsymbol{x})$ and any $\mathcal{T} \subset \mathcal{T}_k(\boldsymbol{x})$
Statistics	$\mathrm{Cor}(x_1, x_2)$	The correlation between x_1 and x_2
	$\mathrm{Var}(x)$	The variance of x
	$\mathrm{Unif}([0, 1]^d)$	The uniform distribution on $[0, 1]^d$

Table S1: Notations of this work.

B Proofs

In this section, we provide the detailed proofs for the main theorems and corollaries. First, we present a series of useful lemmas as follows:

Lemma 5. *For any $a, b, z \in \mathbb{R}$, let $f(x)$ be a continuous and bounded function, and $F(z) = \int_a^z f(t)\,\mathrm{d}t$ be the integral function of $f(x)$ over a to z, and given $\int_a^b f(t)\,\mathrm{d}t = 0$. We denote by $\mu_{a \le x \le z}$ the average of $f(x)$ between a and z, and $\mu_{z \le x \le b}$ the average of $f(x)$ between z and b.*

$$\frac{1}{(z-a)(b-z)} F^2(z) = \frac{(z-a)(b-z)}{(b-a)^2} \left(\mu_{a \le x \le z} - \mu_{z \le x \le b}\right)^2 . \tag{19}$$

Proof of Lemma 5.

$$\begin{aligned} \frac{1}{(z-a)(b-z)} F^2(z) =& \frac{z-a}{b-a} (F(z)/(z-a))^2 + \frac{b-z}{b-a} (F(z)/(b-z))^2 \\ =& \left(\frac{z-a}{b-a} \mu_{a \le x \le z}^2 + \frac{b-z}{b-a} \mu_{z \le x \le b}^2 \right) \left(\frac{z-a}{b-a} + \frac{b-z}{b-a} \right) \\ & - \left(\frac{z-a}{b-a} \mu_{a \le x \le z} + \frac{b-z}{b-a} \mu_{z \le x \le b} \right)^2 \\ =& \frac{(z-a)(b-z)}{(b-a)^2} \left(\mu_{a \le x \le z} - \mu_{z \le x \le b}\right)^2 . \end{aligned} \tag{20}$$

□

Lemma 6. *For any $a, b \in \mathbb{R}$, let $f(x)$ be a continuous and bounded function, and given $\int_a^b f(t)\,\mathrm{d}t = 0$. Let μ_Ω be the average of $f(x)$ in the area C_Ω satisfying the condition Ω. For any z^*, existing a $z^\star$ such that*

$$\frac{(z^*-a)(b-z^*)}{(b-a)^2} \left(\mu_{a \le x \le z^*} - \mu_{z^* \le x \le b}\right)^2 \le \frac{|C_{f(x) \le f(z^\star)}||C_{f(x) \ge f(z^\star)}|}{|C_{f(x)}|^2} \left(\mu_{f(x) \le f(z^\star)} - \mu_{f(x) \ge f(z^\star)}\right)^2 . \tag{21}$$

Proof of Lemma 6. For any z^*, we let $z^\star$ be the value satisfying $\frac{|C_{f(x) \le f(z^\star)}|}{|C_{f(x)}|} = \frac{z^*-a}{b-a}$ and $\frac{|C_{f(x) \ge f(z^\star)}|}{|C_{f(x)}|} = \frac{b-z^*}{b-a}$. Then we just need to prove that

$$\left(\mu_{a \le x \le z^*} - \mu_{z^* \le x \le b}\right)^2 \le \left(\mu_{f(x) \le f(z^\star)} - \mu_{f(x) \ge f(z^\star)}\right)^2 . \tag{22}$$

Essentially, this is to compare the inter-class variance between two child nodes after split in different dimensions (x or $f(x)$). We denote by $Z = \int_{C_L} f(t)\,\mathrm{d}t < 0$ the integral of $f(x)$ in the left node, then the inter-class variance equals to $(Z/p_{\text{left}} - (-Z)/(1-p_{\text{left}}))^2 = Z^2/(p_{\text{left}} \cdot (1-p_{\text{left}}))^2 \propto Z^2$. Because p_{left} is fixed, we just need to consider Z^2. According to the rearrangement inequality, we know that the ordered sum is less than the disordered sum. Therefore, split along the dimension of $f(x)$, i.e., $Z_{f(x) \le f(z^\star)}$ can be view as an ordered sum of $f(x)$, is smaller than the disorder sum $Z_{a \le x \le z^*}$. We have

$$Z_{f(x) \le f(z^\star)} \le Z_{a \le x \le z^*} \le 0 , \tag{23}$$

so $Z^2_{f(x) \le f(z^\star)} \ge Z^2_{a \le x \le z^*}$. Since Eq. (22) is proved, then lemma is proved. □

Lemma 7 (The distance between theoretical and empirical splits)**.** *Assume that Assumption 1 is satisfied and the first-layer forest is consistent. Fix $\xi, \rho > 0$ and $k \in \mathbb{N}^*$. Then in the second layer, there exists $N \in \mathbb{N}^*$ such that, for all $n \ge N$,*

$$\Pr\left[t_\infty \left(\hat{\boldsymbol{t}}_{k,n}(\boldsymbol{x}, \Theta), \mathcal{T}_k^*(\boldsymbol{x}, \Theta) \right) \le \xi \right] \ge 1 - \rho . \tag{24}$$

Proof of Lemma 7. We prove by induction that, for all k, with probability $1-\rho$, for all $\xi > 0$ and all n large enough,

$$d_\infty(\hat{\boldsymbol{t}}_{k,n}(\boldsymbol{x}, \Theta), \mathcal{T}_k^*(\boldsymbol{x}, \Theta)) \le \xi . \tag{25}$$

Call this property H_k. Fix $k > 1$ and assume that H_{k-1} is true. For all $\boldsymbol{t}_{k-1} \in \mathcal{T}_{k-1}(\boldsymbol{x})$, let

$$\hat{t}_{k,n}(\boldsymbol{t}_{k-1}) \in \arg\min_{t_k} \hat{L}_n(\boldsymbol{x}, \boldsymbol{t}_{k-1}, t_k) , \tag{26}$$

and

$$t_k^*(\boldsymbol{t}_{k-1}) \in \arg\min_{t_k} L^*(\boldsymbol{x}, \boldsymbol{t}_{k-1}, t_k) \ , \tag{27}$$

where the minimum is evaluated over $\{t_k \in \mathcal{S}_{C(\boldsymbol{x}, \boldsymbol{t}_{k-1})} \colon t_k^{(1)} \in \mathcal{M}_{\text{try}}\}$. Fix $\rho > 0$. In the rest of the proof, we assume that Θ is fixed.

$$d_\infty \left(\hat{t}_{k,n}(\hat{\boldsymbol{t}}_{k-1,n}), \mathcal{T}_k^*\right) \le d_\infty \left(\hat{t}_{k,n}(\hat{\boldsymbol{t}}_{k-1,n}), t_k^*(\hat{\boldsymbol{t}}_{k-1,n})\right) + d_\infty \left(t_k^*(\hat{\boldsymbol{t}}_{k-1,n}), \mathcal{T}_k^*\right) \ . \tag{28}$$

According to Scornet et al. [29, Lemma 2 and preliminary result in Lemma 3], we have, with probability $\ge 1 - 2\rho$, for all n large enough,

$$d_\infty \left(\hat{t}_{k,n}(\hat{\boldsymbol{t}}_{k-1,n}), t_k^*(\hat{\boldsymbol{t}}_{k-1,n})\right) \le \xi \ . \tag{29}$$

Therefore, we just need to prove that $d_\infty \left(t_k^*(\hat{\boldsymbol{t}}_{k-1,n}), \mathcal{T}_k^*\right) \to 0$ in probability as $n \to \infty$. Let $\{\boldsymbol{t}_{k-1}^{*,i} \colon i \in \mathcal{I}\}$ be the set of best first $k-1$th theoretical splits and $t_k^*(\{\boldsymbol{t}_{k-1}^{*,i})$ be the kth theoretical spits given that the $k-1$ previous ones are $\boldsymbol{t}_{k-1}^{*,i}$.

Let

$$L^{i,*}(\boldsymbol{x}, t_k) = L_k^*(\boldsymbol{x}, \boldsymbol{t}_{k-1}^{*,i}, t_k) \quad \text{and} \quad \hat{L}^*(\boldsymbol{x}, t_k) = L_k^*(\boldsymbol{x}, \hat{\boldsymbol{t}}_{k-1,n}, t_k) \ , \tag{30}$$

$$t_k^*(\boldsymbol{t}_{k-1}^{*,i}) \in \arg\min_{t_k} L^{i,*}(\boldsymbol{x}, t_k) \quad \text{and} \quad t_k^*(\hat{\boldsymbol{t}}_{k-1,n}) \in \arg\min_{t_k} \hat{L}^{i,*}(\boldsymbol{x}, t_k) \ . \tag{31}$$

Then the original problem equals to that:

$$\inf_{i \in \mathcal{I}} d_\infty(t_k^*(\boldsymbol{t}_{k-1}^{*,i}), t_k^*(\hat{\boldsymbol{t}}_{k-1,n})) \to 0, \text{in probability, as } n \to \infty \ . \tag{32}$$

According to Scornet et al. [29, Technical Lemma 2], we just need to prove that, with probability $\ge 1 - \rho$,

$$\inf_i |L^{i,*}(\boldsymbol{x}, t_k^*(\hat{\boldsymbol{t}}_{k-1,n})) - L^{i,*}(\boldsymbol{x}, t_k^*(\boldsymbol{t}_{k-1}^{*,i}))| \le 6\xi \ . \tag{33}$$

$$\begin{aligned}
\inf_i |L^{i,*}(\boldsymbol{x}, t_k^*(\hat{\boldsymbol{t}}_{k-1,n})) - L^{i,*}(\boldsymbol{x}, t_k^*(\boldsymbol{t}_{k-1}^{*,i}))| &\le 2 \inf_i \sup_{t_k} |\hat{L}^*(\boldsymbol{x}, t_k) - L^{i,*}(\boldsymbol{x}, t_k)| \\
\triangleright \text{ According to the continuity of } L_k^* \quad &\le 4\xi + 2 \inf_i \sup_j |\hat{L}^*(\boldsymbol{x}, c'_{j,\boldsymbol{x}}) - L^{i,*}(\boldsymbol{x}, c'_{j,\boldsymbol{x}})| \\
\triangleright \text{ According to Proposition (1.1)} \quad &= 2 \inf_i |L_k^*(\boldsymbol{x}, \hat{\boldsymbol{t}}_{k-1,n}, c'_{d+1,\boldsymbol{x}}) - L_k^*(\boldsymbol{x}, \boldsymbol{t}_{k-1}^{*,i}, c'_{d+1,\boldsymbol{x}})| \\
&\quad + 4\xi \ ,
\end{aligned} \tag{34}$$

where $\mathcal{C}'_{\delta,\boldsymbol{x}} = \{c'_{j,\boldsymbol{x}} \colon 1 \le j \le d+1\}$ is a finite subset such that, for all t_k, $d_\infty(t_k, \mathcal{C}'_{\delta,\boldsymbol{x}}) \le \delta$, by default $d+1$-th dimension is the new feature. When the first-layer forest is consistent, the second-layer CART always theoretically split along the new feature. Since L_k^* is uniformly continuous, by assumption H_{k-1}, $\inf_i \|\hat{\boldsymbol{t}}_{k-1,n} - \boldsymbol{t}_{k-1}^{*,i}\|_\infty \to 0$, we have

$$\inf_i |L^{i,*}(\boldsymbol{x}, t_k^*(\hat{\boldsymbol{t}}_{k-1,n})) - L^{i,*}(\boldsymbol{x}, t_k^*(\boldsymbol{t}_{k-1}^{*,i}))| \le 6\xi \ . \tag{35}$$

The lemma is proved. □

B.1 Proof of Proposition 1 and Proposition 2

Proposition 1 (Priority of the new feature)**.** *Assume the data set follows Assumption 1 and the first-layer forest is consistent. The following results hold for any CART in the second-layer forest.*

1. *In the infinite sample regime ($n = \infty$), we consider a single second-layer CART $h_\infty^{(2)}(\cdot, \Theta)$. All splits in this CART are performed along the new feature only.*

2. *In the finite sample regime ($n < \infty$), we consider a single second-layer CART $h_n^{(2)}(\cdot, \Theta)$. Fix $k \in \mathbb{N}^*$ and $\xi, \rho > 0$. Then, with probability $\ge 1 - \rho$, for all n large enough, we have, the error of the first-layer forest is bound by ξ. As a consequence, the first k splits ($j_{q,n}(\boldsymbol{x}), 1 \le q \le k$) in this CART are performed along the new feature only.*

Proof of Proposition 1. First, we consider the infinite sample regime ($n = \infty$), which the first-layer forest estimation is precise, i.e., $h_{M,\infty}(\boldsymbol{x}) = f(\boldsymbol{x})$.

Recall that, for any cell A, the empirical CART criterion used to split A in the random forest is defined in Eq.(7). For any split (j, z), we denote the theoretical version of $L^*(j, z)$ by

$$\begin{aligned} L^*(j,z) = & \operatorname{Var}[y|\boldsymbol{x} \in A] - \Pr[x^{(j)} < z] \operatorname{Var}[y|x^{(j)} < z \wedge \boldsymbol{x} \in A] \\ & - \Pr[x^{(j)} \geq z] \operatorname{Var}[y|x^{(j)} \geq z \wedge \boldsymbol{x} \in A] \, . \end{aligned} \tag{36}$$

According to the strong law of large numbers, we have $L_n(j, z) \to L^*(j, z)$ almost surely as $n \to \infty$ for all splits $(j, z) \in \mathcal{S}_A$. Thus we have the best theoretical split (j^*, z^*) of the cell A

$$(j^*, z^*) \in \underset{\substack{(j^*,z^*) \in \mathcal{S}_A \\ j \in \mathcal{M}_{try}}}{\arg\max} \; L^*(j,z) \, . \tag{37}$$

In the random forest and deep forest, $\mathcal{M}_{\text{try}}$ is also an important parameter. Unlike random forests, where we give all features the same probability to be selected, deep forests often choose new features with higher probability or even make new features mandatory in order to make use of the representation information brought by new features. Therefore, the deep forest or CART analyzed in this paper selects new features by default. Otherwise, the tree and forest in the second layer will be equivalent to the random forest, because it does not inherit any information in the first layer.

1. Infinite sample region ($n = \infty$)

We start by proving the result in dimension $d = 1$. Letting $C_x = [a, b]$ be any cell, and recalling that $y = f(x^{(1)}) + \epsilon$, then in the infinite sample regime we define the theoretical version of CART's split criterion function on the raw feature as

$$\begin{aligned} L^*(1,z) = & \operatorname{Var}[y|x^{(1)} \in C_x] - \Pr[a \leq x^{(1)} \leq z | x^{(1)} \in C_x] \operatorname{Var}[y | a \leq x^{(1)} \leq z] \\ & - \Pr[z \leq x^{(1)} \leq b | x^{(1)} \in C_x] \operatorname{Var}[y | z \leq x^{(1)} \leq b] \\ = & -\frac{1}{(b-a)^2} \left(\int_a^b f(t) \, \mathrm{d}t \right)^2 + \frac{1}{(b-a)(z-a)} \left(\int_a^z f(t) \, \mathrm{d}t \right)^2 \\ & + \frac{1}{(b-a)(b-z)} \left(\int_z^b f(t) \, \mathrm{d}t \right)^2 . \end{aligned}$$

Let $I = \int_a^b f(t) \, \mathrm{d}t$ and $F(z) = \int_a^z f(t) \, \mathrm{d}t$. Then, the theoretical criterion function

$$L^*(1,z) = \frac{1}{(z-a)(b-z)} \left(F(z) - I \cdot \frac{z-a}{b-a} \right)^2 . \tag{38}$$

According to the consistency of original random forest [29], we have the new feature $h_\infty^{(1)}(x) = f(x)$, which is a consistent estimation of the target function. Thus the theoretical criterion function on the new feature takes the form

$$\begin{aligned} L^*(h,z) = & \operatorname{Var}[y|h \in C_h] - \Pr[h \leq z | h \in C_h] \operatorname{Var}[y|h \leq z] \\ & - \Pr[h \geq z | h \in C_h] \operatorname{Var}[y | h \geq z] \\ = & -\frac{1}{|C_h|^2} \left(\int_{C_h} f(t) \, \mathrm{d}t \right)^2 + \frac{1}{|C_{h \leq z}||C_h|} \left(\int_{C_{h \leq z}} f(t) \, \mathrm{d}t \right)^2 \\ & + \frac{1}{|C_{h \geq z}||C_h|} \left(\int_{C_{h \geq z}} f(t) \, \mathrm{d}t \right)^2 \\ = & -\frac{1}{|C_f|^2} \left(\int_{C_f} f(t) \, \mathrm{d}t \right)^2 + \frac{1}{|C_{f \leq z}||C_f|} \left(\int_{C_{f \leq z}} f(t) \, \mathrm{d}t \right)^2 \\ & + \frac{1}{|C_{f \geq z}||C_f|} \left(\int_{C_{f \geq z}} f(t) \, \mathrm{d}t \right)^2 \end{aligned}$$

Let $I = \int_{C_f} f(t)\,\mathrm{d}t = \int_a^b f(t)\,\mathrm{d}t$ and $G(z) = \int_{C_{f\leq z}} f(t)\,\mathrm{d}t$. Then, the theoretical criterion function becomes

$$L^*(h, z) = \frac{1}{|C_{h\geq z}||C_{h\leq z}|}\left(G(z) - I \cdot \frac{|C_{h\leq z}|}{|C_h|}\right)^2 \ . \tag{39}$$

Without loss of generality, we let $I = 0$, then the optimal split along the raw feature is

$$z^\star = \arg\max_{z\in[a,b]} L^*(1, z) = \arg\max_{z\in[a,b]} \frac{1}{(z-a)(b-z)} G^2(z) \ . \tag{40}$$

We compare the maximum value of the CART-split criterion along the raw feature and the new feature:

$$\begin{aligned} \max_z L^*(1, z) &= \frac{1}{(z^*-a)(b-z^*)} F^2(z^*) \\ \triangleright \text{ According to Lemma 5} \quad &= \frac{(z^*-a)(b-z^*)}{(b-a)^2}(\mu_{a\leq x\leq z^*} - \mu_{z^*\leq x\leq b})^2 \\ \triangleright \text{ According to Lemma 6} \quad &\leq \frac{|C_{h\leq f(z^\star)}||C_{h\geq f(z^\star)}|}{|C_h|^2}(\mu_{h\leq f(z^\star)} - \mu_{h\geq f(z^\star)})^2 \\ \triangleright \text{ According to Lemma 5} \quad &= \frac{1}{|C_{h\leq f(z^\star)}||C_{h\geq f(z^\star)}|} G^2(f(z^\star)) \\ &\leq \max_z L^*(h, z) \ , \end{aligned} \tag{41}$$

and the $d = 1$ case is proved.

Next, we consider the $d > 1$ case, where $A = \prod_{j=1}^d [a_j, b_j] \subset [0, 1]^d$. We know that, for all $1 \leq j \leq d$, there exists a constant I such that

$$\begin{aligned} &\int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} f(\boldsymbol{x})\,\mathrm{d}x^{(1)} \ldots x^{(j-1)} x^{(j+1)} \ldots x^{(d)} \\ =&f_j(x^{(j)}) + \int_{a_1}^{b_1} \cdots \int_{a_d}^{b_d} \sum_{\ell\neq j} f_\ell(x^{(\ell)})\,\mathrm{d}x^{(1)} \ldots x^{(j-1)} x^{(j+1)} \ , \end{aligned} \tag{42}$$

which can be simply denoted as

$$\int_{C_{x^{(-j)}}} f(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}^{(-j)} = f_j(x^{(j)}) + \int_{C_{x^{(-j)}}} \sum_{\ell\neq j} f_\ell(x^{(\ell)})\,\mathrm{d}\boldsymbol{x}^{(-j)} \ . \tag{43}$$

Let $I_j = \int_{C_{x^{(-j)}}} \sum_{\ell\neq j} f_\ell(x^{(\ell)})\,\mathrm{d}\boldsymbol{x}^{(-j)}$, which does not depend on $x^{(j)}$. Since $f(\cdot)$ is an additive model, for all j and all $x^{(j)}$,

$$\int_{a_j}^{z_j} \int_{C_{x^{(-j)}}} f_j(\boldsymbol{x})\,\mathrm{d}\boldsymbol{x}^{(-j)}\,\mathrm{d}x^{(j)} \ = \int_{a_j}^{z_j} f(x^{(j)})\,\mathrm{d}x^{(j)} + (z_j - a_j) I_j \ . \tag{44}$$

Let $z_j = z$, $a_j = a$, $b_j = b$, the theoretical criterion function on the raw feature takes the form

$$\begin{aligned} L^*(j, z) =& \operatorname{Var}[y|x^{(j)} \in C_x] - \Pr[a \leq x^{(j)} \leq z | x^{(j)} \in C_x] \operatorname{Var}[y|a \leq x^{(j)} \leq z] \\ &- \Pr[z \leq x^{(j)} \leq b | x^{(j)} \in C_x] \operatorname{Var}[y|z \leq x^{(j)} \leq b] \\ =& -\frac{1}{(b-a)^2}\left(\int_{C_{x^{(-j)}}} \int_a^b f(x^{(j)})\,\mathrm{d}x^{(j)}\,\mathrm{d}\boldsymbol{x}^{(-j)}\right)^2 \\ &+ \frac{1}{(b-a)(z-a)}\left(\int_{C_{x^{(-j)}}} \int_a^z f(x^{(j)})\,\mathrm{d}x^{(j)}\,\mathrm{d}\boldsymbol{x}^{(-j)}\right)^2 \\ &+ \frac{1}{(b-a)(b-z)}\left(\int_{C_{x^{(-j)}}} \int_z^b f(x^{(j)})\,\mathrm{d}x^{(j)}\,\mathrm{d}\boldsymbol{x}^{(-j)}\right)^2 \\ =& \frac{1}{(z-a)(b-z)}\left(F(z) - I \cdot \frac{z-a}{b-a}\right)^2 + \frac{(b-a)^2 + (b-z)^2 + (z-a)^2}{2(b-a)^2(b-z)(z-a)} I_j \end{aligned}$$

According to the proof of $d = 1$, we can fix $z - a$ and $b - z$. Then we just need consider $\frac{1}{(z-a)(b-z)}\left(F(z) - I \cdot \frac{z-a}{b-a}\right)^2$, which is same as $d = 1$.

Intuitively, in the multi-dimensional case, the correlation between y and $\boldsymbol{x}$ is scattered to each dimension due to the assumption that each dimension is independent and $y = f(\boldsymbol{x}) + \epsilon$ is an additive model. Therefore, when splitting is calculated separately in each dimension, the gain is not as large as that caused by the only dimension splitting in the case of one dimension $L^*(1, z)$ above. According to the result of $d = 1$, we have

$$\max_{j,z} L^*(j, z) < \max_{z} L^*(h, z) \; , \tag{45}$$

and the $d > 1$ case is proved.

2. Finite sample region ($n < \infty$)

Fix $k \in \mathbb{N}^*$ and $\rho, \xi > 0$. According to Lemma 7, with probability $1 - \rho$, for all n large enough, there exists a sequence of theoretical first k splits $\boldsymbol{t}^*_k(\boldsymbol{x}, \Theta)$ such that

$$d_\infty\left(\boldsymbol{t}^*_k(\boldsymbol{x}, \Theta), \hat{\boldsymbol{t}}_{k,n}(\boldsymbol{x}, \Theta)\right) \leq \xi \; . \tag{46}$$

Therefore, with probability $\geq 1 - \rho$, for all n large enough and all $1 \leq j \leq k$, the j-th empirical split $\hat{t}^*_{j,n}(\boldsymbol{x}, \Theta)$ is performed along the same dimension as $t^*_j(\boldsymbol{x}, \Theta)$. According to the result of theoretical criterion splits, the splits are always performed along the new features, which is the most informative variable. Consequently, for all $\boldsymbol{x}, \Theta$ and for all $1 \leq j \leq k$, each empirical split $\hat{t}^*_{j,n}(\boldsymbol{x}, \Theta)$ is performed along the new features only. Then the proposition is proved. □

Proposition 2 (Variation of f in the empirical cell)**.** *Assume that Assumption 1 holds and the first-layer forest is consistent. The following results hold for any CART in the second-layer forest* $h^{(2)}_n(\cdot, \Theta)$. *After splitting along the new feature, the CART will estimate the residual of the first-layer forest estimation. Then for all* $\rho, \xi > 0$, *there exists* $N \in \mathbb{N}^*$ *such that, for all* $n > N$,

$$\Pr\left[\Delta\left(f, C_{\Pi,n}(\boldsymbol{x}, \Theta)\right) \leq \xi\right] \geq 1 - \rho \; . \tag{11}$$

Proof of Proposition 2. According to Proposition 1, we know that in the theoretical version ($n = \infty$), the CART in the second layer will always split along the new feature. Since $f(\boldsymbol{x})$ is uniformly continuous, the result is clear if $\operatorname{diam}(C^*_k(\boldsymbol{x}, \Theta))$ tends to 0 as k tends to infinity. Thus, in the following proof, we assume that $\operatorname{diam}(C^*_k(\boldsymbol{x}, \Theta))$ does not tend to 0. We denote by h the new feature dimension. $(C^*_k(\boldsymbol{x}, \Theta))$ is a decreasing sequence of compact sets, there exist $a_\infty(h, \Theta)$ and $b_\infty(h, \Theta)$ such that

$$C^*_\infty(\boldsymbol{x}, \Theta) \triangleq \cap_{k=1}^{\infty} C^*_k(\boldsymbol{x}, \Theta) = [a_\infty(h, \Theta), b_\infty(h, \Theta)] \; . \tag{47}$$

Since $\operatorname{diam}(C^*_k(\boldsymbol{x}, \Theta))$ does not tend to zero, we have $a_\infty(h, \Theta) < b_\infty(h, \Theta)$. If the criterion L^* is identically zero for all cuts z in $C^*_\infty(\boldsymbol{x}, \Theta)$ then recall Eq. (39), we have

$$G(z) \propto \frac{z - a}{b - a} \; . \tag{48}$$

This proves that $G(z)$ is linear in z, so f is a constant on $[a, b]$. This implies that $\Delta(f, C^*_\infty(\boldsymbol{x}, \Theta)) = 0$. Since f is uniformly continuous,

$$\lim_{k \to \infty} \Delta(f, C^*_k(\boldsymbol{x}, \Theta)) = \Delta(f, C^*_\infty(\boldsymbol{x}, \Theta)) = 0 \; . \tag{49}$$

Fix $\xi, \rho > 0$. Since almost sure convergence implies convergence in probability, according to the result above, there exists $k_0 \in \mathbb{N}^*$ such that

$$\Pr[\Delta(f, C^*_{k_0}(\boldsymbol{x}, \Theta) \leq \xi] \geq 1 - \rho \; . \tag{50}$$

According to Lemma 7, for all $\xi' > 0$, there exists $N \in \mathbb{N}^*$ such that, for all $n \geq N$,

$$\Pr[d_\infty(\hat{\boldsymbol{t}}_{k_0,n}(\boldsymbol{x}, \Theta), \mathcal{T}^*_{k_0}(\boldsymbol{x}, \Theta)) \leq \xi'] \geq 1 - \rho \; . \tag{51}$$

since f is uniformly continuous, we set ξ' sufficiently small such that, for all $\boldsymbol{x} \in [0, 1]^d$, for all $\boldsymbol{t}_{k_0}, \boldsymbol{t}'_{k_0}$ satisfying $d_\infty(\boldsymbol{t}_{k_0}, \boldsymbol{t}'_{k_0}) \leq \xi'$, we have

$$|\Delta(f, C(\boldsymbol{x}, \boldsymbol{t}_{k_0}) - \Delta(f, C(\boldsymbol{x}, \boldsymbol{t}'_{k_0})| \leq \xi \; . \tag{52}$$

Combining Eq. (51) and (52), we obtain

$$\Pr[|\Delta(f, C_{k_0,n}(\boldsymbol{x}, \Theta)) - \Delta(f, C^*_{k_0}(\boldsymbol{x}, \Theta))| \leq \xi] \geq 1 - \rho\,. \tag{53}$$

Then we can obtain the result from $\Delta(f, C) \leq \Delta(f, C')$ whenever $C \subset C'$,

$$\Pr[\Delta(f, C_{\Pi,n}(\boldsymbol{x}, \Theta)) \leq \xi] \geq 1 - 2\rho\,. \tag{54}$$

□

B.2 Proof of Theorem 3

Theorem 3 (Universal consistency)**.** *Let $M \geq 1$. Consider two-layer deep forest $\bar{h}_{2M,n}$ given by Eq. (6) and Breiman's random forest $h_{M,n}$ given by Eq. (4) for the random CARTs satisfying $a_n \to \infty, t_n \to \infty$ and $t_n(\log a_n)^9/a_n \to 0$. Then under the setting described in Section 3 and assume the data set follows Assumption 1,*

1. *[29, Theorem 1] the Breiman's random forest $h_{2M,n}$ is consistent for any $M \geq 1$,*
2. *the two-layer deep forest $\bar{h}_{M,n}$ is consistent for any $M \geq 1$.*

Proof of Theorem 3. **(T3.1).** The universal consistency of Breiman's random forest is proved by Scornet et al. [29, Thoerem 1].

(T3.2). Similar as Scornet et al. [29, Thoerem 1], we can use the bounded variation of f in the empirical cell in Proposition 2 to control the approximation error. Let $\mathcal{H}_n(\Theta)$ be the set of all functions $h : [0,1]^{d+1} \to \mathbb{R}$ piecewise constant on each cell of the partition $\Pi_n(\Theta)$. Thus the second-layer CART estimator $h_n^{(2)}(\boldsymbol{x}, \Theta)$ satisfies

$$h_n^{(2)}(\boldsymbol{x}, \Theta) \in \mathop{\arg\min}_{h \in \mathcal{H}_n(\Theta)} \frac{1}{a_n} \sum_{i \in \mathcal{I}_{n,\Theta}} |h([\boldsymbol{x}_i, h^{(1)}_{M,n}(\boldsymbol{x}_i)]) - y_i|^2\,. \tag{55}$$

Let $(\beta_n)_n$ be a positive sequence, and define the truncated operator T_{β_n} by

$$\begin{cases} T_{\beta_n} u = u, & \text{if } |u| < \beta_n\,, \\ T_{\beta_n} u = \operatorname{sign}(u)\beta_n, & \text{if } |u| \geq \beta_n\,. \end{cases} \tag{56}$$

Then, we define $T_{\beta_n} h_n^{(2)}(\boldsymbol{x}, \Theta)$, $y_L = T_L y$ and $y_{i,L} = T_L y_i$.

For all n large enough, we have

$$\begin{aligned} \mathbb{E} \inf_{\substack{h \in \mathcal{H}_n(\Theta) \\ \|h\|_\infty \leq \beta_n}} \mathbb{E}_{\boldsymbol{x}}[h(\boldsymbol{x}) - f(\boldsymbol{x})]^2 &\leq \mathbb{E} \inf_{\substack{h \in \mathcal{H}_n(\Theta) \\ \|h\|_\infty \leq \|f\|_\infty}} \mathbb{E}_{\boldsymbol{x}}[h(\boldsymbol{x}) - f(\boldsymbol{x})]^2 \\ &\leq \mathbb{E}[\Delta(f, C_{\Pi,n}(\boldsymbol{x}, \Theta))]^2 \\ &\leq \xi^2 + 4\|f\|^2_\infty \Pr[\Delta(f, C_{\Pi,n}(\boldsymbol{x}, \Theta)) \geq \xi]\,. \end{aligned} \tag{57}$$

Connecting with Proposition 2, we have

$$\mathbb{E} \inf_{\substack{h \in \mathcal{H}_n(\Theta) \\ \|h\|_\infty \leq \beta_n}} \mathbb{E}_{\boldsymbol{x}}[h(\boldsymbol{x}) - f(\boldsymbol{x})]^2 \leq 2\xi^2\,. \tag{58}$$

This proves that the approximation error tends to zero.

The proof of estimation error and untruncated estimate is same as Scornet et al. [29, Thoerem 1]. The parameter t_n allows us to control the size of the leaves of CART, which allows us to have enough samples in each leaf node to smooth the impact of noise, so as to control the estimation error.

$$\Pr\left[\sup_{\substack{h \in \mathcal{H}_n(\Theta) \\ \|h\|_\infty \leq \beta_n}} \left| \frac{1}{a_n} \sum_{i \in \mathcal{I}_{n,\Theta}} [h(\boldsymbol{x}_i) - y_{i,L}]^2 - \mathbb{E}[h(\boldsymbol{x}) - y_L]^2 \right| \geq \xi \right] \leq 8\exp\left(-\frac{a_n C_{\xi,n}}{\beta_n^4}\right)\,, \tag{59}$$

where

$$C_{\xi,n} = \frac{\xi^2}{2048} + \mathcal{O}\left(\frac{t_n(\log a_n)^9}{a_n}\right)\,. \tag{60}$$

According to the condition $t_n(\log a_n)^9/a_n \to 0$, we have $C_{\xi,n} \to \frac{\xi^2}{2048}$. Then, we can bound the estimation error as follow

$$\begin{aligned}\mathbb{E}\left[\sup_{\substack{h\in\mathcal{H}_n(\Theta)\\ \|h\|_\infty\leq\beta_n}}\left|\frac{1}{a_n}\sum_{i\in\mathcal{I}_{n,\Theta}}[h(\boldsymbol{x}_i)-y_{i,L}]^2-\mathbb{E}[h(\boldsymbol{x})-y_L]^2\right|\right] &\leq \xi + 16(\beta_n+L)^2\exp\left(-\frac{a_nC_{\xi,n}}{\beta_n^4}\right)\\ &\leq 2\xi\ .\end{aligned} \tag{61}$$

This proves that the estimation error tends to zero. Connecting the approximation and estimation error together with Györfi et al. [46, Theorem 10.2], the consistency of a CART of second-layer deep forest is proved.

The universal consistency can be proved via Biau et al. [25, Proposition 1], which guarantees that the error of forest estimator is no more than twice that of individual randomized CART. □

B.3 Proof of Theorem 4

Theorem 4 (Depth is more powerful than width)**.** *Let $M \geq 1$. Consider two-layer deep forest $\bar{h}_{2M,n}$ given by Eq. (6) and Breiman's random forest $h_{M,n}$ given by Eq. (4) for the random CARTs satisfying $a_n \to \infty, t_n \to \infty$, $t_n = a_n$ and $a_n \log n/n \to 0$. Then under the setting described in Section 3 and assume the data set follows Assumption 1 and 2, the following results hold*

1. *[29, Theorem 2] [30, Theorem 3] The Breiman's random forest $h_{\infty,n}$ is consistent, and for all $M, n \in \mathbb{N}$,*
$$0 \leq R(h_{2M,n}) - R(h_{\infty,n}) \leq \frac{8\|f\|_\infty^2 + 8\sigma^2(1+4\log n)}{M}\ . \tag{16}$$
2. *The two-layer deep forest $\bar{h}_{\infty,n}$ is consistent, and for all $M, n \in \mathbb{N}$, if $\Delta(f, C_{\Pi,n}(\boldsymbol{x},\Theta))$ is small enough, then*
$$0 \leq R(\bar{h}_{2M,n}) - R(\bar{h}_{\infty,n}) \leq \frac{64\|f\|_\infty^2 + 64\sigma^2(1+4\log n)}{M^2}\ . \tag{17}$$

Proof of Theorem 4. **(T4.1).** The consistency of the infinite Breiman's random forest is proved by Scornet et al. [29]. And the convergence rate of the finite random forest with the number of trees M is proved by Scornet [30].

(T4.2). Because each cell contains only one sample in this regime, we define

$$W_{ni}(\boldsymbol{x}) = \mathbb{E}_\Theta[\mathbb{1}_{\boldsymbol{x}_i\in C_{\Pi,n}(\boldsymbol{x},\Theta)}]\ , \tag{62}$$

the infinite two layer deep forest estimation can rewriten as

$$\bar{h}_{\infty,n}(\boldsymbol{x}) = h^{(2)}_{\infty,n}([\boldsymbol{x}, h^{(1)}_{\infty,n}(\boldsymbol{x})]) = \sum_{i=1}^n W_{ni}([\boldsymbol{x}, h^{(1)}_{M,n}(\boldsymbol{x})])y_i\ . \tag{63}$$

Thus,

$$\begin{aligned}\mathbb{E}[\bar{h}_{\infty,n}(\boldsymbol{x}) - f(\boldsymbol{x})] \leq& 2\mathbb{E}\left[\sum_{i=1}^n W_{ni}([\boldsymbol{x}, h^{(1)}_{\infty,n}(\boldsymbol{x})])(y_i - f(\boldsymbol{x}_i))\right]^2\\ &+ 2\mathbb{E}\left[\sum_{i=1}^n W_{ni}([\boldsymbol{x}, h^{(1)}_{\infty,n}(\boldsymbol{x})])(f(\boldsymbol{x}_i) - f(\boldsymbol{x}))\right]^2\\ \triangleq& 2I_n + 2J_n\ .\end{aligned} \tag{64}$$

Similar as Scornet et al. [29], we recall Proposition 2 to control the approximation error of the two-layer deep forest.

$$\begin{aligned}J_n \leq& \mathbb{E}\left[\sum_{i=1}^n \mathbb{1}_{[\boldsymbol{x}_i, h^{(1)}_{\infty,n}(\boldsymbol{x}_1)]\in C_{\Pi,n}([\boldsymbol{x}, h^{(1)}_{\infty,n}(\boldsymbol{x})],\Theta)}\Delta^2(f, C_{\Pi,n}([\boldsymbol{x}, h^{(1)}_{\infty,n}(\boldsymbol{x})],\Theta))\right]\\ \leq& \mathbb{E}[\Delta^2(f, C_{\Pi,n}([\boldsymbol{x}, h^{(1)}_{\infty,n}(\boldsymbol{x})],\Theta))]\\ \leq& \xi(4\|f\|_\infty^2 + 1) \qquad \triangleright \quad \text{According to Proposition 2.}\end{aligned} \tag{65}$$

The proof of estimation error is same as Scornet et al. [29, Thoerem 2]. the estimation error is controlled by forcing the subsampling rate a_n/n to be $o(1/\log n)$. For simplification, we denote $[\boldsymbol{x}, h^{(1)}_{\infty,n}(\boldsymbol{x})]$ as $\mathbf{X}$

$$\begin{aligned} I_n &= \mathbb{E}\left[\sum_{i,j=1}^{n} W_{ni}(\mathbf{X})W_{nj}(\mathbf{X})\left(y_i - f\left(\boldsymbol{x}_i\right)\right)\left(y_j - f\left(\boldsymbol{x}_j\right)\right)\right] \\ &= \mathbb{E}\left[\sum_{i=1} W^2_{ni}(\mathbf{X})\left(y_i - f\left(\boldsymbol{x}_i\right)\right)^2\right] + I'_n \ , \end{aligned} \tag{66}$$

where

$$I'_n = \mathbb{E}\left[\sum_{\substack{i,j \\ i \neq j}} \mathbb{1}_{\mathbf{X}\overset{\Theta}{\leftrightarrow}\mathbf{X}_i}\mathbb{1}_{\mathbf{X}\overset{\Theta'}{\leftrightarrow}\mathbf{X}_j}\left(y_i - f\left(\boldsymbol{x}_i\right)\right)\left(y_j - f\left(\boldsymbol{x}_j\right)\right)\right] \ . \tag{67}$$

By Assumption 2 and Scornet et al. [29, Lemma 4], for all n large enough, $|I'_n| \leq \xi$. Then,

$$\begin{aligned} |I_n| \leq& \xi + \mathbb{E}\left[\max_{1\leq \ell \leq n} W_{n\ell}(\mathbf{X}) \max_{1\leq i\leq n} \varepsilon_i^2\right] \\ \leq& \xi + \max_{1\leq i\leq n} \Pr_{\Theta}\left[\mathbf{X} \overset{\Theta}{\leftrightarrow} \mathbf{X}_i\right]\mathbb{E}\left[\max_{1\leq i\leq n} \varepsilon_i^2\right] \\ \leq& \xi + \frac{a_n}{n}\mathbb{E}\left[\max_{1\leq i\leq n} \varepsilon_i^2\right] \\ \leq& \xi + C\frac{a_n \log n}{n} \leq 2\xi \ . \qquad \triangleright \text{ According to } a_n/n \sim o(1/\log n) \ . \end{aligned} \tag{68}$$

Connecting the approximation and estimation error together, the consistency of an infinite two layer deep forest is proved.

Different from the bagging-style mechanism in random forest, the residual-style mechanism shown in Proposition 2 makes the second-layer forest in DF can reuse the first-layer estimation and focus on the residual learning. According to Scornet [30, Theorem 3.3], we have

$$R(h_{M,n}) - R(h_{\infty,n}) \leq \frac{8}{M} \times \left(\|f\|^2_\infty + \sigma^2(1 + 4\log n)\right) \ , \tag{69}$$

for the first-layer forest estimation. When n is large enough, we have $R(h_{\infty,n}) < \xi$ and

$$R(h_{M,n}) \leq \xi + \frac{8}{M} \times \left(\|f\|^2_\infty + \sigma^2(1 + 4\log n)\right) \ . \tag{70}$$

According to Proposition 1, the first k splits are only along the new feature dimension. This is equivalent to using a piecewise constant function of $h^{(1)}_{M,n}$ to copy the first-layer forest estimation, which is independent of Θ. After the size of piece is small than the first-layer error, the raw features are used to estimate the residual $r(\boldsymbol{x}) = f(\boldsymbol{x}) - h^{(1)}_{M,n}(\boldsymbol{x})$. Thus, we obtain the bound for residual :

$$\|r\|^2_\infty \leq \frac{8}{M} \times \left(\|f\|^2_\infty + \sigma^2(1 + 4\log n)\right) \ . \tag{71}$$

As for the noise of the residual, we first consider the $R(h^{(1)}_{M,n}) \geq \mathbb{E}[\epsilon^2] = \sigma^2$ case: The first-layer estimator is too weak to filter noise, so the noise of the residual is still ϵ. Next, we consider the $R(h^{(1)}_{M,n}) < \mathbb{E}[\epsilon^2] = \sigma^2$ case: Since the first-layer estimator smoothes part of the noise, the noise ϵ' is reduced in the residual. When n is large enough, the size of noise can be bounded by the variation of f in the empirical cell of the first-layer forest

$$\mathbb{E}\epsilon'^2 \leq c\Delta^2(f, C_{\Pi,n}(\boldsymbol{x}, \Theta)) \ . \tag{72}$$

Then we obtain the risk of the finite second-layer forest,

$$\begin{aligned}
R(h_{M,n}^{(2)}) =& \xi + \frac{1}{M} \times \mathbb{E}\left[\mathrm{Var}_{\Theta}\left[\sum_{i=1}^{n} W_{ni}(\boldsymbol{x},\Theta)\left(r(\boldsymbol{x}_i)+\varepsilon_i'\right)\right]\right] \\
\leq& \xi + \frac{1}{M} \times \left[8\|r\|_{\infty}^{2} + 2\mathbb{E}\left[\mathbb{V}_{\Theta}\left[\sum_{i=1}^{n} W_{ni}(\boldsymbol{x},\Theta)\varepsilon_i'\right]\right]\right] \\
\leq& \xi + \frac{1}{M} \times \left[8\|r\|_{\infty}^{2} + 8\sigma'^{2}\mathbb{E}\left[\max_{1\leq i\leq n}\frac{\varepsilon_i'}{\sigma'}\right]^{2}\right] \\
\leq& \xi + \frac{64\|f\|_{\infty}^{2} + 64\sigma^{2}(1+4\log n)}{M^{2}} + \frac{c\Delta^{2}(f, C_{\Pi,n}(\boldsymbol{x},\Theta))}{M}\ .
\end{aligned} \tag{73}$$

Thus, if the variation of f in the empirical cell is small enough, then the two layer deep forest can obtain a faster convergence rate *w.r.t.* M. This theorem is proved. □

C Additional Results for Simulation Experiments

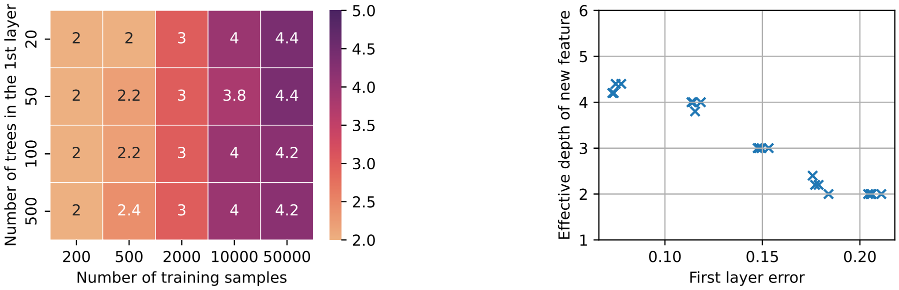
The experimental setting is the same as in Section 7, except that we set $f(\boldsymbol{x})$ to be a nonlinear function, that is,

$$f(\boldsymbol{x}) = \frac{1}{5}\left(\sin 2\pi x_1 + \cos 2\pi x_2 + \sin(2\pi x_3 + \pi/3) + \cos(2\pi x_4 + \pi/3) + \sin 6\pi x_5\right)\ . \tag{74}$$

As plotted in Figure S5, we also observe the same tendency as in Section 7.2, that using inconsistent trees is better in practice, and the 2-layer deep forest (DF) convergences faster *w.r.t.* the number of trees M. We also check the effect depth as in Section 7.1, and Figure S4 convinces us that the new feature has priority in splitting. Furthermore, we set $f(\boldsymbol{x})$ to be an interacted function,

$$f(\boldsymbol{x}) = \frac{1}{5}\left(x_1 + x_2x_3 + x_2^2x_3^{1/2} + x_3\sin 2\pi x_4 + \sin(2\pi x_4)\cos(6\pi x_5 + \pi/4)\right)\ . \tag{75}$$

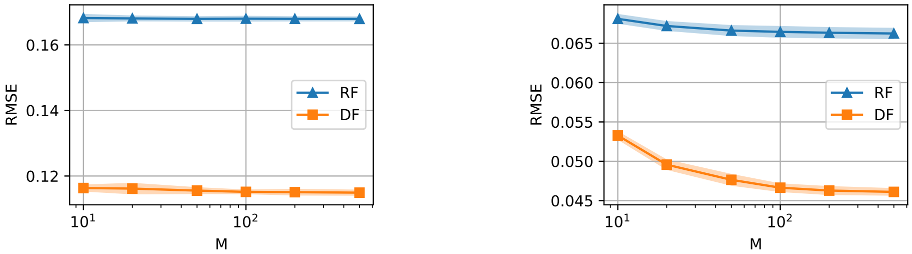
As plotted in Figure S7, using inconsistent trees is better in practice, and the 2-layer deep forest (DF) convergences faster *w.r.t.* the number of trees M. Figure S6 convinces us that the new feature has priority in splitting.



(a) Heatmap of the effective depth of new feature under different settings of the first layer forest.

(b) Effective depth of new feature against the predictive error of the first layer.

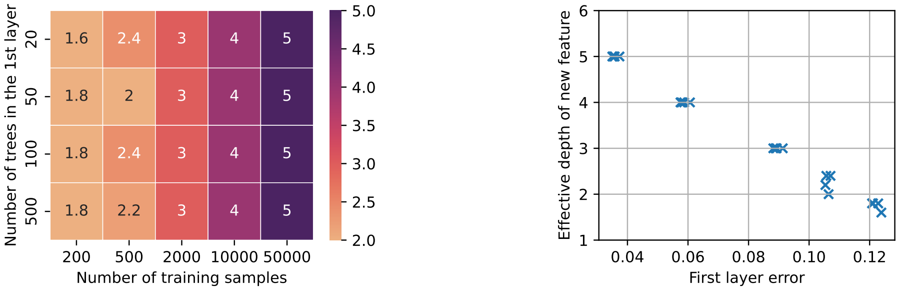
Figure S4: Illustrations of the effective depth of the new feature (the consecutive levels from root node that split on the new feature only). The larger the effective depth, the higher priority the new feature takes in being chosen as the split feature under the CART-split criterion.



(a) Consistent trees ($t_n < a_n$).

(b) Inconsistent trees ($t_n = a_n$).

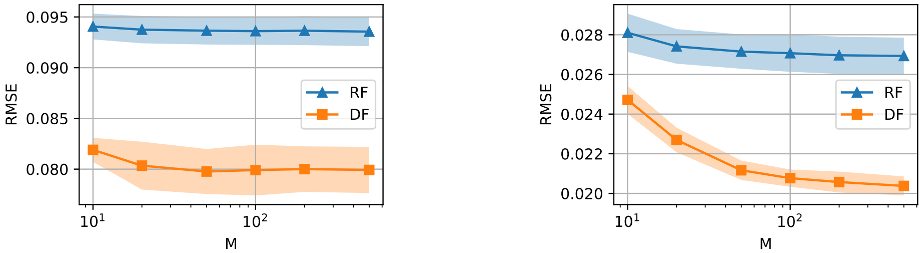
Figure S5: Root mean square error with the increasing of number of trees M.



(a) Heatmap of the effective depth of new feature under different settings of the first layer forest.

(b) Effective depth of new feature against the predictive error of the first layer.

Figure S6: Illustrations of the effective depth of the new feature (the consecutive levels from root node that split on the new feature only). The larger the effective depth, the higher priority the new feature takes in being chosen as the split feature under the CART-split criterion.



(a) Consistent trees ($t_n < a_n$).

(b) Inconsistent trees ($t_n = a_n$).

Figure S7: Root mean square error with the increasing of number of trees M.

D Results on real world regression problems

In this section, we conduct experiments on real-world data sets. It should be noted that when we do experiments on real-world data sets, the number of samples is finite and the true underlying function is unknown, so there is gap from the theoretical analysis of consistency. Even so, the generalization performance on real data sets still shows a tendency for 2-layer deep forest to be more efficient than 1-layer random forest.

Data sets. We conduct experiments on 3 real-world regression problems and the detail statistics of the data sets are shown in Table S2.

1. `housing data set`: This dataset contains information collected by the U.S Census Service concerning housing in the area of Boston Mass. It was obtained from the StatLib archive[1], and has been used extensively throughout the literature to benchmark algorithms. However, these comparisons were primarily done outside of Delve and are thus somewhat suspect.
2. `cadata data set`: This data set gathers information on housing prices using all neighborhood groups in California from the 1990 census. It calculates the distance between the centroids of each block group measured in latitude and longitude. It excludes all block groups reporting zero entries for the independent and dependent variables.
3. `acoustic data set`: This data set is collected from simulation result of COMSOL platform. It aims at predicting the energy focusing effect of an acoustic system based on 21 angle parameters.

Data set	# of samples	# of features
housing	506	13
cadata	20,640	8
acoustic	4,000	21

Table S2: The average test error measured by RMSE of 5 runs on benchmark data sets. DF is better than RF in test error.

Generalization performance. M is set to 500 and the trees are fully grown as is commonly used in the literature. The average RMSE on test set of 5 runs is reported in Table S3.

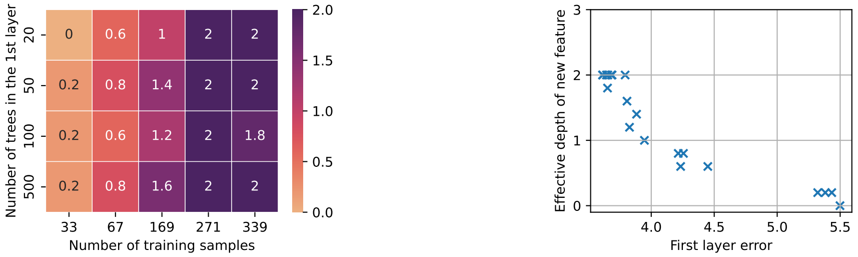
Data set	RF	DF
housing	3.62	**3.56**
cadata	50208	**49363**
acoustic	2.47	**2.34**

Table S3: The average test error measured by RMSE of 5 runs on benchmark data sets. DF is better than RF in test error.

Priority of new features. We vary the number of training samples from 10% to 100% and vary the number of trees M to get different first-layer models. If we check the effective depth in the second-layer tree as shown in Figure S8, S9 and S10, we can also observe that the second layer tree will always choose the new feature to split. This verifies that Proposition 1 also holds in real world data sets.

Convergence rate w.r.t. M. Figure S11 shows that DF enjoys a faster improvement in RMSE with the increasing of M in these three real-world data sets. These experimental results match our theoretical analysis in Theorem 4 that DF has a faster convergence rate *w.r.t.* the number of trees M.

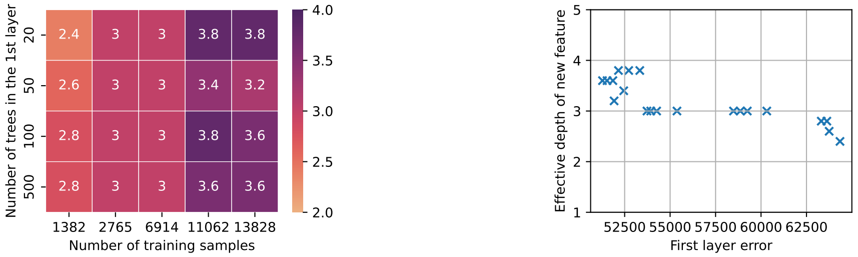
[1] http://lib.stat.cmu.edu/datasets/boston



(a) Heatmap of the effective depth of new feature under different settings of the first layer forest.

(b) Effective depth of new feature against the predictive error of the first layer.

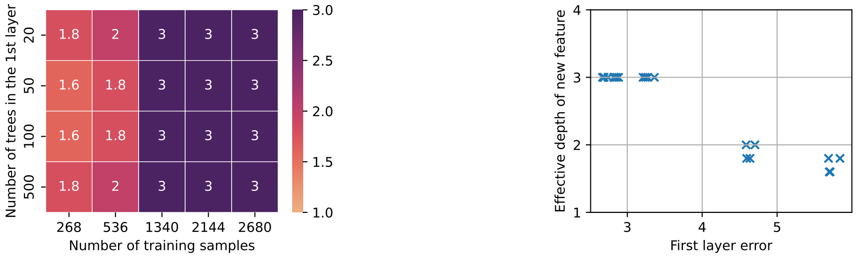
Figure S8: Priority of new features in housing data set.



(a) Heatmap of the effective depth of new feature under different settings of the first layer forest.

(b) Effective depth of new feature against the predictive error of the first layer.

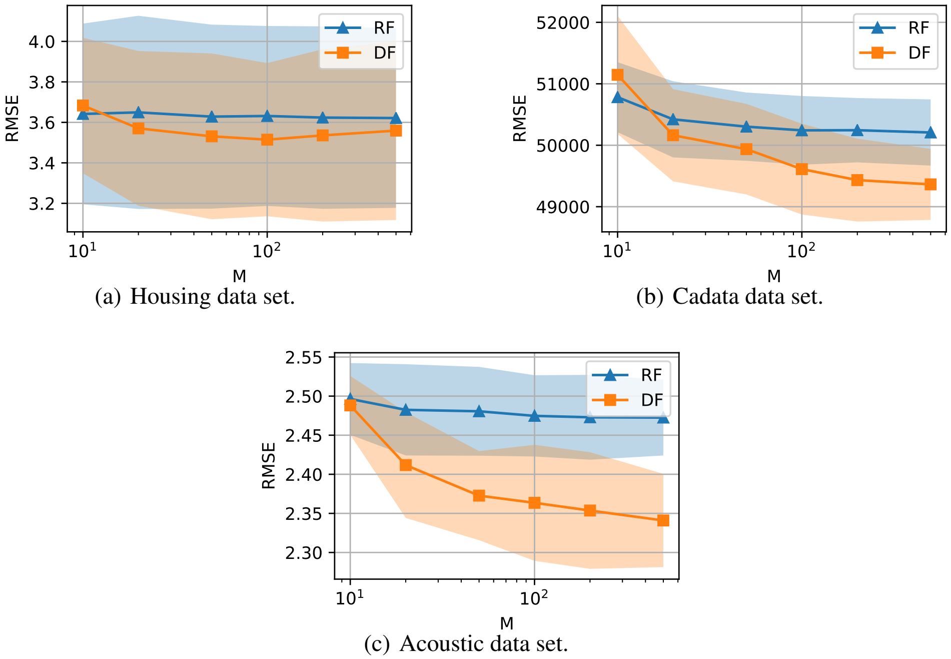
Figure S9: Priority of new features in cadata data sets.



(a) Heatmap of the effective depth of new feature under different settings of the first layer forest.

(b) Effective depth of new feature against the predictive error of the first layer.

Figure S10: Priority of new features in acoustic data set.



(a) Housing data set.

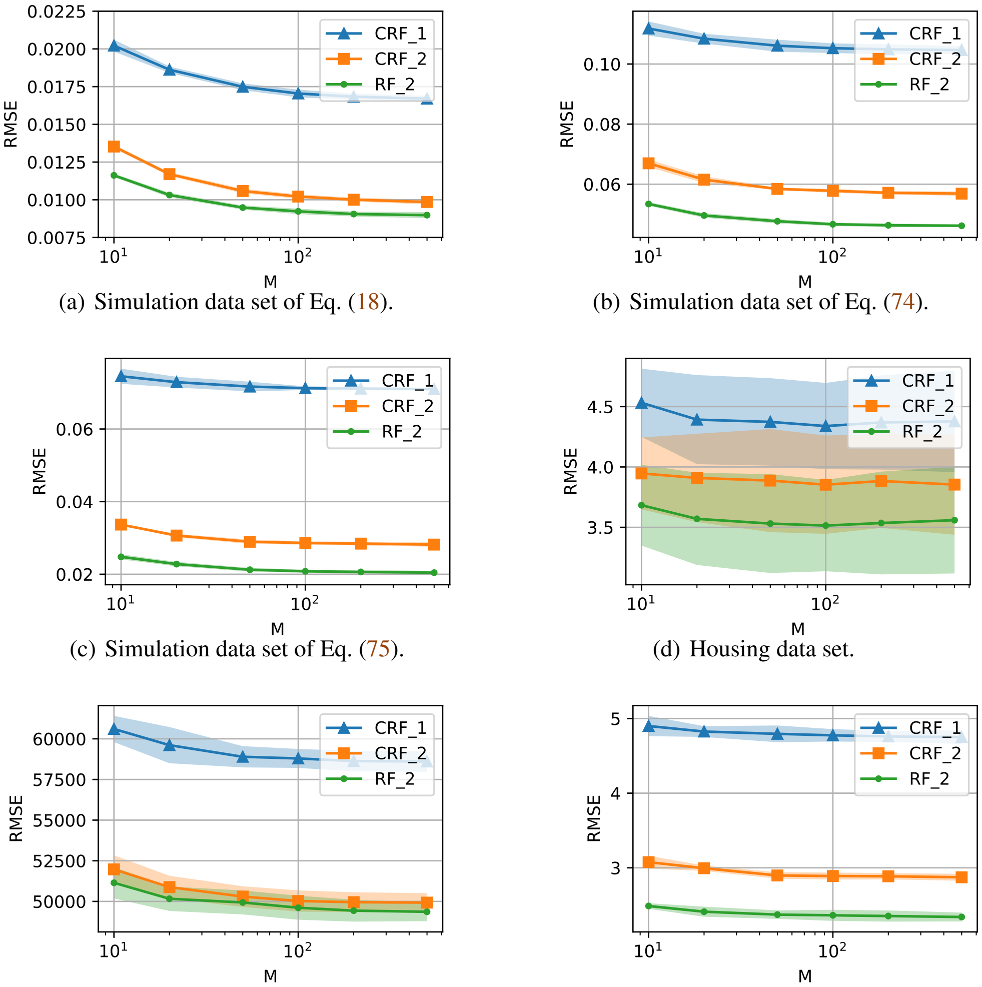
(b) Cadata data set.

(c) Acoustic data set.

Figure S11: Root mean square error with the increasing of number of trees M in real-world data sets.

E Dependence on label information in Assumption 2

In this section, we design a simple comparative experiment to show that the new features make the second-layer forest estimator much less dependent on label information. Specifically, we use a Completely Random Forest (CRF) to replace the random forest in the second layer of the deep forest, and CRF-split criterion does not depend on label information at all. The synthetic and real-world data sets used here are the same as Section 7.1, C and D. In Figure S12, we can find that the performance of the second-layer CRF can be close to the second-layer RF, and significantly outperforms the CRF trained on the original features. This implies that the new features play a positive role in reducing the model's dependence on label information.



(a) Simulation data set of Eq. (18).

(b) Simulation data set of Eq. (74).

(c) Simulation data set of Eq. (75).

(d) Housing data set.

(e) Cadata data set.

(f) Acoustic data set.

Figure S12: Root mean square error with the increasing of number of trees M. CRF_1 represents the completely random forest trained on the original feature space. CRF_2 represents the completely random forest trained on the new feature space. RF_2 represents the Breiman's random forest trained on the new feature space. As the number of samples in the dataset is larger, the performance of CRF and RF at layer 2 is more similar.